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

Lesi Chen1  Haishan Ye2  Luo Luo3,*

1Institute for Interdisciplinary Information Sciences, Tsinghua University
2School of Management, Xi’an Jiaotong University; SGIT AI Lab, State Grid Corporation of China
3School 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 \(\epsilon\)-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 \(\mathcal{O}(\kappa^2 \epsilon^{-2})\) communication rounds, where \(\kappa\) is the condition number and \(N\) is the total number of individuals. 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

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

(1)

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

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

(2)

and the stochastic component \(F_i(x, y; \xi_i)\) indexed by the random variable \(\xi_i\) 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), 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 \(\epsilon\)-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 \(\xi_i\) is finitely sampled from \(\{\xi_{i,1}, \ldots, \xi_{i,n}\}\). That is, we can write the local function as

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

(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 \(\epsilon\)-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 \(\mathcal{O}(\min(\kappa^3 \epsilon^{-3}, \sqrt{n} \kappa^2 \epsilon^{-2}))\).
An Efficient Stochastic Algorithm for Decentralized Nonconvex-Strongly-Concave Minimax Optimization

It is worth noting that the $\mathcal{O}(\min(\epsilon^{-3}, \sqrt{n}\epsilon^{-2}))$ dependency on $\epsilon$ 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; Hendricks et al. 2021; Kovalev et al. 2020; Li et al. 2020, 2022a; 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 $\epsilon$-stationary point of $P(x)$ within $\mathcal{O}(N\epsilon^{-2})$ SFO calls and $\mathcal{O}(\epsilon^{-2})$ communication rounds. Zhang et al. (2021) 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 $\epsilon$-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 $\mathcal{O}(N + \sqrt{m}N\epsilon^{-2})$ SFO complexity to find an $\epsilon$-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 $\mathcal{O}(\kappa^2\epsilon^{-3})$ 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 $\epsilon^{-1}$, 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 not have an 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 $y$ 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 $\mathcal{Y} = \{y \in \mathbb{R}^{d_y} : |y|_\infty \leq c\}$ for some positive constant $c$. In distributionally robust optimization (Yan et al. 2019), the variable $y$ represents the probability distribution in the simplex $\mathcal{Y} = \{y \in \mathbb{R}^{d_y} : \sum_k y_k = 1, y_k \geq 0\}$. 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 $\mathcal{O}(N + \sqrt{N}\kappa\epsilon^{-2})$ and the communication complexity of $\mathcal{O}(\kappa^2\epsilon^{-2})$. 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 $\epsilon$, up to logarithmic factors.
- For the online case, DREAM achieves the $\mathcal{O}(\kappa^2\epsilon^{-3})$ SFO complexity and the $\mathcal{O}(\kappa^2\epsilon^{-2})$ communication complexity. The SFO complexity of DREAM is optimal in the dependency of $\epsilon$ (Arjevani et al.
We compare our theoretical results with previous work in Table I.

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 $1 = [1, \cdots, 1]^\top \in \mathbb{R}^m$. We define aggregated variables $x \in \mathbb{R}^{m \times d_x}$, $y \in \mathbb{R}^{m \times d_y}$ and $z \in \mathbb{R}^{m \times (d_z + d_y)}$ for all agents as
\[
x = \begin{bmatrix} x(1) \\ \vdots \\ x(m) \end{bmatrix}, \quad y = \begin{bmatrix} y(1) \\ \vdots \\ y(m) \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} z(1) \\ \vdots \\ z(m) \end{bmatrix},
\]
where row vectors $x(i) \in \mathbb{R}^{d_x}$ and $y(i) \in \mathbb{R}^{d_y}$ are local variables on the $i$-th agent; and we also denote $z(i) = [x(i); y(i)] \in \mathbb{R}^d$ with $d = d_x + d_y$. We use the lowercase with the bar to represent mean vector, e.g.,
\[
\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x(i), \quad \bar{y} = \frac{1}{m} \sum_{i=1}^{m} y(i) \quad \text{and} \quad \bar{z} = \frac{1}{m} \sum_{i=1}^{m} z(i).
\]

Similarly, we also introduce the aggregated gradient as
\[
\nabla f(z) = \begin{bmatrix} \nabla f_1(x(1), y(1))^\top \\ \vdots \\ \nabla f_m(x(m), y(m))^\top \end{bmatrix} \in \mathbb{R}^{m \times d}.
\]

We use $1[\cdot]$ 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 $P(x) \triangleq \max_{y \in Y} f(x, y)$ 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 $F_i(x, y; \xi)$ on each agent is $L$-average smooth for some $L > 0$. That is, we have
\[
\mathbb{E} \| \nabla F_i(x, y; \xi) - \nabla F_i(x', y; \xi) \|^2 \\
\leq L^2 (\|x - x'\|^2 + \|y - y'\|^2)
\]
for any $(x, y), (x', y') \in \mathbb{R}^{d_x \times d_y}$ and random index $\xi$.

**Assumption 2.3.** We suppose each local function $f_i(x, y)$ is $\mu$-strongly-concave in $y$. That is, there exists some constant $\mu > 0$ such that we have
\[
f_i(x, y) \leq f_i(x, y') + \nabla_y f_i(x, y')^\top (y - y') - \frac{\mu}{2} \| y - y' \|^2
\]
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 $\mu$ 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 $y^*(x) \triangleq \arg \max_{y \in Y} f(x, y)$.

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

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

The goal of algorithms is to find an $\epsilon$ 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) $\nabla F_i(x, y; \xi)$ is unbiased and has bounded variance. That is, we have
\[
\mathbb{E}[\nabla F_i(x, y; \xi)] = \nabla f_i(x, y)
\]
and
\[
\mathbb{E} \| \nabla F_i(x, y) - \nabla f_i(x, y; \xi) \|^2 \leq \sigma^2
\]
with $\sigma^2 < +\infty$. 

---

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

1: Initialize: $a^{(-1)} = a^{(0)}$
2: $\eta_0 = 1/(1 + \sqrt{1 - A^2(W)})$
3: for $k = 0, 1, \ldots, K$ do
4: \[ a^{(k+1)} = (1 + \eta_0)Wa^{(k)} - \eta_0 a^{(k-1)} \]
5: end for
6: Output: $a^{(K)}$

---

Lesi Chen1, Haishan Ye2, Luo Luo3,*
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 $\kappa^{p}$ and $\kappa^{q}$ are used when the polynomial dependency on $\kappa$ is not explicitly provided [Tsaknakis et al. 2020; Zhang et al. 2021b]. The notation $\tilde{O}(\cdot)$ 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 $\mathcal{Y} = \mathbb{R}^{k}$. 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.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Algorithm</th>
<th>#SFO</th>
<th>#Communication</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GT-DA</td>
<td>$\tilde{O}(N\kappa^{p}\epsilon^{-2})$</td>
<td>$\tilde{O}(\kappa^{q}\epsilon^{-2})$</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Tsaknakis et al. [2020]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online</td>
<td>GT-GDA</td>
<td>$\mathcal{O}(N\kappa^{p}\epsilon^{-2})$</td>
<td>$\mathcal{O}(\kappa^{q}\epsilon^{-2})$</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [2021b]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GT-SRVR</td>
<td>$\mathcal{O}(N + \sqrt{mN}\kappa^{p}\epsilon^{-2})$</td>
<td>$\mathcal{O}(\kappa^{q}\epsilon^{-2})$</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [2021b]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DREAM (Ours)</td>
<td>$\mathcal{O}(N + \sqrt{N}\kappa^{p}\epsilon^{-2})$</td>
<td>$\tilde{O}(\kappa^{2}\epsilon^{-2})$</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Theorem 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online</td>
<td>DM-HSGD</td>
<td>$\mathcal{O}(\kappa^{3}\epsilon^{-3})$</td>
<td>$\mathcal{O}(\kappa^{3}\epsilon^{-3})$</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>Xian et al. [2021]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DREAM (Ours)</td>
<td>$\mathcal{O}(\kappa^{3}\epsilon^{-3})$</td>
<td>$\tilde{O}(\kappa^{2}\epsilon^{-2})$</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Theorem 3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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: $W1 = W^{\top}1 = 1$.

e. positive semidefinite: $W \succeq O_{m}$.

Note that if a matrix $W$ satisfies Assumption 2.5 a-d, we can let Assumption 2.5 e 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 $a \in \mathbb{R}^{m \times d}$, for the standard mixing iterate given by $a(k+1) = Wa(k)$, we have

$$\|Wa - 1a\| \leq (1 - \delta)^{K}\|a - 1\a\|,$$

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 $\delta$. 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 $\frac{1}{m}1^{\top}a(K) = \tilde{a}(0)$ and

$$\|a(K) - 1a(0)\| \leq c_{1}(1 - c_{2}\sqrt{\delta})K\|\tilde{a}(0) - 1a(0)\|,$$

where $\tilde{a}(0) = \frac{1}{m}1^{\top}a(0)$, $\|\cdot\|$ 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
\[ \Pi(y) = \arg \min_{y' \in \mathcal{Y}} \|y' - y\|^2 \quad \text{and} \quad \Pi'(y) = \arg \min_{y' \in \mathcal{Y}} \|y' - y\|^2 \]
for \( y \in \mathbb{R}^{d_y} \) and \( y \in \mathbb{R}^{m \times d_y} \) 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 \( 0 < \eta \leq 1/L \).

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 \( \epsilon \)-stationary point of \( P(x) \) within at most \( O(k^3 \epsilon^{-3}) \) and \( O(N + \sqrt{N} \kappa^2 \epsilon^{-2}) \) SFO calls for the online and offline setting respectively; and both of two settings require at most \( O(k^3 \epsilon^{-3} \log m/\sqrt{\delta}) \) communication rounds.

3.1 Method Overview

DREAM constructs stochastic recursive gradients \( g_t(i) \) \( \text{[Fang et al., 2018; Li et al., 2021; Luo et al., 2020; Nguyen et al., 2017]} \) to estimate the local gradients, i.e. \( g_t(i) \approx \nabla f(x_t(i)) \). This step (Line 22 in Algorithm 2) reduces the variance of stochastic gradient estimators on each agent, leading to the optimal \( O(\epsilon^{-3}) \) dependency in the SFO upper bound. DREAM then applies the gradient tracking technique \( \text{[Di Lorenzo and Scutari, 2016; Qu and Li, 2019]} \) to track the average gradient over the network via the gradient tracker \( s_t(i) \), i.e. \( s_t(i) \approx \nabla f(z_t(i)) \). The gradient tracker \( s_t(i) \) 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 \( \text{[Nedic et al., 2017; Song et al., 2023]} \). With the gradient tracker, DREAM applies the two-timescale gradient descent ascent \( \text{[Lin et al., 2020a]} \) 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 \( \text{[Luo et al., 2020; Xian et al., 2021; Zhang et al., 2021b]} \), we propose a novel Lyapunov function as follows:
\[
\Phi_t = \Phi_{t-1} + \frac{1}{m} C_t + \frac{\eta}{m p} V_t + \frac{\eta}{p} U_t,
\]
where
\[
\Phi_t = P(x_t) - P^* + \alpha(P(x_t) - f(x_t, y_t)),
\]
\[
C_t = \|z_t - 1/\alpha \|_2^2 + \eta^2 \|s_t - 1/\alpha \|_2^2,
\]
\[
V_t = \frac{1}{m} \|g_t - \nabla f(z_t)\|^2,
\]
\[
U_t = \left\| \frac{1}{m} \sum_{i=1}^m (g_t(i) - \nabla f_t(z_t(i))) \right\|^2.
\]

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

- \( \Phi_t \) measures the optimization error. The gradient descent ascent step ensures that \( \Phi_t \) decrease monotonically at each iteration \( \text{[Chen et al., 2022; Yang et al., 2020]} \).
- \( C_t \) measures the consensus error, which can be bounded by the properties of gradient tracking and consensus steps \( \text{[Song et al., 2023]} \).
- \( V_t \) and \( U_t \) measure the variance of \( g_t \), which can be bounded by the property of martingale \( \text{[Fang et al., 2018; Luo et al., 2020]} \). 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 \( \Phi_t \) characterizes the sub-optimality of the maximization problem \( \max_y \min_y f(\bar{x}_t, y) \) 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, \( \text{[Xian et al., 2021; Zhang et al., 2021b]} \) 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 \( \eta \leq 1/(4L) \) and
\[
\gamma = \frac{\alpha}{1 + \alpha} 128 \kappa^2.
\]
Algorithm 2 Decentralized Recursive-gradient dEscent Ascent Method (DREAM)

1: Notations: Let $z_t = [x_t, y_t] \in \mathbb{R}^{m \times d}$ and $s_t = [u_t, v_t] \in \mathbb{R}^{m \times d}$.

2: Input: initial parameter $z_0 \in \mathbb{R}^d$, 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 $K'$.

3: $z_0 = 1z_0$

4: parallel for $i = 1, \ldots, m$ do

5: Sample $S_0(i) = \begin{cases} \{\xi_{i,1}, \ldots, \xi_{i,b'}\} \text{ i.i.d., online case;} \\ \{\xi_{i,1}, \ldots, \xi_{i,n}\} \text{, offline case;} \end{cases}$

6: $g_0(i) = \frac{1}{b'} \sum_{\xi_{i,j} \in S_0(i)} \nabla F_i(z_0(i); \xi_{i,j})$

7: end parallel for

8: $s_0 = \text{FastMix}(g_0, K_0)$

9: for $t = 0, \ldots, T - 1$ do

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

11: $x_{t+1} = \text{FastMix}(x_t - \gamma u_t, K)$

12: $y_{t+1} = \text{FastMix}(\Pi(y_t + \eta v_t), K)$

13: parallel for $i = 1, \ldots, m$ do

14: if $\zeta_t = 1$ do

15: Sample $S_t'(i) = \begin{cases} \{\xi_{i,1}, \ldots, \xi_{i,b'}\} \text{ i.i.d., online case;} \\ \{\xi_{i,1}, \ldots, \xi_{i,n}\} \text{, offline case;} \end{cases}$

16: $g_{t+1}(i) = \frac{1}{b'} \sum_{\xi_{i,j} \in S_t'(i)} \nabla F_i(z_{t+1}(i); \xi_{i,j})$

17: else

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

19: Sample $S_t(i) = \{\xi_{i,1}, \ldots, \xi_{i,n}\}$ i.i.d.

20: $g_{t+1}(i) = g_t(i) + \frac{\omega_t(i)}{b q} \sum_{\xi_{i,j} \in S_t(i)} (\nabla F_i(z_{t+1}(i); \xi_{i,j}) - \nabla F_i(z_t(i); \xi_{i,j}))$

21: end if

22: end parallel for

23: $s_{t+1} = \begin{cases} \text{FastMix}(s_t + g_{t+1} - g_t, K'), & \text{if } \zeta_t = 1; \\ \text{FastMix}(s_t + g_{t+1} - g_t, K), & \text{if } \zeta_t = 0; \end{cases}$

24: end for

25: Output: $x_{out}$ by uniformly sampling from $\{x_0(1), x_0(2), \cdots, x_{T-1}(m)\}$.
Then for any $\alpha > 0$ it holds that
\[
\mathbb{E}[\Psi_{t+1}] \leq \Psi_t - \frac{\gamma}{2} \|
abla f(\bar{x}_t)\|^2 - \frac{1}{8\gamma} \mathbb{E}[\|ar{x}_{t+1} - \bar{x}_t\|^2
\]
\[-\frac{\alpha}{16\eta} \mathbb{E}[\|ar{z}_{t+1} - \bar{z}_t\|^2 + 6\alpha\eta U_t + \frac{3\alpha}{m\eta} C_t].
\]

If both $C_t$ and $U_t$ are sufficiently small, the above lemma indicates that the optimization error decreases by roughly $(\gamma\eta/2)\|
abla f(\bar{x}_t)\|^2$ 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:
\[
\rho_0 = c_1(1 - c_2\sqrt{\delta})^K_0,
\]
\[
\rho' = c_1(1 - c_2\sqrt{\delta})^{K'}
\]
\[
\rho = c_1(1 - c_2\sqrt{\delta})^{K}.
\]

Then we can bound the consensus error as follows.

**Lemma 3.2.** For Algorithm 2, let $\rho^2 \leq 1/24$. Then
\[
\mathbb{E}[C_{t+1}] \leq 12\rho^2C_t + 6\rho^2\eta^2mV_t + 2\rho^2m\mathbb{E}[\|\bar{z}_{t+1} - \bar{z}_t\|^2
\]
\[+ \frac{6\rho^2m\eta^2\sigma^2}{b^2}\mathbb{I}[b' < n],\]

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

**Lemma 3.4.** For Algorithm 2 we have
\[
\mathbb{E}[U_{t+1}] \leq (1 - p)U_t + \frac{4(1 - p)L^2}{m^2bq}C_t
\]
\[+ \frac{\rho^2}{m^2b^2}\mathbb{I}[b' < n] + \frac{3(1 - p)L^2}{mbq}\mathbb{E}[\|\bar{z}_{t+1} - \bar{z}_t\|^2].
\]

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 3, we set parameters
\[
\eta = \frac{1}{48L}, \quad b = \left\lceil \sqrt{\frac{b^2}{m}} \right\rceil, \quad q = \frac{1}{5} \sqrt{\frac{b^2}{m}}, \quad p = \frac{bq}{bq + b'}
\]
\[T = \left[\frac{16\psi_0}{\gamma\eta^2} + 2\right], \quad K_0 = \left\lceil \frac{\log (16c_1/(\gamma m^2))}{c_2\sqrt{\delta}} \right\rceil,
\]
\[K = \left\lceil \frac{5\log(c_1(m/b') + 1)}{c_2\sqrt{\delta}} \right\rceil, \quad K' = \left\lceil \frac{5\log(c_1m)}{c_2\sqrt{\delta}} \right\rceil.
\]

where $c_1, c_2$ are defined in Proposition 2.3, $\gamma = \Theta(\kappa^{-2})$ follows (8), $\alpha = 1/8$ and
\[
b' = \begin{cases} \left\lceil \frac{32\sigma^2/(\gamma mc^2)}{n} \right\rceil, & \text{for online case,} \\ n, & \text{for offline case.} \end{cases}
\]

Then the output satisfies $\mathbb{E}\|\nabla P(x_{out})\| \leq \epsilon$ within the overall SFO complexity
\[
\mathcal{O}(\kappa^2\epsilon^{-2} + \kappa^3L\sigma\epsilon^{-3}), \quad \text{for online case;}
\]
\[
\mathcal{O}(mn + \sqrt{mnk^2L}\epsilon^{-2}), \quad \text{for offline case;}
\]
and communication complexity $\mathcal{O}(\kappa^2 L^{-2} \log m) / \sqrt{\delta}$.

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 https://github.com/TrueNobility303/DREAM.

1https://github.com/TrueNobility303/DREAM
We compare the performance of our proposed method with mixing matrix \( W \) where \( l = 0 \) tune for \( \lambda \) since \( \tau \) matrix on a ring graph with 8 nodes (\( m = 8 \)) and \( \nu = 0.0 \), which leads to a low consensus rate.

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

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

\[
f_i(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} (y_i l_{ij}(x) - V(y) + g(x)),
\]

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

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

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

We set mixing matrix \( W \) as the \( \pi \)-lazy random walk matrix on a ring graph with 8 nodes (\( m = 8 \)) and \( \tau = 0.999 \), 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 \( \text{DREAM} \) with \( \text{DM-HSGD} \) (Xian et al. 2021), \( \text{GT-DA} \) (Tsaknakis et al. 2020), \( \text{GT-GDA} \) and \( \text{GT-SRVR} \) (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 \( \text{DREAM} \), we tune \( b, b' \) from \{64, 128, 256, 512\}. For \( \text{GT-SRVR} \), we let the epoch length \( Q = \lceil \sqrt{n} \rceil \) as suggested by the authors. For \( \text{GT-DA} \), we let the number of inner loops \( R = 4 \) by following Tsaknakis et al. (2020). For each algorithm, we tune \( \eta \) and \( \gamma \) 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 \( \text{DREAM} \) performs apparently better than baselines.

5 Conclusion and Future Work

In this paper, we have proposed an efficient algorithm \( \text{DREAM} \) for decentralized stochastic nonconvex-strongly-concave minimax optimization. The theoretical analysis showed \( \text{DREAM} \) achieves the best-known SFO complexity of \( O(\min(\kappa^3 e^{-3}, \kappa^2 \sqrt{N} e^{-2})) \) and communication complexity of \( O(\kappa^2 e^{-2}) \). The numerical experiments on robust logistic regression show the empirical performance of \( \text{DREAM} \) 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...
Lesi Chen¹, Haishan Ye², Luo Luo³,*

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

References


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


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

(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]
A Some Useful Lemmas

We first provide some useful lemmas.

Lemma A.1. For any $a_1, \ldots, a_m \in \mathbb{R}^d$, 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 $z \in \mathbb{R}^{m \times d}$ and $\bar{z} = \frac{1}{m} 1^T z$, we have

\[
\| z - \bar{z} \| \leq \| z \|,
\]

Lemma A.3. Under Assumption 2.2, we have

\[
\| \nabla f(z) - \nabla f(z') \| \leq L \| z - z' \|
\]

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

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

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

\[
\bar{s}_0 = \frac{1}{m} 1^T s_0 = \frac{1}{m} 1^T g_0.
\]

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

\[
\bar{s}_{k+1} = \bar{s}_k + \bar{g}_{k+1} - \bar{g}_k
\]

\[
= \frac{1}{m} 1^T g(z_k) + \bar{g}_{k+1} - \frac{1}{m} 1^T g(z_k)
\]

\[
= \bar{g}_{k+1}
\]

\[
= \frac{1}{m} 1^T g_{k+1},
\]

which finished the proof.

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

\[
\left\| \bar{s}_t - \nabla f(\bar{z}_t) \right\|^2 \leq 2 \left\| \frac{1}{m} \sum_{i=1}^{m} (g_t(i) - \nabla f_t(z_t(i))) \right\|^2 + \frac{2L^2}{m} \| z_t - \bar{z}_t \|^2.
\]
We denote $\bar{y}_t$ for general convex and compact set $\mathcal{Y}$. We further illustrate this below.

We denote $y_{t+1/2} = y_t + \eta_t u_t$ by following Xian et al. (2021)'s notation. The procedure of DM-HSGD guarantees that $\bar{y}_{t+1}$ is the constrained reduced gradient

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

(11)

where $\bar{y}_t = \arg \max_{y \in \mathcal{Y}} f(\bar{x}_t, y)$ 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 $y_{t+1/2} = y_t + \eta_y u_t$ by following Xian et al. (2021)'s notation. The procedure of DM-HSGD guarantees that $\bar{y}_{t+1}$ is the constrained reduced gradient

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

(12)
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

$$\tilde{y}_{t+1} = \bar{\Pi}(y_{t+1/2}) = \tilde{y}_{t+1/2},$$  

which leads to

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

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

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

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

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 $y_{t+1}' = \Pi(y_t + \eta v_t)$. By Proposition 2.3 the update rules of $x_t, y_t$ means

$$\tilde{x}_{t+1} = \tilde{x}_t - \gamma \eta \tilde{u}_t \quad \text{and} \quad \tilde{y}_{t+1} = \tilde{y}_{t+1}' = \bar{\Pi}(y_t + \eta v_t).$$  

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

$$\mathbb{E}[P(\tilde{x}_{t+1})]$$  

$$\leq \mathbb{E} \left[ P(\tilde{x}_t) + \nabla P(\tilde{x}_t)^\top (\tilde{x}_{t+1} - \tilde{x}_t) + \frac{L_P}{2} ||\tilde{x}_{t+1} - \tilde{x}_t||^2 \right]$$  

$$= \mathbb{E} \left[ P(\tilde{x}_t) - \gamma \eta \nabla P(\tilde{x}_t)^\top \tilde{u}_t + \frac{L_P \gamma^2 \eta^2}{2} ||\tilde{u}_t||^2 \right]$$  

$$= \mathbb{E} \left[ P(\tilde{x}_t) - \frac{\gamma \eta}{2} ||\nabla P(\tilde{x}_t)||^2 - \left( \frac{\gamma \eta}{2} - \frac{L_P \gamma^2 \eta^2}{2} \right) ||\tilde{u}_t||^2 + \frac{\gamma \eta}{2} ||\nabla P(\tilde{x}_t) - \tilde{u}_t||^2 \right]$$  

$$\leq \mathbb{E} \left[ P(\tilde{x}_t) - \frac{\gamma \eta}{2} ||\nabla P(\tilde{x}_t)||^2 - \left( \frac{\gamma \eta}{2} - \frac{L_P \gamma^2 \eta^2}{2} \right) ||\tilde{u}_t||^2 \right]$$  

$$+ \mathbb{E} \left[ \frac{\gamma \eta}{2} \sum_{i=1}^m ||\nabla P(\tilde{x}_t) - \nabla_x f(\tilde{x}_t, y_t(i))||^2 + \frac{\gamma \eta}{2} \sum_{i=1}^m ||\tilde{u}_t - \nabla_x f(\tilde{x}_t, y_t(i))||^2 \right]$$  

$$\leq \mathbb{E} \left[ P(\tilde{x}_t) - \frac{\gamma \eta}{2} ||\nabla P(\tilde{x}_t)||^2 - \left( \frac{\gamma \eta}{2} - \frac{L_P \gamma^2 \eta^2}{2} \right) ||\tilde{u}_t||^2 \right]$$  

$$+ \mathbb{E} \left[ \frac{4 \kappa^2 \gamma \eta}{m} \sum_{i=1}^m ||G_y(\tilde{x}_t, y_t(i))||^2 + \frac{\gamma \eta}{2} \sum_{i=1}^m ||\tilde{u}_t - \nabla_x f(\tilde{x}_t, y_t(i))||^2 \right]$$  

$$\leq \mathbb{E} \left[ P(\tilde{x}_t) - \frac{\gamma \eta}{2} ||\nabla P(\tilde{x}_t)||^2 - \left( \frac{\gamma \eta}{2} - \frac{L_P \gamma^2 \eta^2}{2} \right) ||\tilde{u}_t||^2 \right]$$  

$$+ \mathbb{E} \left[ \frac{8 \kappa^2 \gamma \eta}{m} \sum_{i=1}^m ||v'(i)||^2 + \frac{8 \kappa^2 \gamma \eta}{m} \sum_{i=1}^m ||v_t(i) - \nabla_y f(\tilde{x}_t, y_t(i))||^2 + \frac{\gamma \eta}{2} \sum_{i=1}^m ||\tilde{u}_t - \nabla_x f(\tilde{x}_t, y_t(i))||^2 \right].$$
where $\mathbf{v}'_t(i) = (y'_t+1(i) - y_t(i))/\eta$. 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

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

Recall that $y'_t+1 = \Pi(y_t + \eta \mathbf{v}_t)$ and the first-order optimality of the projection, we have

$$
(y_t(i) + \eta \mathbf{v}_t(i) - y'_t+1(i))^T (y - y'_t+1) \leq 0
$$

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

$$
\mathbf{v}_t(i)^T (\bar{y}_t - y'_t+1(i))
\leq \frac{1}{\eta} (\bar{y}_t - y'_t+1(i))^T(y'_t+1(i) - y_t(i))
\leq \frac{1}{2\eta} \|\bar{y}_t - y_t(i)\|^2 - \frac{1}{2\eta} \|\bar{y}_t - y'_t+1(i)\|^2 - \frac{1}{2\eta} \|y'_t+1(i) - y_t(i)\|^2.
$$

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

$$
\mathbb{E}[-f(\bar{x}_t+1, y'_t+1(i))]
\leq \mathbb{E}[-f(\bar{x}_t+1, y_t(i)) - \nabla_y f(\bar{x}_t+1, y_t(i))^T (y'_t+1(i) - y_t(i)) + \frac{L}{2} \|y'_t+1(i) - y_t(i)\|^2]
\leq \mathbb{E}[-f(\bar{x}_t, y_t(i)) - \nabla_x f(\bar{x}_t, y_t(i))^T (\bar{x}_t+1 - \bar{x}_t) + \frac{L}{2} \|\bar{x}_t+1 - \bar{x}_t\|^2]
+ \mathbb{E}[\nabla_y f(\bar{x}_t+1, y_t(i))^T (y'_t+1(i) - y_t(i)) + \frac{L}{2} \|y'_t+1(i) - y_t(i)\|^2]
\leq \mathbb{E}[-f(\bar{x}_t, \bar{y}_t) + \nabla_y f(\bar{x}_t, y_t(i))^T (\bar{y}_t - y_t(i)) - \nabla_x f(\bar{x}_t, y_t(i))^T (\bar{x}_t+1 - \bar{x}_t) + \frac{L}{2} \|\bar{x}_t+1 - \bar{x}_t\|^2]
+ \mathbb{E}[\nabla_y f(\bar{x}_t, y_t(i))^T (\bar{y}_t - y_t(i)) + \frac{L}{2} \|\bar{y}_t - y_t(i)\|^2]
= \mathbb{E}[-f(\bar{x}_t, \bar{y}_t) + \nabla_y f(\bar{x}_t, y_t(i))^T (\bar{y}_t - y'_t+1(i)) - \nabla_x f(\bar{x}_t, y_t(i))^T (\bar{x}_t+1 - \bar{x}_t) + \frac{L}{2} \|\bar{x}_t+1 - \bar{x}_t\|^2]
+ \mathbb{E}[\nabla_y f(\bar{x}_t, y_t(i))^T (\bar{y}_t - y'_t+1(i)) + \frac{L}{2} \|\bar{y}_t - y'_t+1(i)\|^2]
\leq \mathbb{E}[-f(\bar{x}_t, \bar{y}_t) - \frac{1}{2\eta} \|y'_t+1(i) - y_t(i)\|^2 - \nabla_x f(\bar{x}_t, y_t(i))^T (\bar{x}_t+1 - \bar{x}_t) + \frac{L}{2} \|\bar{x}_t+1 - \bar{x}_t\|^2]
+ \mathbb{E}[\nabla_y f(\bar{x}_t, y_t(i))^T (\bar{y}_t - y_t(i)) + \frac{L}{2} \|\bar{y}_t - y_t(i)\|^2]
\leq \mathbb{E}[-f(\bar{x}_t, \bar{y}_t) - \left(\frac{\eta}{4} - \frac{\eta^2 L}{2}\right) \|\mathbf{v}'_t(i)\|^2 + \left(\frac{\gamma^2 \eta^2 L}{2} + \gamma^2 \eta^3 L^2 + \frac{3 \gamma^2 \eta}{2}\right) \|\bar{y}_t\|^2]
+ \mathbb{E}[\frac{\eta}{2} \|\nabla_x f(\bar{x}_t, y_t(i)) - \bar{u}_t\|^2 + \frac{\eta}{2} \|\nabla_y f(\bar{x}_t, y_t(i)) - \bar{v}_t(i)\|^2 + \frac{1}{2\eta} \|\bar{y}_t - y_t(i)\|^2].
$$
where $v_t^i = (y_{t+1}^i - y_t^i)/\eta$. 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

$$
(\nabla y f(\bar{x}_t, y_t(i)) - \nabla y f(\bar{x}_{t+1}, y_t(i)))^T (y_{t+1}^i - y_t^i) \\
\leq \frac{1}{4\eta} ||y_{t+1}^i - y_t^i||^2 + \eta ||\nabla y f(\bar{x}_{t+1}, y_t(i)) - \nabla y f(\bar{x}_t, y_t(i))|| \\
\leq \frac{1}{4\eta} ||y_{t+1}^i - y_t^i||^2 + \eta L^2 ||\bar{x}_{t+1} - \bar{x}_t||^2 \\
= \frac{\eta}{4} ||v_t^i||^2 + \gamma^2 \eta^3 L^2 ||\bar{u}_t||^2.
$$

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

$$
E[-f(\bar{x}_{t+1}, \bar{y}_{t+1})] \leq E \left[ -f(\bar{x}_t, \bar{y}_t) - \left( \frac{\eta}{4} - \frac{\eta^2 L}{2} \right) \frac{1}{m} \sum_{i=1}^{m} ||v_t^i||^2 + \left( \frac{\gamma \eta^2 L}{2} + \gamma^2 \eta^3 L^2 + \frac{3\gamma^2 \eta}{2} \right) ||\bar{u}_t||^2 \right] \\
+ \frac{1}{m} \sum_{i=1}^{m} E \left[ \frac{\eta}{2} ||\nabla_x f(\bar{x}_t, y_t(i)) - \bar{u}_t||^2 + \frac{\eta}{2} ||\nabla y f(\bar{x}_t, y_t(i)) - v_t^i||^2 + \frac{1}{2\eta} ||\bar{y}_t - y_t(i)||^2 \right].
$$

(18)

Note that $\bar{y}_{t+1} = \bar{y}_{t+1}$. Adding (15) multiplying $(1 + \alpha)$ along with (18) multiplying $\alpha$, we get

$$
E[P_{t+1}] \\
\leq P_t + (1 + \alpha)E \left[ -\frac{\gamma \eta}{2} ||\nabla P(\bar{x}_t)||^2 - \left( \frac{\gamma \eta^2 L}{2} - \frac{L \rho \gamma^2 \eta^2}{2} \right) ||\bar{u}_t||^2 \right] + (1 + \alpha) \times \\
E \left[ \frac{8\kappa^2 \gamma \eta}{m} \sum_{i=1}^{m} ||v_t^i||^2 + \frac{8\kappa^2 \gamma \eta}{m} \sum_{i=1}^{m} ||v_t^i - \nabla_y f(\bar{x}_t, y_t(i))||^2 + \frac{\gamma \eta}{m} \sum_{i=1}^{m} ||\bar{u}_t - \nabla_x f(\bar{x}_t, y_t(i))||^2 \right] \\
+ \alpha \left[ \left( \frac{\eta}{4} - \frac{\eta^2 L}{2} \right) \frac{1}{m} \sum_{i=1}^{m} ||v_t^i||^2 + \left( \frac{\gamma \eta^2 L}{2} + \gamma^2 \eta^3 L^2 + \frac{3\gamma^2 \eta}{2} \right) ||\bar{u}_t||^2 \right] \\
+ \frac{\alpha}{m} \sum_{i=1}^{m} E \left[ \frac{\eta}{2} ||\nabla_x f(\bar{x}_t, y_t(i)) - \bar{u}_t||^2 + \frac{\eta}{2} ||\nabla y f(\bar{x}_t, y_t(i)) - v_t^i||^2 + \frac{1}{2\eta} ||\bar{y}_t - y_t(i)||^2 \right].
$$

Rearranging the above result leads to

$$
E[\Psi_{t+1}] \leq E \left[ \Psi_t - \frac{(1 + \alpha) \gamma \eta}{2} ||\nabla P(\bar{x}_t)||^2 \\
- \left( (1 + \alpha) \left( \frac{\gamma \eta^2 L}{2} - \frac{L \rho \gamma^2 \eta^2}{2} \right) - \alpha \left( \frac{L \gamma^2 \eta^2}{2} + \gamma^2 \eta^3 L^2 + \frac{3\gamma^2 \eta}{2} \right) \right) E[||\bar{u}_t||^2] \\
- \alpha \left( \frac{\eta}{4} - \frac{L \rho \eta^2}{2} \right) - (1 + \alpha)8\kappa^2 \gamma \eta \frac{1}{m} \sum_{i=1}^{m} E[||v_t^i||^2] \\
+ \left( (1 + \alpha) \gamma + \frac{\alpha}{2} \right) \frac{\eta}{2} m \sum_{i=1}^{m} E[||\bar{u}_t - \nabla_x f(\bar{x}_t, y_t(i))||^2] \\
+ \left( (1 + \alpha)8\kappa^2 \gamma + \frac{\alpha}{2} \right) \frac{\eta}{2} m \sum_{i=1}^{m} E[||v_t^i - \nabla_y f(\bar{x}_t, y_t(i))||^2] \\
+ \frac{\alpha}{2\eta m} \sum_{i=1}^{m} E[||\bar{y}_t - y_t(i)||^2].
$$
Also, the fact where we use the definition of \( \gamma \) in (\ref{eq:gamma}) means

\[
\alpha \left( \frac{\eta}{4} - \frac{L \eta^3}{2} \right) - (1 + \alpha)8\kappa^2\gamma \eta \geq \frac{\alpha \eta}{16}
\]
as well as

\[
(1 + \alpha) \left( \frac{\gamma \eta}{2} - \frac{L \rho \gamma^2 \eta^2}{2} \right) - \alpha \left( \frac{L \gamma^2 \eta^2}{2} + \gamma^2 \eta^3 L + \frac{3 \gamma^2 \eta}{2} \right) \geq \frac{\gamma \eta}{4}.
\]

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

\[
8(1 + \alpha)\kappa^2\gamma + \frac{\alpha}{2} \leq \alpha \quad \text{and} \quad (1 + \alpha)8\kappa^2\gamma + \frac{\alpha}{2} \leq \alpha.
\]

Therefore, we obtain the optimization bound as

\[
E[\Psi_{t+1}] \leq E \left[ \Psi_t - \frac{\gamma \eta}{2} \| \nabla P(x_t) \|^2 - \frac{\gamma \eta}{4} \| \bar{u}_t \|^2 - \frac{\alpha \eta}{16m} \sum_{i=1}^{m} \| v_i^t(i) \|^2 \right] + E \left[ \frac{\alpha \eta}{m} \sum_{i=1}^{m} \| \bar{u}_t - \nabla_x f(\bar{x}_t, y_t(i)) \|^2 + \frac{\alpha \eta}{m} \sum_{i=1}^{m} \| v_i(i) - \nabla_y f(\bar{x}_t, y_t(i)) \|^2 + \frac{\alpha}{2 \eta m} \sum_{i=1}^{m} \| \bar{y}_t - y_t(i) \|^2 \right].
\]

(19)

Note that it holds that

\[
\sum_{i=1}^{m} \| v_t(i) - \nabla_y f(\bar{x}_t, y_t(i)) \|^2 \leq 3m \| \bar{v}_t - \nabla_y f(\bar{x}_t, \bar{y}_t) \|^2 + 3 \| v_t - 1 \bar{v}_t \|^2 + 3L^2 \| y_t - 1 \bar{y}_t \|^2
\]

and

\[
\sum_{i=1}^{m} \| \bar{u}_t - \nabla_x f(\bar{x}_t, y_t(i)) \|^2 \leq 2m \| \bar{u}_t - \nabla_x f(\bar{x}_t, \bar{y}_t) \|^2 + 2L^2 \| y_t - 1 \bar{y}_t \|^2,
\]

where we use the \( L \)-smoothness and Young’s inequality. Then we combine Lemma \ref{lem:main} to get

\[
\frac{\alpha \eta}{m} \sum_{i=1}^{m} \| \bar{u}_t - \nabla_x f(\bar{x}_t, y_t(i)) \|^2 + \frac{\alpha \eta}{m} \sum_{i=1}^{m} \| v_t(i) - \nabla_y f(\bar{x}_t, y_t(i)) \|^2
\]

\[
\leq 6\alpha \eta L^2 \| y_t - 1 \bar{y}_t \|^2 + \frac{3\alpha \eta}{m} \| s_t - \nabla f(\bar{x}_t, \bar{y}_t) \|^2 + \frac{3\alpha \eta}{m} \| v_t - 1 \bar{v}_t \|^2
\]

\[
\leq \frac{\alpha}{2 \eta m} \| y_t - 1 \bar{y}_t \|^2 + 3\alpha \eta \| s_t - \nabla f(\bar{x}_t, \bar{y}_t) \|^2 + \frac{3\alpha \eta}{m} \| v_t - 1 \bar{v}_t \|^2,
\]

(20)

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

\[
\frac{1}{m} \sum_{i=1}^{m} \| v_i^t(i) \|^2 \geq \| \bar{v}_t \|^2.
\]

Hence, we have

\[
E[\Psi_{t+1}] \leq E \left[ \Psi_t - \frac{\gamma \eta}{2} \| \nabla P(x_t) \|^2 - \frac{\gamma \eta}{4} \| \bar{u}_t \|^2 - \frac{\alpha \eta}{16} \| \bar{v}_t \|^2 \right] + E \left[ 3\alpha \eta \| s_t - \nabla_x f(\bar{x}_t, \bar{y}_t) \|^2 + \frac{3\alpha \eta}{m} \| v_t - 1 \bar{v}_t \|^2 + \frac{\alpha}{\eta m} \| y_t - \bar{y}_t \|^2 \right].
\]

Combining with Lemma \ref{lem:main} and using \( \| y_t - 1 \bar{y}_t \| \leq \| z_t - 1 \bar{s}_t \| \) and \( \eta \leq 1/(4L) \), we obtain

\[
E[\Psi_{t+1}] \leq E \left[ \Psi_t - \frac{\gamma \eta}{2} \| \nabla P(x_t) \|^2 - \frac{\gamma \eta}{4} \| \bar{u}_t \|^2 - \frac{\alpha \eta}{16} \| \bar{v}_t \|^2 \right] + E \left[ \frac{3\alpha \eta}{m} \| v_t - 1 \bar{v}_t \|^2 + 6\alpha \left\| \frac{1}{m} \sum_{i=1}^{m} (g_t(i) - \nabla f_i(z_t(i))) \right\|^2 + \frac{2\alpha}{\eta m} \| z_t - \bar{z}_t \|^2 \right].
\]
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
\[
E[\Psi_{t+1}] \leq E \left[ \Psi_t - \frac{\gamma \eta}{2} \| \nabla P(\bar{x}_t) \|^2 - \frac{1}{8 \gamma \eta} \| \bar{x}_{t+1} - \bar{x}_t \|^2 - \frac{\alpha}{16 \eta} \| \bar{z}_{t+1} - \bar{z}_t \|^2 \right] \\
+ E \left[ \frac{3 \eta}{m} \| v_t - 1 \bar{v}_t \|^2 + 6 \alpha \eta \right] \left( \frac{1}{m} \sum_{i=1}^m (g_i(i) - \nabla f_i(z_i(i))) \right)^2 + \frac{2 \alpha}{\eta m} \| z_t - 1 \bar{z}_t \|^2.
\]
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
\[
\| x_{t+1} - 1 \bar{x}_{t+1} \| \\
\leq \rho \| x_t - \gamma \eta u_t - 1 (\bar{x}_t - \eta \bar{u}_t) \| \\
\leq \rho \left( \| x_t - 1 \bar{x}_t \| + \gamma \eta \| u_t - 1 \bar{u}_t \| \right),
\]
where the last step is due to triangle inequality. Similarly, we define the notation $\Pi(\cdot) = \frac{1}{m} 11^T(\cdot)$ for convenience. Then for variable $y$, we can verify that
\[
\| y_{t+1} - 1 \bar{y}_{t+1} \| \\
\leq \rho \left( \| y_t + \eta v_t - \frac{1}{m} 11^T (y_t + \eta v_t) \| \right) \\
\leq \rho \left( \| y_t + \eta v_t - \Pi (1 \bar{y}_t + \eta 1 \bar{v}_t) \| + \rho \left( \| \Pi (1 \bar{y}_t + \eta 1 \bar{v}_t) - 1 \Pi (y_t + \eta v_t) \| \right) \\
\leq 2 \rho \| y_t + \eta v_t - 1 (\bar{y}_t + \eta \bar{v}_t) \| \\
\leq 2 \rho (\| y_t - 1 \bar{y}_t \| + \eta \| v_t - 1 \bar{v}_t \|),
\]
where in the third inequality we use the non-expansiveness of projection and Lemma 11 in Ye et al. (2020b), i.e.
\[
\left\| 1 \Pi(x) - \Pi(1 \bar{x}) \right\| \leq \| x - 1 \bar{x} \|.
\]
Consequently, we use Young’s inequality together with $\gamma \in (0, 1]$ to obtain
\[
\| z_{t+1} - 1 \bar{z}_{t+1} \|^2 = \| x_{t+1} - 1 \bar{x}_{t+1} \|^2 + \| y_{t+1} - 1 \bar{y}_{t+1} \|^2 \\
\leq 8 \rho^2 \| y_t - 1 \bar{y}_t \|^2 + 8 \rho^2 \eta^2 \| v_t - 1 \bar{v}_t \|^2 + 2 \rho^2 \| x_t - 1 \bar{x}_t \|^2 + 2 \rho^2 \eta^2 \| y_t - 1 \bar{y}_t \|^2 \\
\leq 8 \rho^2 \| z_t - 1 \bar{z}_t \|^2 + 8 \rho^2 \eta^2 \| s_t - 1 \bar{s}_t \|^2.
\]
Furthermore, if $24 \rho^2 \leq 1$, we have
\[
\| z_{t+1} - z_t \|^2 \leq 3 \| z_{t+1} - 1 \bar{z}_{t+1} \|^2 + 3 \| z_t - 1 \bar{z}_t \|^2 + 3 \| 1 \bar{z}_t - 1 \bar{z}_{t+1} \|^2 \\
\leq (24 \rho^2 + 3) \| z_t - 1 \bar{z}_t \|^2 + 24 \rho^2 \eta^2 \| s_t - 1 \bar{s}_t \|^2 + 3 \| 1 \bar{z}_t - 1 \bar{z}_{t+1} \|^2 \\
\leq 4 \| z_t - 1 \bar{z}_t \|^2 + \eta^2 \| s_t - 1 \bar{s}_t \|^2 + 3m \| \bar{z}_t - \bar{z}_{t+1} \|^2.
\]
We let $\rho_t = \rho'$ for $\zeta_t = 1$ and $\rho_t = \rho$ otherwise. The update of $g_{t+1}(i)$ means
\[
\mathbb{E}\left[\rho_t^2 \|g_{t+1}(i) - g_t(i)\|^2\right] = \rho^2 p \mathbb{E}\left[\frac{1}{b'} \sum_{\xi_t,i \in S_t(i)} \nabla F_i(z_{t+1}(i); \xi_t,i) - g_t(i)\right]^2 + \frac{\rho^2(1-p)}{bq} \mathbb{E}\|\nabla F_i(z_{t+1}(i); \xi_t,i) - \nabla F_i(z_t(i); \xi_t,1)\|^2
\]
\[
\leq 3\rho^2 p \mathbb{E}\left[\frac{1}{b'} \sum_{\xi_t,i \in S_t(i)} \nabla F_i(z_{t+1}(i); \xi_t,i) - \nabla f_i(z_{t+1}(i))\right]^2 + 3\rho^2 p \mathbb{E}\|\nabla f_i(z_{t+1}(i)) - \nabla f_i(z_t(i))\|^2
\]
\[
+ 3\rho^2 p \mathbb{E}\|\nabla f_i(z_t(i)) - g_t(i)\|^2 + \frac{\rho^2(1-p)L^2}{bq} \mathbb{E}\|z_{t+1}(i) - z_t(i)\|^2
\]
\[
\leq \frac{3\rho^2 \sigma^2}{b'} [b' < n] + 3\rho^2 p L^2 \mathbb{E}\|z_{t+1}(i) - z_t(i)\|^2 + 3\rho^2 p \mathbb{E}\|\nabla f_i(z_t(i)) - g_t(i)\|^2 + \left(1 - \frac{p}{bq} + 3p\right) \rho^2 L^2 \mathbb{E}\|z_{t+1}(i) - z_t(i)\|^2,
\]
where the first inequality is based on update rules and Assumption 2.2, 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
\[
\mathbb{E}\left[\rho_t^2 \|g_{t+1} - g_t\|^2\right] \leq \frac{3\rho^2 \sigma^2}{b'} [b' < n] + 3\rho^2 p \mathbb{E}\|\nabla f(z_t) - g_t\|^2 + \left(1 - \frac{p}{bq} + 3p\right) \rho^2 L^2 \mathbb{E}\|z_{t+1} - z_t\|^2.
\]
Let $c = \max\{1/(bq), 1\}$ and plug it into (22), then
\[
\mathbb{E}\left[\rho_t^2 \|g_{t+1} - g_t\|^2\right] \leq \frac{3\rho^2 cm^2 \sigma^2}{b'} [b' < n] + 3\rho^2 p \mathbb{E}\|\nabla f(z_t) - g_t\|^2 + c \rho^2 L^2 \mathbb{E}\|z_{t+1} - z_t\|^2
\]
\[
\leq \frac{3\rho^2 cm^2 \sigma^2}{b'} [b' < n] + 3\rho^2 p \mathbb{E}\|\nabla f(z_t) - g_t\|^2 + 16 c \rho^2 L^2 \mathbb{E}\|z_t - \bar{z}_t\|^2 + 4 c \rho^2 L^2 \mathbb{E}\|s_t - \bar{s}_t\|^2 + 12 c \rho^2 m L^2 \mathbb{E}\|z_{t+1} - \bar{z}_t\|^2.
\]
Furthermore, we have
\[
\|s_{t+1} - \bar{s}_{t+1}\|
\leq \rho_t \left\|s_t + g_{t+1} - g_t - \frac{1}{m} 11^T (s_t + g_{t+1} - g_t)\right\|
\leq \rho_t \left\|s_t - \bar{s}_t\right\| + \rho_t \left\|g_{t+1} - g_t - \frac{1}{m} 11^T (g_{t+1} - g_t)\right\|
\leq \rho_t \left\|s_t - \bar{s}_t\right\| + \rho_t \left\|g_{t+1} - g_t\right\|,
\]
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 $\eta \leq 1/(4L)$, we have
\[
\eta^2 \mathbb{E}\|s_{t+1} - \bar{s}_{t+1}\|^2
\leq 2 \rho^2 \eta^2 \|s_t - \bar{s}_t\|^2 + 2 \eta^2 \mathbb{E}\|g_{t+1} - g_t\|^2
\leq 4 c \rho^2 \eta^2 \mathbb{E}\|s_t - \bar{s}_t\|^2 + 4 c \rho^2 \mathbb{E}\|z_t - \bar{z}_t\|^2
\]
\[
+ 6 \rho^2 \eta^2 \mathbb{E}\|\nabla f(z_t) - g_t\|^2 + 2 c \rho^2 m \mathbb{E}\|\bar{z}_{t+1} - \bar{z}_t\|^2 + \frac{6 \rho^2 m \rho^2 \sigma^2}{b'} \left\|b' < n\right\|.
\]
Combining (25) and (21), we obtain
\[
\mathbb{E}\left[\|z_{t+1} - \bar{z}_{t+1}\|^2 + \eta^2 \|s_{t+1} - \bar{s}_{t+1}\|^2\right]
\leq 12 c \rho^2 \mathbb{E}\|z_t - \bar{z}_t\|^2 + \eta^2 \|s_t - \bar{s}_t\|^2 + 6 \rho^2 \eta^2 \mathbb{E}\|\nabla f(z_t) - g_t\|^2 + 2 c \rho^2 m \mathbb{E}\|\bar{z}_{t+1} - \bar{z}_t\|^2 + \frac{6 \rho^2 m \rho^2 \sigma^2}{b'} \left\|b' < n\right\|,
\]
which finishes our proof. \qed
The Proof of Lemma 3.3

Proof. The update of $g_{t+1}(i)$ means

$$
E \left\| g_{t+1}(i) - \nabla f_i(z_{t+1}(i)) \right\|^2
= p E \left\| \frac{1}{b} \sum_{\xi, j \in S_t(i)} \nabla F_i(z_{t+1}(i); \xi, j) - \nabla f_i(z_{t+1}(i)) \right\|^2
+ (1 - p) E \left\| g_t(i) + \frac{\omega(i)}{bq} \sum_{\xi, j \in S_t(i)} (\nabla F_i(z_{t+1}(i); \xi, j) - \nabla f_i(z_t(i); \xi, j) - \nabla f_i(z_{t+1}(i)) + \nabla f_i(z_t(i))) \right\|^2
\leq \frac{m \sigma^2}{b'} I[b' < n] + (1 - p) E \left\| g_t(i) - \nabla f_i(z_t(i)) \right\|^2
+ \frac{\omega(i)}{bq} \sum_{\xi, j \in S_t(i)} (\nabla F_i(z_{t+1}(i); \xi, j) - \nabla F_i(z_t(i); \xi, j) - \nabla f_i(z_{t+1}(i)) + \nabla f_i(z_t(i))) \right\|^2
\leq \frac{m \sigma^2}{b'} I[b' < n] + (1 - p) E \left\| g_t(i) - \nabla f_i(z_t(i)) \right\|^2
+ \frac{1 - p}{bq} E \left\| \nabla F_i(z_{t+1}(i); \xi, j) - \nabla F_i(z_t(i); \xi, j) \right\|^2
\leq \frac{m \sigma^2}{b'} I[b' < n] + (1 - p) E \left\| g_t(i) - \nabla f_i(z_t(i)) \right\|^2
+ \frac{(1 - p) L^2}{bq} E \| z_{t+1} - z_t \|^2,
$$

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 $i = 1, \ldots, m$ for above result and using (22), we obtain

$$
E \left\| g_{t+1} - \nabla f(z_{t+1}) \right\|^2
\leq \frac{m \sigma^2}{b'} I[b' < n] + (1 - p) E \| g_t - \nabla f(z_t) \|^2
+ \frac{(1 - p) L^2}{bq} E \| z_{t+1} - z_t \|^2
\leq \frac{m \sigma^2}{b'} I[b' < n] + (1 - p) E \| g_t - \nabla f(z_t) \|^2
+ \frac{(1 - p) L^2}{bq} E \| z_t - \bar{z}_t \|^2 + \eta^2 \| s_t - \bar{s}_t \|^2 + 3m \| \bar{z}_{t+1} - \bar{z}_t \|^2,
$$

which is the variance bound as claimed by the definition of $V_t$. \hfill \Box

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

$$
E \left\| \sum_{i=1}^m a_i \right\|^2 = \sum_{i=1}^m E \| a_i \|^2,
$$

where each $a_1, \ldots, a_m$ are independent with zero mean.
The Proof of Theorem 3.1

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

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

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

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

Telescoping for \( t = 0, 1, \ldots, T - 1 \), we obtain

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

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

\[
E \|\nabla P(x_{out})\|^2 = \frac{1}{mT} \sum_{t=0}^{T-1} \sum_{i=1}^{m} \|\nabla P(x_t(i))\|^2 \leq \frac{2}{mT} \sum_{t=0}^{T-1} \sum_{i=1}^{m} \left( \|\nabla P(\bar{x}_t)\|^2 + \|\nabla P(x_t(i)) - \nabla P(\bar{x}_t)\|^2 \right) \leq \frac{2}{mT} \sum_{t=0}^{T-1} \sum_{i=1}^{m} \left( \|\nabla P(\bar{x}_t)\|^2 + L^2 \|x_t(i) - \bar{x}_t\|^2 \right) = \frac{2}{T} \sum_{t=0}^{T-1} \|\nabla P(\bar{x}_t)\|^2 + \frac{2L^2}{mT} \sum_{t=0}^{T-1} \|x_t - 1\bar{x}_t\|^2 \leq \frac{2}{T} \sum_{t=0}^{T-1} \|\nabla P(\bar{x}_t)\|^2 + \frac{2L^2}{mT} \sum_{t=0}^{T-1} C_t, 
\]

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

\[
E[\|\nabla P(x_{out})\|^2] \leq \frac{4}{\gamma T} \Phi_0 - \frac{16\alpha}{\gamma^2} - 2L^2 \leq \frac{2}{mT} \sum_{t=0}^{T-1} C_t + \frac{12\sigma^2}{\gamma mb'} \|b' < n\|. 
\]
Recall the choice of $\gamma$ in [5]; $\eta \leq 1/(4L)$ and $pT \geq 2$. Hence, we have

$$\mathbb{E}\|\nabla P(x_{out})\|^2 \leq \frac{4}{\gamma\eta T} \Phi_0 + \frac{12\sigma^2}{\gamma mb'}[b' < n]$$

$$= \frac{4}{\gamma\eta T} \left( \Psi_0 + \frac{\eta}{mp} V_0 + \frac{\eta}{p} U_0 + \frac{\eta}{m} C_0 \right) + \frac{12\sigma^2}{\gamma mb'}[b' < n]$$

$$= \frac{4}{\gamma\eta T} \left( \Psi_0 + \frac{2\sigma^2}{mb'p} [b' < n] + \frac{\eta}{m} \|s_0 - 1s_0\|^2 \right) + \frac{12\sigma^2}{\gamma mb'}[b' < n]$$

$$\leq \frac{4}{\gamma\eta T} \left( \Psi_0 + \frac{2\sigma^2}{mb'p} [b' < n] + \frac{\eta\rho^2}{m} \|g_0 - 1g_0\|^2 \right) + \frac{12\sigma^2}{\gamma mb'}[b' < n]$$

$$\leq \frac{8}{\gamma\eta T} \Psi_0 + \frac{16\sigma^2}{\gamma mb'} [b' < n] + \frac{4\rho^2}{m} \|g_0 - 1g_0\|^2 \frac{1}{\gamma m}.$$ 

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

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

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

Plug in the choice of $b', q$ yields the SFO complexity as claimed. Next, recalling the definition of $\rho, \rho', \rho_0$ in [6], we know the total number of communication rounds is

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

\[\Box\]

**H Future Directions and Subsequent Works**

A future direction is to establish decentralized stochastic algorithms with better dependency on the condition number $\kappa$ 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; 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 $Y$). 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 $\mathcal{O}(N\kappa^2\delta\epsilon^{-2})$, 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–Lojasiewicz (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 $\mathcal{O}(\kappa^4\epsilon^{-3})$ is worse than our $\mathcal{O}(\kappa^2\epsilon^{-2})$. 

Lesi Chen¹, Haishan Ye², Luo Luo³,*