Shuffling Gradient-Based Methods for Nonconvex-Concave Minimax Optimization

Quoc Tran-Dinh
Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill
quoctd@email.unc.edu

Trang H. Tran
School of OR and Information Engineering
Cornell University, Ithaca, NY
htt27@cornell.edu

Lam M. Nguyen
IBM Research, Thomas J. Watson Research Center
Yorktown Heights, NY
LamNguyen.MLTD@ibm.com

Abstract

This paper aims at developing novel shuffling gradient-based methods for tackling two classes of minimax problems: *nonconvex-linear* and *nonconvex-strongly concave* settings. The first algorithm addresses the nonconvex-linear minimax model and achieves the state-of-the-art oracle complexity typically observed in nonconvex optimization. It also employs a new shuffling estimator for the "hyper-gradient", departing from standard shuffling techniques in optimization. The second method consists of two variants: *semi-shuffling* and *full-shuffling* schemes. These variants tackle the nonconvex-strongly concave minimax setting. We establish their oracle complexity bounds under standard assumptions, which, to our best knowledge, are the best-known for this specific setting. Numerical examples demonstrate the performance of our algorithms and compare them with two other methods. Our results show that the new methods achieve comparable performance with SGD, supporting the potential of incorporating shuffling strategies into minimax algorithms.

1 Introduction

Minimax problems arise in various applications across generative machine learning, game theory, robust optimization, online learning, and reinforcement learning (e.g., [1, 2, 3, 5, 12, 13, 17, 19, 21, 25, 35, 40]). These models often involve stochastic settings or large finite-sum objective functions. To tackle these problems, existing methods frequently adapt stochastic gradient descent (SGD) principles to develop algorithms for solving the underlying minimax problems [4, 13]. For instance, in generative adversarial networks (GANs), early algorithms employed stochastic gradient descent-ascent methods where two routines, each using an SGD loop, ran iteratively [13]. However, practical implementations of SGD often incorporate shuffling strategies, as seen in popular deep learning libraries like TensorFlow and PyTorch. This has motivated recent research on developing shuffling techniques specifically for optimization algorithms [4, 5, 8, 16, 26, 32, 38]. Our work builds upon this trend by developing shuffling methods for two specific classes of minimax problems.

Problem statement. In this paper, we study the following minimax optimization problem:

$$\min_{w\in\mathbb{R}^p}\max_{u\in\mathbb{R}^q}\Big\{\mathcal{L}(w,u) := f(w) + \mathcal{H}(w,u) - h(u) \equiv f(w) + \tfrac{1}{n}\sum_{i=1}^{n}\mathcal{H}_i(w,u) - h(u)\Big\}, \tag{1}$$

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

where $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed, and convex function, $\mathcal{H}_i : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ are smooth for all $i \in [n] := \{1, 2, \cdots, n\}$, and $h : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ is also a proper, closed, and convex function. In this paper, we will focus on two classes of problems in (1), overlapped to each other.

(NL) $\mathcal{H}_i$ is nonconvex in w and linear in u as $\mathcal{H}_i(w, u) := \langle F_i(w), Ku \rangle$ for a given function $F_i : \mathbb{R}^p \to \mathbb{R}^m$ and a matrix $K \in \mathbb{R}^{q \times m}$ for all $i \in [n]$ and $(w, u) \in \mathrm{dom}\,(\mathcal{L})$.

(NC) $\mathcal{H}_i$ is nonconvex in w and $\mathcal{H}_i(w, \cdot) - h(\cdot)$ is strongly concave in u for all $(w, u) \in \mathrm{dom}\,(\mathcal{L})$.

Although (NC) looks more general than (NL), both cases can be overlapped, but one is not a special case of the other. Under these two settings, our approach will rely on a *bilevel optimization* approach, where the lower-level problem is to solve $\max_u \mathcal{L}(w, u)$, while the upper-level one is $\min_w \mathcal{L}(w, u)$.

Challenges. The setting (NL) is a special case of stochastic nonconvex-concave minimax problems because the objective term $\mathcal{H}(w, u) := \langle F(w), Ku \rangle$ is linear in u. It is equivalent to the compositional model (CO) described below. However, if h is only merely convex and not strongly convex (e.g., the indicator of a standard simplex), then Φ_0 in (CO) becomes nonsmooth regardless of F's properties. This presents our first challenge. A natural approach to address this issue, as discussed in Section 2, is to smooth Φ_0. The second challenge arises from the composition between the outer function h^* and the finite sum $F(\cdot)$ in (CO). Unlike standard finite-sum optimization, this composition prevents any direct use of existing techniques, requiring a novel approach for algorithmic development and analysis. The third challenge involves unbiased estimators for gradients or "hyper-gradients" in minimax problems. Most existing methods rely on unbiased estimators for objective gradients, with limited work exploring biased estimators. While biased estimators can be used, they require variance reduction properties (see, e.g., [10]). The setting (NC) faces the same second and third challenges as the setting (NL). Additionally, when reformulating it as a minimization problem using a bilevel optimization approach (3), constructing a shuffling estimator for the "hyper-gradient" $\nabla\Phi_0$ becomes unclear. This requires solving the lower-level maximization problem (2). Therefore, it remains an open question whether shuffling gradient-type methods can be extended to this bilevel optimization approach to address (1). In this paper, we address the following research question:

Can we efficiently develop shuffling gradient methods to solve (1) *for both* (NL) *and* (NC) *settings?*

Our attempt to tackle this question leads to a novel way of constructing shuffling estimators for the hyper-gradient $\nabla\Phi_0$ or its smoothed counterpart. This allows us to develop two shuffling gradient-based algorithms with rigorous theoretical guarantees on oracle complexity, matching state-of-the-art complexity results in shuffling-type algorithms for nonconvex optimization.

Related work. Shuffling optimization algorithms have gained significant attention in optimization and machine communities, demonstrating advantages over standard SGDs, see, e.g., [4, 5, 8, 16, 26, 32, 38]. Nevertheless, applying these techniques to minimax problems like (1) remains challenging, with limited existing literature (e.g., [3, 8, 11]). Das *et al.* in [8] explored a specific case of (1) without nonsmooth terms f and h, assuming strong monotonicity and L-Lipschitz continuity of the gradient $\nabla\mathcal{H} := [\nabla_w\mathcal{H}, -\nabla_u\mathcal{H}]$ of the joint objective $\mathcal{H}$. Their algorithm simplifies to a shuffling variant of fixed-point iteration or a gradient descent-ascent scheme, not applicable to our settings. Cho and Yun in [3] built upon [8] by relaxing the strong monotonicity to Polyak-Łojasiewicz (PŁ) conditions. This work is perhaps the most closely related one to our algorithm, Algorithm 2, for the (NC) setting. Note that the method in [3] exploits Nash's equilibrium perspective with a simultaneous update, which is different from our alternative update. Moreover, [3] only considers the noncomposite case with $f = 0$ and $h = 0$. Though we only focus on a nonconvex-strongly-concave setting (NC), our results here can be extended to the PŁ condition as in [3]. Very recently, Konstantinos *et al.* in [11] introduced shuffling extragradient methods for variational inequalities, which encompass convex-concave minimax problems as a special case. However, this also falls outside the scope of our work due to the nonconvexity of (1) in w. Again, all the existing works in [3, 8, 11] utilize a Nash's equilibrium perspective, while ours leverages a bilevel optimization technique. Besides, in contrast to our sampling-without-replacement approach, stochastic and randomized methods (i.e. using i.i.d. sampling strategies) have been extensively studied for minimax problems, see, e.g., [9, 14, 15, 18, 22, 23, 31, 37, 42]. A comprehensive comparison can be found, e.g., in [3].

Contribution. Our main contribution can be summarized as follows.

(a) For setting (NL), we suggest to reformulate (1) into a compositional minimization and exploit a smoothing technique to treat this reformulation. We propose a new way of constructing shuffling estimators for the "hyper-gradient" $\nabla\Phi_\gamma$ (cf. (10)) and establish their properties.

(b) We propose a novel shuffling gradient-based algorithm (*cf.* Algorithm 1) to approximate an ϵ-KKT point of (1) for the setting (NL). Our method requires $\mathcal{O}(n\epsilon^{-3})$ evaluations of F_i and ∇F_i under the strong convexity of h, and $\mathcal{O}(n\epsilon^{-7/2})$ evaluations of F_i and ∇F_i without the strong convexity of h, for a desired accuracy $\epsilon > 0$.

(c) For setting (NC), we develop two variants of the shuffling gradient method: *semi-shuffling* and *full-shuffling* schemes (*cf.* Algorithm 2). The semi-shuffling variant combines both gradient ascent and shuffling gradient methods to construct a new algorithm, which requires $\mathcal{O}(n\epsilon^{-3})$ evaluations of both $\nabla_w \mathcal{H}_i$ and $\nabla_u \mathcal{H}_i$. The full-shuffling scheme allows to perform both shuffling schemes on the maximization and the minimization alternatively, requiring either $\mathcal{O}(n\epsilon^{-3})$ or $\mathcal{O}(n\epsilon^{-4})$ evaluations of $\nabla_u \mathcal{H}_i$ depending on our assumptions, while maintaining $\mathcal{O}(n\epsilon^{-3})$ evaluations of $\nabla_w \mathcal{H}_i$ for a given desired accuracy $\epsilon > 0$.

If a random shuffling strategy is used in our algorithms, then the oracle complexity in all the cases presented above is improved by a factor of $\sqrt{n}$. Our settings (NL) and (NC) of (1) are different from existing works [3, 8, 11], as we work with general nonconvexity in w, and linearity or [strong] concavity in u, and both f and h are possibly nonsmooth. Our algorithms are not reduced or similar to existing shuffling methods for optimization, but we use shuffling strategies to form estimators for the hyper-gradient $\nabla \Phi_0$ in (5). The oracle complexity in both settings (NL) and (NC) is similar to the ones in nonconvex optimization and in a special case of (1) from [3] (up to a constant factor).

Paper outline. The rest of this paper is organized as follows. Section 2 presents our bilevel optimization approach to (1) and recalls necessary preliminary results. Section 3 develops our shuffling algorithm to solve the setting (NL) of (1) and establishes its convergence. Section 4 proposes new shuffling methods to solve the setting (NC) and investigates their convergence. Section 5 presents numerical experiments, while technical proofs and supporting results are deferred to Supp. Docs.

Notations. For a function f, we use $\mathrm{dom}\,(f)$ to denote its effective domain, and ∇f for its gradient or Jacobian. If f is convex, then ∇f denotes a subgradient, ∂f is its subdifferential, and prox_f is its proximal operator. We use $\mathcal{F}_t$ to denote $\sigma(w_0, w_1, \cdots, w_t)$, a σ-algebra generated by random vectors $w_0, w_1, \cdots, w_t$, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$ is a conditional expectation, and $\mathbb{E}[\cdot]$ is the full expectation. As usual, $\mathcal{O}(\cdot)$ denotes Big-O notation in the theory of algorithm complexity.

2 Bilevel Optimization Approach and Preliminary Results

Our approach relies on a bilevel optimization technique [9] in contrast to Nash's game viewpoint [24], which treats the maximization as a lower level and the minimization as an upper level problem.

2.1 Bilevel optimization approach

The minimax model (1) is split into a *lower-level* (*i.e. a follower*) *maximization problem* of the form:

$$\begin{aligned} \Phi_0(w) &:= \max_{u \in \mathbb{R}^q} \big\{ \mathcal{H}(w,u) - h(u) \equiv \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathcal{H}_i(w,u) - h(u) \big\}, \\ u_0^*(w) &:= \operatorname*{argmax}_{u \in \mathbb{R}^q} \big\{ \mathcal{H}(w,u) - h(u) \equiv \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathcal{H}_i(w,u) - h(u) \big\}. \end{aligned} \tag{2}$$

For Φ_0 defined by (2), then the *upper-level* (*i.e. the leader*) *minimization problem* can be written as

$$\Psi_0^\star := \min_{w \in \mathbb{R}^p} \Big\{ \Psi_0(w) := \Phi_0(w) + f(w) \Big\}. \tag{3}$$

Clearly, this approach is sequential, and only works if Φ_0 is well-defined, i.e. (2) is globally solvable. Hence, the concavity of $\mathcal{H}(w,\cdot) - h(\cdot)$ w.r.t. to u is crucial for this approach as stated below. However, this assumption can be relaxed to a global solvability of (2) combined with a PŁ condition as in [3].

Assumption 1 (Basic)**.** *Problems* (1) *and* (3) *satisfy the following assumptions for all* $i \in [n]$*:*

(a) $\Psi_0^\star := \inf_w \Psi_0(w) > -\infty$.
(b) $\mathcal{H}_i$ *is differentiable w.r.t.* $(w,u) \in \mathrm{dom}\,(\mathcal{L})$ *and* $\mathcal{H}_i(w,\cdot)$ *is concave in* u *for any* w.
(c) *Both* $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ *and* $h : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ *are proper, closed, and convex.*

This assumption remains preliminary. To develop our algorithms, we will need more conditions on $\mathcal{H}_i$ and possibly on f and h, which will be stated later. In addition, we can work with a sublevel set

$$\mathcal{L}_{\Psi_0}(w_0) := \{ w \in \mathrm{dom}\,(\Psi_0) : \Psi_0(w) \le \Psi_0(w_0) \} \tag{4}$$

of Ψ_0 for a given initial point w_0 from our methods. If $u_0^*(w)$ is uniquely well-defined for given $w \in \mathcal{L}_{\Psi_0}(w_0)$, then by the well-known Danskin's theorem, Φ_0 is differential at w and its gradient is

$$\nabla \Phi_0(w) = \nabla_w \mathcal{H}(w, u_0^*(w)) = \frac{1}{n}\sum_{i=1}^n \nabla_w \mathcal{H}_i(w, u_0^*(w)). \tag{5}$$

We adopt the term "hyper-gradient" from bilevel optimization to name $\nabla\Phi_0$ in this paper.

2.2 Technical assumptions and properties of Φ_0 for nonconvex-linear setting (NL)

(a) ***Compositional minimization formulation.*** If $\mathcal{H}_i(w, u) := \langle F_i(w), Ku\rangle$ as in setting (NL), then (1) is equivalently reformulated into the following *nonconvex compositional minimization* problem:

$$\min_{w\in\mathbb{R}^p} \Big\{ \Psi_0(w) := f(w) + \Phi_0(w) = f(w) + h^*\Big(\frac{1}{n}\sum_{i=1}^n K^T F_i(w)\Big)\Big\}, \tag{CO}$$

where $h^*(v) := \sup_u\{\langle v, u\rangle - h(u)\}$, the Fenchel conjugate of h, and $\Phi_0(w) = h^*(K^T F(w))$. If h is not strongly convex, then h^* is convex but possibly nonsmooth.

(b) ***Technical assumptions.*** To develop our algorithms, we also need the following assumptions.

Assumption 2. *h is μ_h-strongly convex with $\mu_h \geq 0$, and $\mathrm{dom}(h)$ is bounded by $M_h < +\infty$.*

Assumption 3 (For F_i)**.** *For setting* (NL) *with $\mathcal{H}_i(w, u) := \langle F_i(w), Ku\rangle$ ($i \in [n]$), assume that*

- (a) *F_i is continuously differentiable, and its Jacobian ∇F_i is L_{F_i}-Lipschitz continuous.*
- (b) *F_i is also M_{F_i}-Lipschitz continuous or equivalently, its Jacobian ∇F_i is M_{F_i}-bounded.*
- (c) *There exists a positive constant $\sigma_J \in (0, +\infty)$ such that*

$$\frac{1}{n}\sum_{i=1}^n \|\nabla F_i(w) - \nabla F(w)\|^2 \leq \sigma_J^2, \quad \forall w \in \mathrm{dom}\,(F)\,. \tag{6}$$

Assumption 2 allows $\mu_h = 0$ that also covers the non-strong convexity of h. Assumption 3 is rather standard to develop gradient-based methods for solving (1). Under Assumption 3, the finite-sum F is also M_F-Lipschitz continuous and the Jacobian ∇F of F is also L_F-Lipschitz continuous with

$$M_F := \max\{M_{F_i} : i \in [n]\} \quad \text{and} \quad L_F := \max\{L_{F_i} : i \in [n]\}. \tag{7}$$

Condition (6) can be relaxed to the form $\frac{1}{n}\sum_{i=1}^n \|\nabla F_i(w) - \nabla F(w)\|^2 \leq \sigma_J^2 + \Theta_J\|\nabla\Phi_0(w)\|^2$ for some $\Theta_J \geq 0$, where $\nabla\Phi_0$ is a [sub]gradient of Φ_0 or Φ_γ (its smoothed approximation). Moreover, under Assumption 3, if $\mu_h > 0$, then ∇h^* is L_{h^*}-Lipschitz continuous with $L_{h^*} := \frac{1}{\mu_h}$. Thus it is possible (see [9]) to prove that Φ_0 is differentiable, and $\nabla\Phi_0$ is also L_{Φ_0}-Lipschitz continuous with $L_{\Phi_0} := M_h\|K\|L_F + \frac{M_F^2\|K\|^2}{\mu_h}$ as a consequence of Lemma 4 when $\gamma \downarrow 0^+$ in Supp. Doc. A.

(c) ***Smoothing technique for lower-level maximization problem*** (2)**.** If h is only merely convex (i.e. $\mu_h = 0$), then (2) may not be uniquely solvable, leading to the possible non-differentiability of Φ_0. Let us define the following convex function:

$$\phi_0(v) := \max_{u\in\mathbb{R}^q}\{\langle v, Ku\rangle - h(u)\} = h^*(K^T v). \tag{8}$$

Then, Φ_0 in (2) or (CO) can be written as $\Phi_0(w) = \phi_0(F(w)) = \phi_0\left(\frac{1}{n}\sum_{i=1}^n F_i(w)\right)$. Our goal is to smooth ϕ_0 if h is not strongly convex, leading to

$$\begin{cases} \phi_\gamma(v) := \max\limits_u \{\langle v, Ku\rangle - h(u) - \gamma b(u)\}, \\ u_\gamma^*(v) := \operatorname*{argmax}\limits_u \{\langle v, Ku\rangle - h(u) - \gamma b(u)\}, \end{cases} \tag{9}$$

where $\gamma > 0$ is a given smoothness parameter and $b : \mathbb{R}^q \to \mathbb{R}$ is a proper, closed, and 1-strongly convex function such that $\mathrm{dom}(h) \subseteq \mathrm{dom}(b)$. We also denote $D_b := \sup\{\|\nabla b(u)\| : u \in \mathrm{dom}\,(h)\}$. In particular, if we choose $b(u) := \frac{1}{2}\|u - \bar{u}\|^2$ for a fixed $\bar{u}$, then $u_\gamma^*(v) = \mathrm{prox}_{h/\gamma}(\bar{u} - K^T v)$.

Using ϕ_γ, problem (CO) can be approximated by its smoothed formulation:

$$\min_{w\in\mathbb{R}^p}\Big\{\Psi_\gamma(w) := f(w) + \Phi_\gamma(w) = f(w) + \phi_\gamma(F(w)) \equiv f(w) + \phi_\gamma\Big(\frac{1}{n}\sum_{i=1}^n F_i(w)\Big)\Big\}. \tag{10}$$

To develop our method, one key step is to approximate the hyper-gradient of Φ_γ in (10), where

$$\nabla\Phi_\gamma(w) = \nabla F(w)^T \nabla\phi_\gamma(F(w)) = \frac{1}{n}\sum_{i=1}^n \nabla F_i(w)^T \nabla\phi_\gamma(F(w)). \tag{11}$$

Then, $\nabla\Phi_\gamma$ is L_{Φ_γ}-Lipschitz continuous with $L_{\Phi_\gamma} := M_h\|K\|L_F + \frac{M_F^2\|K\|^2}{\mu_h+\gamma}$ (see Lemma 4).

2.3 Technical assumptions and properties of Φ_0 for the nonconvex-strongly-concave setting

To develop our shuffling gradient-based algorithms for solving (1) under the nonconvex-strongly-concave setting (NC), we impose the following assumptions.

Assumption 4 (For $\mathcal{H}_i$)**.** *$\mathcal{H}_i$ for all $i \in [n]$ in (1) satisfies the following conditions:*

(a) *For any given w such that $(w, u) \in \mathrm{dom}(\mathcal{H})$, $\mathcal{H}_i(w, \cdot)$ is μ_H-strongly concave w.r.t. u.*
(b) *$\nabla \mathcal{H}_i$ is (L_w, L_u)-Lipschitz continuous, i.e. for all $(w, u), (\hat{w}, \hat{u}) \in \mathrm{dom}(\mathcal{H})$:*

$$\|\nabla \mathcal{H}_i(w, u) - \nabla \mathcal{H}_i(\hat{w}, \hat{u})\|^2 \le L_w^2 \|w - \hat{w}\|^2 + L_u^2 \|u - \hat{u}\|^2. \tag{12}$$

(c) *There exist two constants $\Theta_w \ge 0$ and $\sigma_w \ge 0$ such that for $(w, u) \in \mathrm{dom}(\mathcal{H})$, we have*

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_w \mathcal{H}_i(w, u) - \nabla_w \mathcal{H}(w, u)\|^2 \le \Theta_w \|\nabla_w \mathcal{H}(w, u)\|^2 + \sigma_w^2. \tag{13}$$

There exist two constants $\Theta_u \ge 0$ and $\sigma_u \ge 0$ such that for all $(w, u) \in \mathrm{dom}(\mathcal{H})$, we have

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_u \mathcal{H}_i(w, u) - \nabla_u \mathcal{H}(w, u)\|^2 \le \Theta_u \|\nabla_u \mathcal{H}(w, u)\|^2 + \sigma_u^2. \tag{14}$$

Assumption 4(a) makes sure that our lower-level maximization of (1) is well-defined. Assumption 4(b) and (c) are standard in shuffling gradient-type methods as often seen in nonconvex optimization [9].

Lemma 1 (Smoothness of Φ_0)**.** *Under Assumptions 2 and 4, $u_0^*(\cdot)$ in (2) is κ-Lipschitz continuous with $\kappa := \frac{L_u}{\mu_H + \mu_h}$. Moreover, $\nabla \Phi_0$ in (5) is L_{Φ_0}-Lipschitz continuous with $L_{\Phi_0} := (1 + \kappa) L_w$.*

2.4 Approximate KKT points and approximate stationary points

(a) ***Exact and approximate KKT points and stationary points.*** A pair $(w^\star, u^\star) \in \mathrm{dom}(\mathcal{L})$ is called a KKT (Karush-Kuhn-Tucker) point of (1) if

$$0 \in \nabla_w \mathcal{H}(w^\star, u^\star) + \partial f(w^\star) \quad \text{and} \quad 0 \in -\nabla_u \mathcal{H}(w^\star, u^\star) + \partial h(u^\star). \tag{15}$$

Given a tolerance $\epsilon > 0$, **our goal** is to find an ϵ-approximate KKT point $(\widehat{w}, \widehat{u})$ of (1) defined as

$$r_w \in \nabla_w \mathcal{H}(\widehat{w}, \widehat{u}) + \partial f(\widehat{w}), \quad r_u \in -\nabla_u \mathcal{H}(\widehat{w}, \widehat{u}) + \partial h(\widehat{u}), \quad \text{and} \quad \mathbb{E}\big[\|[r_w, r_u]\|^2\big] \le \epsilon^2. \tag{16}$$

A vector $w^\star \in \mathrm{dom}(\Psi_0)$ is said to be a stationary point of (3) if

$$0 \in \nabla \Phi_0(w^\star) + \partial f(w^\star). \tag{17}$$

Since f is possibly nonsmooth, we can define a stationary point of (3) via a gradient mapping as:

$$\mathcal{G}_\eta(w) := \eta^{-1}\big(w - \mathrm{prox}_{\eta f}(w - \eta \nabla \Phi_0(w))\big), \tag{18}$$

where $\eta > 0$ is given. It is well-known that $\mathcal{G}_\eta(w^\star) = 0$ iff $w^\star$ is a stationary point of (3). Again, since we cannot exactly compute $w^\star$, we expect to find an ϵ-stationary point $\widehat{w}_T$ of (3) such that $\mathbb{E}\big[\|\mathcal{G}_\eta(\widehat{w}_T)\|^2\big] \le \epsilon^2$ for a given tolerance $\epsilon > 0$.

(b) ***Constructing an approximate stationary point and KKT point from algorithms.*** Our algorithms below generate a sequence $\{\widetilde{w}_t\}_{t \ge 0}^T$ such that $\frac{1}{T+1}\sum_{t=0}^T \mathbb{E}\big[\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2\big] \le \epsilon^2$. Hence, we construct an ϵ-stationary point $\widehat{w}_T$ using one of the following two options:

$$\widehat{w}_T := \widetilde{w}_{t_*}, \ \text{ where } \begin{cases} t_* := \mathrm{argmin}\{\|\mathcal{G}_\eta(\widetilde{w}_t)\| : 0 \le t \le T\}, & (\textit{Option 1}) \ \text{ or} \\ t_* \text{ is uniformly randomly chosen from } \{0, 1, \cdots, T\} & (\textit{Option 2}). \end{cases} \tag{19}$$

Clearly, we have $\mathbb{E}\big[\|\mathcal{G}_\eta(\widehat{w}_T)\|^2\big] \le \frac{1}{T+1}\sum_{t=0}^T \mathbb{E}\big[\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2\big] \le \epsilon^2$. We need the following result.

Lemma 2. (a) *If $(w^\star, u^\star)$ is a KKT point of (1), then $w^\star$ is a stationary point of (3). Conversely, if $w^\star$ is a stationary point of (3), then $(w^\star, u_0^*(w^\star))$ is a KKT point of (1).*

(b) *If $\widehat{w}_T$ is an ϵ-stationary point of (3) and $\nabla \Phi_0$ is L_{Φ_0}-Lipschitz continuous, then $(\overline{w}_T, \overline{u}_T)$ is an $\hat{\epsilon}$-KKT point of (1), where $\overline{w}_T := \mathrm{prox}_{\eta f}(\widehat{w}_T - \eta \nabla \Phi_0(\widehat{w}_T))$, $\overline{u}_T := u_0^*(\overline{w}_T)$, and $\hat{\epsilon} := (1 + L_{\Phi_0}\eta)\epsilon$.*

(c) *If $\widehat{w}_T$ is an ϵ-stationary point of (10), then $(\overline{w}_T, \overline{u}_T)$ is an $\hat{\epsilon}$-KKT point of (1), where $\overline{w}_T := \mathrm{prox}_{\eta f}(\widehat{w}_T - \eta \nabla \Phi_\gamma(\widehat{w}_T))$, $\overline{u}_T := u_\gamma^*(F(\overline{w}_T))$, and $\hat{\epsilon} := \max\{(1 + L_{\Phi_\gamma}\eta)\epsilon, \gamma D_b\}$.*

Lemma 2 allows us to construct an $\hat{\epsilon}$-approximate KKT point $(\overline{w}_T, \overline{u}_T)$ of (1) from an ϵ-stationary point $\widehat{w}_T$ of either (3) or its smoothed problem (10), where $\hat{\epsilon} = \mathcal{O}(\max\{\epsilon, \gamma\})$.

2.5 Technical condition to handle the possible nonsmooth term f

To handle the nonsmooth term f of (1) in our algorithms we require one more condition as in [5].

Assumption 5. *Let Φ_γ be defined by (10), which reduces to Φ_0 given by (2) as $\gamma \downarrow 0^+$, and $\mathcal{G}_\eta$ be defined by (18). Assume that there exist two constants $\Lambda_0 \geq 1$ and $\Lambda_1 \geq 0$ such that:*

$$\|\nabla\Phi_\gamma(w)\|^2 \leq \Lambda_0\|\mathcal{G}_\eta(w)\|^2 + \Lambda_1, \quad \forall w \in \mathrm{dom}\,(\Phi_0)\,. \tag{20}$$

If $f = 0$, then $\mathcal{G}_\eta(w) \equiv \nabla\Phi_\gamma(w)$, and Assumption 5 automatically holds with $\Lambda_0 = 1$ and $\Lambda_1 = 0$. If $f \neq 0$, then it is crucial to have $\Lambda_0 \geq 1$ in (20). Let us consider two examples to see why?

(i) If f is M_f-Lipschitz continuous (e.g., ℓ_1-norm), then (20) also holds with $\Lambda_0 := 1 + \nu > 1$ and $\Lambda_1 := \frac{1+\nu}{\nu} M_f$ for a given $\nu > 0$.

(ii) If $f = \delta_{\mathcal{W}}$, the indicator of a nonempty, closed, convex, and bounded set $\mathcal{W}$, then Assumption 5 also holds by the same reason as in Example (i) (see Supp. Doc. A).

3 Shuffling Gradient Method for Nonconvex-Linear Minimax Problems

We first propose a new construction using shuffling techniques to approximate the true gradient $\nabla\Phi_\gamma$ in (11) for any $\gamma \geq 0$. Next, we propose our algorithm and analyze its convergence.

3.1 The shuffling gradient estimators for $\nabla\Phi_\gamma$

Challenges. To evaluate $\nabla\Phi_\gamma(w)$ in (11), we need to evaluate both $\nabla F(w)$ and $F(w)$ at each w. However, in SGD or shuffling gradient methods, we want to approximate both quantities at each iteration. Note that this gradient can be written in a finite-sum $\frac{1}{n}\sum_{i=1}^n \nabla F_i(w)^T\nabla\phi_\gamma(F(w))$ (see (11)), but every summand requires $\nabla\phi_\gamma(F(w))$, which involves the full evaluation of F.

Our estimators. Let $F_{\pi^{(t)}(i)}(w_{i-1}^{(t)})$ and $\nabla F_{\hat{\pi}^{(t)}(i)}(w_{i-1}^{(t)})$ be the function value and the Jacobian component evaluated at $w_{i-1}^{(t)}$ respectively for $i \in [n]$, where $\pi^{(t)} = (\pi^{(t)}(1), \pi^{(t)}(2), \cdots, \pi^{(t)}(n))$ and $\hat{\pi}^{(t)} = (\hat{\pi}^{(t)}(1), \hat{\pi}^{(t)}(2), \cdots, \hat{\pi}^{(t)}(n))$ are two permutations of $[n] := \{1, 2, \cdots, n\}$. We want to use these quantities to approximate the function value $F(w_0^{(t)})$ and its Jacobian $\nabla F(w_0^{(t)})$ of F at $w_0^{(t)}$, respectively, where $w_0^{(t)}$ the iterate vector at the beginning of each epoch t.

For function value $F(w_0^{(t)})$, we suggest the following approximation at each *inner iteration* $i \in [n]$:

$$\textbf{Option 1:} \qquad F_i^{(t)} := \frac{1}{n}\Big[\sum_{j=1}^i F_{\pi^{(t)}(j)}(w_{j-1}^{(t)}) + \sum_{j=i+1}^n F_{\pi^{(t)}(j)}(w_0^{(t)})\Big]. \tag{21}$$

Alternative to (21), for all $i \in [n]$, we can simply choose another option:

$$\textbf{Option 2:} \qquad F_i^{(t)} := \frac{1}{n}\sum_{j=1}^n F_j(w_0^{(t)}) = \frac{1}{n}\sum_{j=1}^n F_{\pi^{(t)}(j)}(w_0^{(t)}). \tag{22}$$

For Jacobian $\nabla F(w_0^{(t)})$, we suggest to use the following standard shuffling estimator for all $i \in [n]$:

$$\nabla F_i^{(t)} := \nabla F_{\hat{\pi}^{(t)}(i)}(w_{i-1}^{(t)}). \tag{23}$$

For $F_i^{(t)}$ from (21) (or (22)) and for $\nabla F_i^{(t)}$ from (23), we form an approximation of $\nabla\Phi_\gamma(w_0^{(t)})$ as

$$\widetilde{\nabla}\Phi_\gamma(w_{i-1}^{(t)}) := (\nabla F_i^{(t)})^T\nabla\phi_\gamma(F_i^{(t)}) \equiv (\nabla F_i^{(t)})^T K u_\gamma^*(F_i^{(t)}). \tag{24}$$

Discussion. The estimator $F_i^{(t)}$ for F requires $n - i$ more function evaluations $F_{\pi^{(t)}(j)}(w_0^{(t)})$ at each epoch t. The first option (21) for F uses $2n$ function evaluations F_i, while the second one in (22) only needs n function evaluations at each epoch $t \geq 0$. However, (21) uses the most updated information up to the *inner iteration* i compared to (22), which is expected to perform better. The Jacobian estimator $\nabla F_i^{(t)}$ is standard and only uses one sample or a mini-batch at each iteration i.

3.2 The shuffling gradient-type algorithm for nonconvex-linear setting (NL)

We propose Algorithm 1, a shuffling gradient-type method, to approximate a stationary point of (10).

Discussion. First, the cost per epoch of Algorithm 1 consists of either $2n$ or n function evaluations F_i, and n Jacobian evaluations ∇F_i. Compare to standard shuffling gradient-type methods, e.g., in [8], Algorithm 1 has either n more evaluations of F_i or the same cost. Second, when implementing

Algorithm 1 (Shuffling Proximal Gradient-Based Algorithm for Solving (10))

Initialization: Choose an initial point $\widetilde{w}_0 \in \mathrm{dom}\,(\Phi_0)$ and a smoothness parameter $\gamma > 0$.
for $t = 1, 2, \cdots, T$ **do**
  Set $w_0^{(t)} := \widetilde{w}_{t-1}$;
  Generate two permutations $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ of $[n]$ (identically or randomly and independently)
  for $i = 1, \cdots, n$ **do**
    Evaluate $F_i^{(t)}$ by either (21) or (22) using $\pi^{(t)}$, and $\nabla F_i^{(t)}$ by (23) using $\hat{\pi}^{(t)}$.
    Solve (9) to get $u_\gamma^*(F_i^{(t)})$ and form $\widetilde{\nabla}\Phi_\gamma(w_{i-1}^{(t)}) := (\nabla F_i^{(t)})^T K u_\gamma^*(F_i^{(t)})$.
    Update $w_i^{(t)} := w_{i-1}^{(t)} - \frac{\eta_t}{n}\widetilde{\nabla}\Phi_\gamma(w_{i-1}^{(t)})$;
  end for
  Compute $\widetilde{w}_t := \mathrm{prox}_{\eta_t f}(w_n^{(t)})$;
end for

Algorithm 1, we do not need to evaluate the full Jacobian $\nabla F_i^{(t)}$, but rather the product of matrix $(\nabla F_i^{(t)})^T$ and vector $\nabla\Phi_\gamma(F_i^{(t)})$ as $\widetilde{\nabla}\Phi_\gamma(w_{i-1}^{(t)}) := (\nabla F_i^{(t)})^T\nabla\Phi_\gamma(F_i^{(t)})$. Evaluating this matrix-vector multiplication is much more efficient than evaluating the full Jacobian $\nabla F_i^{(t)}$ and $\nabla\Phi_\gamma(F_i^{(t)})$ individually. Third, thanks to Assumption 5, the proximal step $\widetilde{w}_t := \mathrm{prox}_{\eta_t f}(w_n^{(t)})$ is only required at the end of each epoch t. This significantly reduces the computational cost if $\mathrm{prox}_{\eta_t f}$ is expensive.

3.3 Convergence Analysis of Algorithm 1 for Nonconvex-Linear Setting (NL)

Now, we are ready to state the convergence result of Algorithm 1 in a short version: Theorem 1. The full version of this theorem is Theorem 6, which can be found in Supp. Doc. B.

Theorem 1. *Suppose that Assumptions 1, 2, 3, and 5 holds for the setting* (NL) *of* (1) *and* $\epsilon > 0$ *is a sufficiently small tolerance. Let* $\{\widetilde{w}_t\}$ *be generated by Algorithm 1 after* $T = \mathcal{O}(\epsilon^{-3})$ *epochs using arbitrarily permutations* $\pi^{(t)}$ *and* $\hat{\pi}^{(t)}$ *and a learning rate* $\eta_t = \eta := \mathcal{O}(\epsilon)$ *(see Theorem 6 in Supp. Doc. B for the exact formulas of* T *and* η*). Then, we have* $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_{\eta_t}(\widetilde{w}_t)\|^2 \le \epsilon^2$.

Alternatively, if $\{\widetilde{w}_t\}$ *is generated by Algorithm 1 after* $T := \mathcal{O}(n^{-1/2}\epsilon^{-3})$ *epochs using two random and independent permutations* $\pi^{(t)}$ *and* $\hat{\pi}^{(t)}$ *and a learning rate* $\eta_t = \eta := \mathcal{O}(n^{1/2}\epsilon)$ *(see Theorem 6 in Supp. Doc. B for the exact formulas). Then, we have* $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\mathcal{G}_{\eta_t}(\widetilde{w}_t)\|^2] \le \epsilon^2$.

Our first goal is to approximate a stationary point $w^\star$ of (CO) as $\mathbb{E}[\|\mathcal{G}_\eta(\widehat{w})\|^2] \le \epsilon^2$, while Algorithm 1 only provides an ϵ-stationary of (10). For a proper choice of γ, it is also an ϵ-stationary point of (3).

Corollary 1. *Let* $\widehat{w}_T$ *defined by* (19) *be generated from* $\{\widetilde{w}_t\}$ *of Algorithm 1. Under the conditions of Theorem 1 and any permutations* $\pi^{(t)}$ *and* $\hat{\pi}^{(t)}$*, the following statements hold.*

(a) *If* h *is* μ_h*-strongly convex with* $\mu_h > 0$*, then we can set* $\gamma = 0$*, and Algorithm 1 requires* $\mathcal{O}(n\epsilon^{-3})$ *evaluations of* F_i *and* ∇F_i *to achieve an* ϵ*-stationary* $\widehat{w}_T$ *of* (3).

(b) *If* h *is only convex (i.e.* $\mu_h = 0$*), then we can set* $\gamma := \mathcal{O}(\epsilon)$*, and Algorithm 1 needs* $\mathcal{O}(n\epsilon^{-7/2})$ *evaluations of* F_i *and* ∇F_i *to achieve an* ϵ*-stationary* $\widehat{w}_T$ *of* (3).

If, in addition, $\pi^{(t)}$ *and* $\hat{\pi}^{(t)}$ *are sampled uniformly at random without replacement and independently, and* $\Lambda_1 = \mathcal{O}(n^{-1})$*, then the numbers of evaluations of* F_i *and* ∇F_i *are reduced by a factor of* $\sqrt{n}$*.*

4 Shuffling Method for Nonconvex-Strongly Concave Minimax Problems

In this section, we develop shuffling gradient-based methods to solve (1) under the nonconvex-strongly concave setting (NC). Since this setting does not cover the nonconvex-linear setting (NL) in Section 3 as a special case, we need to treat it separately using different ideas and proof techniques.

4.1 The construction of algorithm

Unlike the linear case with $\mathcal{H}_i(w, u) = \langle F_i(w), Ku\rangle$ in Section 3, we cannot generally compute the solution $u_0^*(\widetilde{w}_{t-1})$ in (2) exactly for a given $\widetilde{w}_{t-1}$. We can only approximate $u_0^*(\widetilde{w}_{t-1})$ by some $\widetilde{u}_t$. This leads to another level of inexactness in an approximate "hyper-gradient" $\widetilde{\nabla}\Phi_0(w_{i-1}^{(t)})$ defined by

$$\widetilde{\nabla}\Phi_0(w_{i-1}^{(t)}) := \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(i)}(w_{i-1}^{(t)}, \widetilde{u}_t). \tag{25}$$

There are different options to approximate $u_0^*(\widetilde{w}_{t-1})$. We propose two options below, but other choices are possible, including accelerated gradient ascent methods and stochastic algorithms [6, 20].

(a_1) **Gradient ascent scheme for the lower-level problem.** We apply a standard gradient ascent scheme to update $\widetilde{u}_t$: *Starting from $s = 0$ with $u_0^{(t)} := \widetilde{u}_{t-1}$, at each epoch $s = 1, \cdots, S$, we update*

$$\widehat{u}_s^{(t)} := \mathrm{prox}_{\hat{\eta}_t h}\big(\widehat{u}_{s-1}^{(t)} + \tfrac{\hat{\eta}_t}{n}\textstyle\sum_{i=1}^n \nabla_u \mathcal{H}_i(\widetilde{w}_{t-1}, \widehat{u}_{s-1}^{(t)})\big), \tag{26}$$

for a given learning rate $\hat{\eta}_t > 0$. Then, we finally output $\widetilde{u}_t := \widehat{u}_S^{(t)}$ to approximate $u_0^*(\widetilde{w}_{t-1})$.

To make our method more flexible, we allow to perform either only *one iteration* (i.e. $S = 1$) or *multiple iterations* (i.e. $S > 1$) of (26). Each iteration s requires n evaluations of $\nabla_u \mathcal{H}_i$.

(a_2) **Shuffling gradient ascent scheme for the lower-level problem.** We can also construct $\widetilde{u}_t$ by a *shuffling gradient ascent scheme*. Again, we allow to run either only *one epoch* (i.e. $S = 1$) or *multiple epochs* (i.e. $S > 1$) of the shuffling algorithm to update $\widetilde{u}_t$, leading to the following scheme: *Starting from $s := 1$ with $\widehat{u}_0^{(t)} := \widetilde{u}_{t-1}$, at each epoch $s = 1, 2, \cdots, S$, having $\widehat{u}_{s-1}^{(t)}$, we generate a permutation $\pi^{(s,t)}$ of $[n]$ and run a shuffling gradient ascent scheme as*

$$\left\{\begin{array}{l} u_0^{(s,t)} := \widehat{u}_{s-1}^{(t)}, \\ \textit{For } i = 1, 2, \cdots, n, \textit{ update} \\ \qquad u_i^{(s,t)} := u_{i-1}^{(s,t)} + \frac{\hat{\eta}_t}{n}\nabla_u \mathcal{H}_{\pi^{(s,t)}(i)}(\widetilde{w}_{t-1}, u_{i-1}^{(s,t)}), \\ \widehat{u}_s^{(t)} := \mathrm{prox}_{\hat{\eta}_t h}(u_n^{(s,t)}). \end{array}\right. \tag{27}$$

At the end of the S-th epoch, we output $\widetilde{u}_t := \widehat{u}_S^{(t)}$ as an approximation to $u_0^*(\widetilde{w}_{t-1})$. Here, we use the same learning rate $\hat{\eta}_t$ for all epochs $s \in [S]$. Each epoch s requires n evaluations of $\nabla_u \mathcal{H}_i$.

(b) **Shuffling gradient descent scheme for the upper-level minimization problem.** Having $\widetilde{u}_t$ from either (26) or (27), we run a *shuffling gradient descent epoch* to update $\widetilde{w}_t$ from $\widetilde{w}_{t-1}$ as

$$\left\{\begin{array}{l} w_0^{(t)} := \widetilde{w}_{t-1}, \\ \text{For } i = 1, 2, \cdots, n, \text{ update} \\ \qquad w_i^{(t)} := w_{i-1}^{(t)} - \frac{\eta_t}{n}\widetilde{\nabla}\Phi_0(w_{i-1}^{(t)}) \equiv w_{i-1}^{(t)} - \frac{\eta_t}{n}\nabla_w \mathcal{H}_{\hat{\pi}^{(t)}(i)}(w_{i-1}^{(t)}, \widetilde{u}_t), \\ \widetilde{w}_t := \mathrm{prox}_{\eta_t f}(w_n^{(t)}). \end{array}\right. \tag{28}$$

These two steps (26) (or (27)) in u and (28) in w are implemented alternatively for $t = 1, \cdots, T$.

(c) **The full algorithm.** Combining both steps (26) (or (27)) and (28), we can present an *alternating shuffling proximal gradient algorithm* for solving (1) as in Algorithm 2.

Algorithm 2 (Alternating Shuffling Proximal Gradient Algorithm for Solving (1) under setting (NC))

1: **Initialization:** Choose an initial point $(\widetilde{w}_0, \widetilde{u}_0) \in \mathrm{dom}(\mathcal{L})$.
2: **for** $t = 1, 2, \cdots, T$ **do**
3: Compute $\widetilde{u}_t$ using either (26) or (27).
4: Set $w_0^{(t)} := \widetilde{w}_{t-1}$ and generate a permutation $\hat{\pi}^{(t)}$ of $[n]$.
5: **for** $i = 1, \cdots, n$ **do**
6: Evaluate $\widetilde{\nabla}\Phi_0(w_{i-1}^{(t)}) := \nabla_w \mathcal{H}_{\hat{\pi}^{(t)}(i)}(w_{i-1}^{(t)}, \widetilde{u}_t)$.
7: Update $w_i^{(t)} := w_{i-1}^{(t)} - \frac{\eta_t}{n}\widetilde{\nabla}\Phi_0(w_{i-1}^{(t)})$.
8: **end for**
9: Compute $\widetilde{w}_t := \mathrm{prox}_{\eta_t f}(w_n^{(t)})$.
10: **end for**

Discussion. Algorithm 2 has a similar form as Algorithm 1, where $u_0^*(\widetilde{w}_{t-1})$ is approximated by $\widetilde{u}_t$. In Algorithm 1, $u_0^*(\widetilde{w}_{t-1})$ is approximated by $u_\gamma^*(F_i^{(t)})$. Moreover, Algorithm 1 solves the smoothed problem (10) of (3), while Algorithm 2 directly solves (3). Depending on the choice of method to approximate $u_0^*(\widetilde{w}_{t-1})$, we obtain different variants of Algorithm 2. We have proposed two variants:

- **Semi-shuffling variant:** We use (26) for computing $\widetilde{u}_t$ to approximate $u_0^*(\widetilde{w}_{t-1})$.
- **Full-shuffling variant:** We use (27) for computing $\widetilde{u}_t$ to approximate $u_0^*(\widetilde{w}_{t-1})$.

Note that Algorithm 2 works in an alternative manner, where it approximates $u_0^*(\widetilde{w}_{t-1})$ up to a certain accuracy before updating $\widetilde{w}_t$. This alternating update is very natural and has been widely applied to solve minimax optimization as well as bilevel optimization problems, see, e.g., [1, 9, 13].

4.2 Convergence analysis

Now, we state the convergence of both variants of Algorithm 2: *semi-shuffling* and *full-shuffling* variants. The full proof of the following theorems can be found in Supp. Doc. C.

(a) ***Convergence of the semi-shuffling variant.*** Our first result is as follows.

Theorem 2. *Suppose that Assumptions 1, 2, 4, and 5 hold for* (1)*, and $\mathcal{G}_\eta$ is defined by* (18)*.*

Let $\{(\widetilde{w}_t, \widetilde{u}_t)\}$ be generated by Algorithm 2 using the ***gradient ascent scheme*** (26) *with $\eta := \mathcal{O}(\epsilon)$ explicitly given in Theorem 8 of Supp. Doc. C, $\hat{\eta} \in (0, \frac{2}{L_u+\mu_h}]$, $S := \mathcal{O}\big(\frac{1}{\hat{\eta}}\big(\mu_h + \frac{4L_u\mu_H}{L_u+\mu_H}\big)^{-1}\big) = \mathcal{O}(1)$, and $T := \mathcal{O}(\epsilon^{-3})$ explicitly given in Theorem 8. Then, we have $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \epsilon^2$.*

Consequently, Algorithm 2 requires $\mathcal{O}(n\epsilon^{-3})$ evaluations of both $\nabla_w\mathcal{H}_i$ and $\nabla_u\mathcal{H}_i$ to achieve an ϵ-stationary point $\widehat{w}_T$ of (3) *computed by* (19)*.*

Note that Theorem 2 holds for both $S > 1$ and $S = 1$ (i.e. we perform only one iteration of (26)).

(b) ***Convergence of the full-shuffling variant – The case $S > 1$ with multiple epochs.*** We state our results for two separated cases: only $\mathcal{H}_i$ is μ_H-strongly convex, and only h is μ_h-strongly convex.

Theorem 3 (Strong convexity of $\mathcal{H}_i$)**.** *Suppose that Assumptions 1, 2, 4, and 5 hold, and $\mathcal{H}_i$ is μ_H-strongly concave with $\mu_H > 0$ for $i \in [n]$, but h is only merely convex.*

Let $\{(\widetilde{w}_t, \widetilde{u}_t)\}$ be generated by Algorithm 2 using S epochs of the ***shuffling routine*** (27) *and fixed learning rates $\eta_t = \eta := \mathcal{O}(\epsilon)$ as given in Theorem 8 of Supp. Doc. C for a given $\epsilon > 0$, $\hat{\eta}_t := \hat{\eta} = \mathcal{O}(\epsilon)$, $S := \lfloor \frac{\ln(7/2)}{\mu_H\hat{\eta}} \rfloor$, and $T := \mathcal{O}(\epsilon^{-3})$. Then, we have $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \epsilon^2$.*

Consequently, Algorithm 2 requires $\mathcal{O}(n\epsilon^{-3})$ evaluations of $\nabla_w\mathcal{H}_i$ and $\mathcal{O}(n\epsilon^{-4})$ evaluations of $\nabla_u\mathcal{H}_i$ to achieve an ϵ-stationary point $\widehat{w}_T$ of (3) *computed by* (19)*.*

Theorem 4 (Strong convexity of h)**.** *Suppose that Assumptions 1, 2, 4, and 5 hold for* (1)*, and h is μ_h-strongly convex with $\mu_h > 0$, but $\mathcal{H}_i$ is only merely concave for all $i \in [n]$. Then, under the same settings as in Theorem 3, but with $S := \lfloor \frac{\ln(7/2)}{\mu_h\hat{\eta}} \rfloor$, the conclusions of Theorem 3 still hold.*

(c) ***Convergence of the full-shuffling variant – The case $S = 1$ with one epoch.*** Both Theorems 3 and 4 require $\mathcal{O}(n\epsilon^{-4})$ evaluations of $\nabla_u\mathcal{H}_i$. To improve this complexity, we need two additional assumptions but can perform only one epoch of (27), i.e. $S = 1$.

Assumption 6. *Let $\hat{\mathcal{G}}_\eta(u) := \eta^{-1}(u - \mathrm{prox}_{\eta h}(u + \eta\nabla_u\mathcal{H}(w,u)))$ be the gradient mapping of $\psi(w,\cdot) := -\mathcal{H}(w,\cdot) + h(\cdot)$. Assume that there exist $\hat{\Lambda}_0 \geq 1$ and $\hat{\Lambda}_1 \geq 0$ such that*

$$\|\nabla_u\mathcal{H}(w,u)\|^2 \leq \hat{\Lambda}_0\|\hat{\mathcal{G}}_\eta(u)\|^2 + \hat{\Lambda}_1, \quad \forall (w,u) \in \mathrm{dom}\,(\mathcal{L})\,. \tag{29}$$

Clearly, if $h = 0$, then $\hat{\mathcal{G}}_\eta(u) = -\nabla_u\mathcal{H}(w,u)$ and (20) automatically holds for $\hat{\Lambda}_0 = 1$ and $\hat{\Lambda}_1 = 0$. Assumption 6 is similar to Assumption 5, and it is required to handle the prox operator of h in (27).

Assumption 7. *For f in* (1)*, there exists $L_f \geq 0$ such that*

$$f(y) \leq f(x) + \langle f'(x), y - x\rangle + \tfrac{L_f}{2}\|y - x\|^2, \quad \forall x, y \in \mathrm{dom}\,(f)\,,\ f'(x) \in \partial f(x). \tag{30}$$

Clearly, if f is L_f-smooth, then (30) holds. If f is also convex, then (30) implies that f is L_f-smooth.

Under these additional assumptions, we have the following result.

Theorem 5. *Suppose that Assumptions 1, 2, 4, 5, 6, and 7 hold and $\mathcal{G}_\eta$ is defined by* (18)*.*

Let $\{(\widetilde{w}_t, \widetilde{u}_t)\}$ be generated by Algorithm 2 using ***one epoch*** *($S = 1$) of the* ***shuffling routine*** (27)*, and fixed learning rates $\eta_t = \eta := \mathcal{O}(\epsilon)$ as in Theorem 9 of Supp. Doc. C for a given $\epsilon > 0$, $\hat{\eta}_t := \hat{\eta} = 30\kappa^2\eta$, and $T := \mathcal{O}(\epsilon^{-3})$, where $\kappa := \frac{L_u}{\mu_H+\mu_h}$. Then, we have $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \epsilon^2$.*

Consequently, Algorithm 2 requires $\mathcal{O}(n\epsilon^{-3})$ evaluations of both $\nabla_w \mathcal{H}_i$ and of $\nabla_u \mathcal{H}_i$ to achieve an ϵ-stationary point $\widehat{w}_T$ of (3) computed by (19).

Similar to Algorithm 1, if $\pi^{(s,t)}$ and $\hat{\pi}^{(t)}$ are generated randomly and independently, $\Lambda_1 = \mathcal{O}(1/n)$, and $\hat{\Lambda}_1 = \mathcal{O}(1/n)$, then our complexity stated above can be improved by a factor of $\sqrt{n}$. Nevertheless, we omit this analysis. Finally, we can combine each Theorem 2, 3, 4 or 5 and Lemma 2 to construct an $\hat{\epsilon}$-KKT point of (1). Theorem 5 has a better complexity than Theorems 3 and 4, but requires stronger assumptions. Algorithm 2 is also different from the one in [3] both in terms of algorithmic form and the underlying problem to be solved, while achieving the same oracle complexity.

5 Numerical Experiments

We perform some experiments to illustrate Algorithm 1 and compare it with two existing and related algorithms. Further details and additional experiments can be found in Supp. Doc. D.

We consider the following regularized stochastic minimax problem studied, e.g., in [9, 33]:

$$\min_{w\in\mathbb{R}^p} \Big\{ \max_{1\le j\le m} \big\{ \tfrac{1}{n}\textstyle\sum_{i=1}^n F_{i,j}(w) \big\} + \tfrac{\lambda}{2}\|w\|^2 \Big\}, \tag{31}$$

where $F_{i,j} : \mathbb{R}^p \times \Omega \to \mathbb{R}_+$ can be viewed as the loss of the j-th model for data point $i \in [n]$. If we define $\phi_0(v) := \max_{1\le j\le m}\{v_j\}$ and $f(w) := \frac{\lambda}{2}\|w\|^2$, then (31) can be reformulated into (3). Since $v_j \ge 0$, we have $\phi_0(v) := \max_{1\le j\le m}\{v_j\} = \|v\|_\infty = \max_{\|u\|_1\le 1}\langle v, u\rangle$, which is nonsmooth. Thus we can smooth ϕ_0 as $\phi_\gamma(v) := \max_{\|u\|_1\le 1}\{\langle v,u\rangle - (\gamma/2)\|u\|^2\}$ using $b(u) := \frac{1}{2}\|u\|^2$.

Here, we apply our problem (31) to solve a model selection problem in binary classification with nonnegative nonconvex losses, see, e.g., [41]. Each function $F_{i,j}$ belongs to 4 different nonconvex losses ($m = 4$): $F_{i,1}(w,\xi) := 1 - \tanh(b_i\langle a_i, w\rangle)$, $F_{i,2}(w,\xi) := \log(1 + \exp(-b_i\langle a_i, w\rangle)) - \log(1 + \exp(-b_i\langle a_i, w\rangle - 1))$, $F_{i,3}(w,\xi) := (1 - 1/(\exp(-b_i\langle a_i, w\rangle) + 1))^2$, and $F_{i,4}(w,\xi) := \log(1 + \exp(-b_i\langle a_i, w\rangle))$ (see [41] for more details), where (a_i, b_i) represents data samples.

We implement 4 algorithms: our `SGM` with 2 options, `SGD` from [10], and `Prox-Linear` from [11]. We test these algorithms on two datasets from LIBSVM [6]. We set $\lambda := 10^{-4}$ and update the smooothing parameter γ_t as $\gamma_t := \frac{1}{2(t+1)^{1/3}}$. The learning rate η for all algorithms is finely tuned from $\{100, 50, 10, 5, 1, 0.5, 0.1, 0.05, 0.01, 0.001, 0.0001\}$, and the results are shown in Figure 1 for **w8a** and **rcv1** datasets using $k_b = 32$ blocks. The details of this experiment is given in Supp. Doc. D.

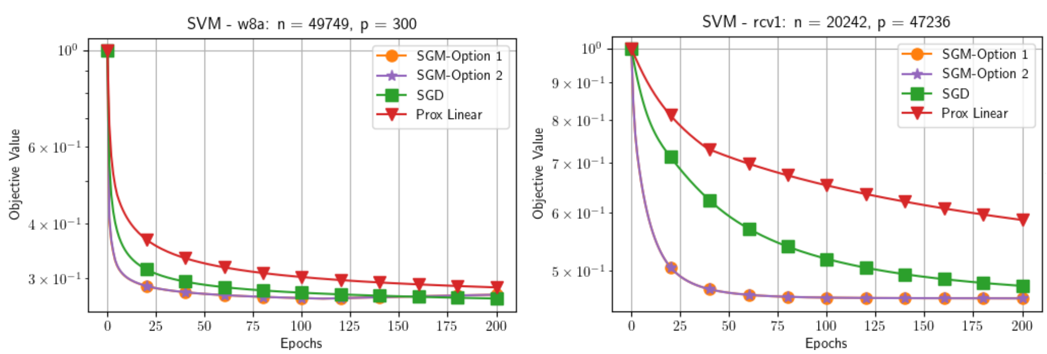

Figure 1: The performance of 4 algorithms for solving (31) on two datasets after 200 epochs.

As shown in Figure 1, the two variants of our `SGM` have a comparable performance with `SGD` and `Prox-Linear`, providing supportive evidence for using shuffling strategies in minimax algorithms.

6 Conclusions

This work explores a bilevel optimization approach to address two prevalent classes of nonconvex-concave minimax problems. These problems find numerous applications in practice, including robust learning and generative AIs. Motivated by the widespread use of shuffling strategies in implementing gradient-based methods within the machine learning community, we develop novel shuffling-based algorithms for solving these problems under standard assumptions. The first algorithm uses a non-standard shuffling strategy and achieves the state-of-the-art oracle complexity typically observed in nonconvex optimization. The second algorithm is also new, flexible, and offers a promising possibility for further exploration. Our results are expected to provide theoretical justification for incorporating shuffling strategies into minimax optimization algorithms, especially in nonconvex settings.

Acknowledgments and Disclosure of Funding

This work was partly supported by the National Science Foundation (NSF): NSF-RTG grant No. NSF DMS-2134107 and the Office of Naval Research (ONR), grant No. N00014-23-1-2588.

References

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

[2] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[4] A. Beznosikov, E. Gorbunov, H. Berard, and N. Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 172–235. PMLR, 2023.

[5] K. Bhatia and K. Sridharan. Online learning with dynamics: A minimax perspective. *Advances in Neural Information Processing Systems*, 33:15020–15030, 2020.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[7] H. Cho and C. Yun. SGDA with shuffling: faster convergence for nonconvex-PŁ minimax optimization. *The 11th International Conference on Learning Representations*, pp. 1–10, 2022.

[8] A. Das, B. Schölkopf, and M. Muehlebach. Sampling without replacement leads to faster rates in finite-sum minimax optimization. *Advances in Neural Information Processing Systems*, 35:6749–6762, 2022.

[9] S. Dempe. *Foundations of Bilevel Programming*. Springer Science & Business Media, 2002.

[10] D. Driggs, J. Liang, and C.-B. Schönlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, vol. 23, no. 24, pp. 1–43, 2022.

[11] K. Emmanouilidis, R. Vidal, and N. Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. In *International Conference on Artificial Intelligence and Statistics*, pages 3682–3690. PMLR, 2024.

[12] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *International Conference on Learning Representations*, pp. 1–10, 2019.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022.

[15] E. Y. Hamedani, A. Jalilzadeh, N. S. Aybat, and U. V. Shanbhag. Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. *arXiv preprint arXiv:1806.04118*, 2018.

[16] J. Z. HaoChen and S. Sra. Random shuffling beats SGD after finite epochs. *International Conference on Machine Learning*, pp. 2624–2633, 2019.

[17] E. Ho, A. Rajagopalan, A. Skvortsov, S. Arulampalam, and M. Piraveenan. Game theory in defence applications: A review. *Sensors*, 22(3):1032, 2022.

[18] Y. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

[19] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys (CSUR)*, 54(8):1–49, 2021.

[20] G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

[21] F. Lin, X. Fang, and Z. Gao. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control & Optimization*, 12(1):159, 2022.

[22] N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

[23] L. Luo, H. Ye, and T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, vol. 33, pp. 20566–20577, 2020.

[24] Z. Luo, J. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, 1996.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[26] Q. Meng, W. Chen, Y. Wang, Z.-M. Ma, and T.-Y. Liu. Convergence analysis of distributed stochastic gradient descent with shuffling. *Neurocomputing*, 337:46–57, 2019.

[27] K. Mishchenko, A. Khaled, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

[28] K. Mishchenko, A. Khaled, and P. Richtárik. Proximal and federated random reshuffling. In *International Conference on Machine Learning*, pages 15718–15749. PMLR, 2022.

[29] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

[30] L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.

[31] B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

[32] I. Safran and O. Shamir. How good is SGD with random shuffling? *Conference on Learning Theory*, pp. 3250–3284, 2020.

[33] A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optim. Methods Softw.*, 17(3):523–542, 2002.

[34] Q. Tran-Dinh, D. Liu, and L. M. Nguyen. Hybrid variance-reduced SGD algorithms for nonconvex-concave minimax problems. *The 34th Conference on Neural Information Processing Systems (NeurIPs 2020)*, 2020.

[35] J. Wang, T. Zhang, S. Liu, P.-Y. Chen, J. Xu, M. Fardad, and B. Li. Adversarial attack generation empowered by min-max optimization. *Advances in Neural Information Processing Systems*, 34:16020–16033, 2021.

[36] M. Wang, E. Fang, and L. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math. Program.*, 161(1-2):419–449, 2017.

[37] J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[38] B. Ying, K. Yuan, and A. H. Sayed. Convergence of variance-reduced stochastic learning under random reshuffling. *arXiv preprint arXiv:1708.01383*, 2(3):6, 2017.

[39] J. Zhang and L. Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Mathematical Programming*, pp. 1–43, 2022.

[40] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[41] L. Zhao, M. Mammadov, and J. Yearwood. From convex to nonconvex: a loss function analysis for binary classification. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1281–1288. IEEE, 2010.

[42] R. Zhao. Optimal stochastic algorithms for convex-concave saddle-point problems. *arXiv preprint arXiv:1903.01687*, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims made in the abstract reflects our contribution stated in the introduction, see the "Contribution" paragraph in the introduction section. Our contribution consists of two algorithms, Algorithm 1 and Algorithm 2, and their theoretical convergence guarantees stated in the subsequent theorems.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This paper has limitation as it only focuses on two classes of minimax problems defined in (1). Yes, we only consider two classes of minimax problems: nonconvex-linear (NL) and convex-strongly concave (NC), covered by our assumption, Assumptions 1 to 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We state all required assumptions in Assumptions 1 to 5. Our theoretical results stated in each theorem also refers to these assumptions when required. Our full proofs are given in Supp. Docs. due to space limit, and we believe that our technical proofs are correct.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the details of our experiments, including mathematical models, the detailed implementation of algorithms, the choice of parameters, and datasets. We also upload the code with examples to run and verify.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our data is available online from LIBSVM. The code is implemented in Python. The code for all experiments is also provided with instruction.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Supp. Doc. D provides all the details of our experiments, including how to select parameters, and how to report our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not have such a result to report.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments were run on a MacBook Pro. 2.8GHz Quad-Core Intel Core I7, 16Gb Memory specified at the beginning of Supp. Doc. D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our data is publicly available online from LIBSVM.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: We do not yet know if our paper has an immediate broader impact. However, since our problems and our algorithms are sufficiently general, we hope they will create broader impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not have our own real data or specific model that has a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Our code is open-source and will be made available online under a standard public license.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: It does not have new asset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: It does not relate to crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: It does not require any approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

Supplementary Document:
Shuffling Gradient-Based Methods for Nonconvex-Concave Minimax Optimization

Due to space limit, some results in the main text are not fully presented and clearly clarified. This supplementary document provides further details of our results in the main text. It also provides and proves technical lemmas used in this paper, presents the full proofs of our theoretical results, and additional examples and details of our numerical experiments.

A Technical Results and Proofs

This section gives the details of results related to minimax problem (1), and discusses the underlying technical assumptions and the properties of related functions and quantities used in this paper.

(a) **Elementary facts.** We recall the following facts, which will be used in the sequel.

[F_1] If $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper, closed, and μ_h-strongly convex, and $\mathrm{prox}_{\eta h}$ is the proximal operator of ηh for any $\eta > 0$, then for any $u, \hat{u} \in \mathrm{dom}\,(h)$, we have

$$\|\mathrm{prox}_{\eta h}(u) - \mathrm{prox}_{\eta h}(\hat{u})\|^2 \leq \tfrac{1}{1+2\mu_h\eta}\|u - \hat{u}\|^2. \tag{32}$$

[F_2] For any proper, closed, and convex function $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ and $\eta > 0$, we have

$$x - \mathrm{prox}_{\eta h}(x) \in \eta\partial h(\mathrm{prox}_{\eta h}(x)).$$

[F_3] Consider the lower level maximization problem (2) as

$$u_0^*(w) := \operatorname*{argmax}_{u\in\mathbb{R}^q}\big\{\mathcal{H}(w,u) - h(u) \equiv \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathcal{H}_i(w,u) - h(u)\big\}.$$

Then, under Assumption 1, its optimality condition can be written as

$$\nabla_u \mathcal{H}(w, u_0^*(w)) \in \partial h(u_0^*(w)). \tag{33}$$

(b) **Details of Assumption 5 and Assumption 6.** Both Assumptions 5 and 6 look relatively technical, though they have been used in previous works such as [5]. Both assumptions are the same, but one for f and the other for h, and thus we only discuss Assumption 5.

Note that [5] did not provide any example to motivate Assumption 5 for the case $f \neq 0$. Assumption 5 extends the one from [5] so that it holds for certain cases, including the two examples described after Assumption 5. Here, we further elaborate these examples in detail.

(i) *Example 1*. If f is M_f-Lipschitz continuous (e.g., the ℓ_1-norm), then (20) in Assumption 5 also holds. Indeed, since f is M_f-Lipschitz continuous, it is obvious that ∂f is M_f-bounded, and hence, by the fact $[F_2]$ above, we have $\|\mathrm{prox}_{\eta f}(u) - u\| \leq \eta M_f$ for any u. Using this inequality, and the definition of $\mathcal{G}_\eta$ in (18), we can easily show that

$$\|\nabla\Phi_\gamma(w) - \mathcal{G}_\eta(w)\| = \gamma^{-1}\|\mathrm{prox}_{\eta f}(w - \gamma\nabla\Phi_\gamma(w)) - (w - \gamma\nabla\Phi_\gamma(w))\| \leq M_f.$$

Then, for any $\nu > 0$, by Young's inequality, we have $\|\nabla\Phi_\gamma(w)\|^2 \leq (1+\nu)\|\mathcal{G}_\eta(w)\|^2 + \frac{1+\nu}{\nu}\|\nabla\Phi_\gamma(w) - \mathcal{G}_\eta(w)\|^2 \leq (1+\nu)\|\mathcal{G}_\eta(w)\|^2 + \frac{1+\nu}{\nu}M_f$. Hence, Assumption 5 holds for $\Lambda_0 := 1+\nu$ and $\Lambda_1 := \frac{1+\nu}{\nu}M_f$.

(ii) *Example 2*. It is also easy to check that if $f = \delta_{\mathcal{W}}$, the indicator of a nonempty, closed, convex, and bounded set $\mathcal{W}$, then for any $w \in \mathcal{W}$, we also have $\|\mathrm{prox}_{\eta f}(w) - w\| = \|\mathrm{proj}_{\mathcal{W}}(w) - w\| \leq 2\mathrm{diam}(\mathcal{W})$, where $\mathrm{diam}(\mathcal{W})$ is the diameter of $\mathcal{W}$. Hence, by the same proof as in *Example 1*, Assumption 5 also holds.

(c) **Technical results.** The following lemma summarizes the properties of ϕ_γ defined by (9), which was proved in [9]. It will be used in the sequel for analyzing Algorithm 1.

Lemma 3. *Let ϕ_0 and ϕ_γ be defined by* (8) *and* (9)*, respectively. Then, under Assumption 3:*

(a) $\mathrm{dom}(h)$ *is bounded by M_h iff ϕ_γ is M_{Φ_0}-Lipschitz continuous with $M_{\phi_0} := M_h\|K\|$.*

(b) *ϕ_γ is L_{ϕ_γ}-smooth with $L_{\phi_\gamma} := \frac{\|K\|^2}{\mu_h+\gamma}$ (i.e. $\nabla\phi_\gamma$ is L_{ϕ_γ}-Lipschitz continuous).*

(c) *$\phi_\gamma(v) \leq \phi_0(v) \leq \phi_\gamma(v) + \gamma B_{\phi_0}$ for any v, where $B_{\phi_0} := \sup\{b(u) : u \in \mathrm{dom}\,(h)\}$.*

(d) *For any $\hat{\gamma} \geq \gamma > 0$ and v, we have $\phi_\gamma(v) \leq \phi_{\hat{\gamma}}(v)+(\hat{\gamma}-\gamma)b(u^*_\gamma(v)) \leq \phi_{\hat{\gamma}}(v)+(\hat{\gamma}-\gamma)B_{\phi_0}$.*

(d) **The smoothness of Φ_γ and Φ_0.** One key step to develop our algorithms is to show that Φ_γ defined by (10) and Φ_0 in (2) are L-smooth (i.e. their gradient is Lipschitz continuous). The following lemma shows the L_{Φ_γ}-smoothness of Φ_γ defined in (10), whose proof is given in [9, Lemma A.3.].

Lemma 4 (Smoothness of Φ_γ)**.** *Under Assumption 3, $\nabla\Phi_\gamma$ of Φ_γ defined by* (11) *is L_{Φ_γ}-Lipschitz continuous with $L_{\Phi_\gamma} := M_h\|K\|L_F + \frac{M_F^2\|K\|^2}{\mu_h+\gamma}$, where $\gamma \geq 0$ such that $\mu_h + \gamma > 0$.*

Consequently, for any $w, \hat{w} \in \mathrm{dom}\,(\Phi_\gamma)$, we have

$$-\tfrac{L_{\Phi_\gamma}}{2}\|\hat{w} - w\|^2 \leq \Phi_\gamma(\hat{w}) - \Phi_\gamma(w) - \langle\nabla\Phi_\gamma(w), \hat{w} - w\rangle \leq \tfrac{L_{\Phi_\gamma}}{2}\|\hat{w} - w\|^2. \tag{34}$$

Alternatively, Lemma 1 in the main text can be expanded in detail as follows.

Lemma 5. *Under Assumption 4, let $u^*_0(\cdot)$ and Φ_0 be defined by* (2)*. Then, $u^*_0(\cdot)$ is κ-Lipschitz continuous with $\kappa := \frac{L_u}{\mu_H+\mu_h} > 0$, i.e.:*

$$\|u^*_0(w) - u^*_0(\hat{w})\| \leq \kappa\|w - \hat{w}\|, \quad \forall w, \hat{w} \in \mathrm{dom}\,(\Phi_0)\,. \tag{35}$$

Moreover, Φ_0 is L_{Φ_0}-smooth, i.e. $\|\nabla\Phi_0(w) - \nabla\Phi_0(\hat{w})\| \leq L_{\Phi_0}\|w - \hat{w}\|$ for all $w, \hat{w} \in \mathrm{dom}\,(\Phi_0)$, where $L_{\Phi_0} := (1 + \kappa)L_w$. Consequently, for all $w, \hat{w} \in \mathrm{dom}\,(\Phi_0)$, we have

$$-\frac{L_{\Phi_0}}{2}\|\hat{w} - w\|^2 \leq \Phi_0(\hat{w}) - \Phi_0(w) - \langle\nabla\Phi_0(w), \hat{w} - w\rangle \leq \frac{L_{\Phi_0}}{2}\|\hat{w} - w\|^2. \tag{36}$$

This lemma is proven similar to the one, e.g., in [7], and we omit it here.

(e) **Proof of Lemma 2 – Approximate stationary and KKT points.** Now, we provide the proof of Lemma 2 in the main text.

Proof of Lemma 2. (a) If $(w^\star, u^\star)$ is a KKT point of (1), then

$$0 \in \nabla_w\mathcal{H}(w^\star, u^\star) + \partial f(w^\star) \quad \text{and} \quad 0 \in -\nabla_u\mathcal{H}(w^\star, u^\star) + \partial h(u^\star).$$

Since $\mathcal{H}(w^\star, \cdot) - h(\cdot)$ is concave, $0 \in -\nabla_u\mathcal{H}(w^\star, u^\star) + \partial h(u^\star)$ implies that $u^\star \in \underset{u}{\mathrm{argmax}}\{\mathcal{H}(w^\star, u) - h(u)\}$. For Φ_0 defined by (2), by Danskin's theorem, we have $\nabla\Phi_0(w^\star) = \nabla_w\mathcal{H}(w^\star, u^\star)$. Hence, combining this relation and $0 \in \nabla_w\mathcal{H}(w^\star, u^\star) + \partial f(w^\star)$, we have $0 \in \nabla\Phi_0(w^\star) + \partial f(w^\star)$, which shows that $w^\star$ is a stationary point of (3). The converse statement is proved similarly, and we omit.

(b) If $\widehat{w}_T$ is an ϵ-stationary point of (3), then using a shorthand $g_T := \mathcal{G}_\eta(\widehat{w}_T)$, we have $\mathbb{E}[\|g_T\|^2] \leq \epsilon^2$. From (18), we have $g_T = \eta^{-1}(\widehat{w}_T - \mathrm{prox}_{\eta f}(\widehat{w}_T - \eta\nabla\Phi_0(\widehat{w}_T)))$, which is equivalent to $g_T \in \nabla\Phi_0(\widehat{w}_T) + \partial f(\widehat{w}_T - \eta g_T)$. Let us define $\overline{w}_T$ as in Lemma 2 and e_T as follows:

$$\begin{cases} \overline{w}_T \ := \ \widehat{w}_T - \eta g_T = \mathrm{prox}_{\eta f}(\widehat{w}_T - \eta\nabla\Phi_0(\widehat{w}_T))), \\ e_T \ := \ g_T + \nabla\Phi_0(\overline{w}_T) - \nabla\Phi_0(\widehat{w}_T). \end{cases} \tag{37}$$

Then, $g_T \in \nabla\Phi_0(\widehat{w}_T) + \partial f(\overline{w}_T)$ is equivalent to $e_T \in \nabla\Phi_0(\overline{w}_T) + \partial f(\overline{w}_T) = \nabla_w\mathcal{H}(\overline{w}_T, u^*_0(\overline{w}_T)) + \partial f(\overline{w}_T)$. On the other hand, from (33), we have $0 \in -\nabla_u\mathcal{H}(\overline{w}_T, u^*_0(\overline{w}_T)) + \partial h(u^*_0(\overline{w}_T))$. By the triangle inequality, and the L_{Φ_0}-Lipschitz continuity of $\nabla\Phi_0$, we have

$$\begin{aligned} \|e_T\| &\overset{(37)}{\leq} \|g_T\| + \|\nabla\Phi_0(\overline{w}_T) - \nabla\Phi_0(\widehat{w}_T)\| \\ &\leq \|g_T\| + L_{\Phi_0}\|\overline{w}_T - \widehat{w}_T\| \\ &\overset{(37)}{\leq} (1 + L_{\Phi_0}\eta)\|g_T\|. \end{aligned}$$

Hence, we get

$$\mathbb{E}[\|e_T\|^2] \leq (1+L_{\Phi_0}\eta)^2\mathbb{E}[\|g_T\|^2] \leq (1+L_{\Phi_0}\eta)^2\epsilon^2.$$

This concludes that if $\widehat{w}_T$ is an ϵ-stationary point of (3), then $(\overline{w}_T, u_0^*(\overline{w}_T))$ is an $\hat{\epsilon}$-KKT point of (1) with $\hat{\epsilon} := (1+L_{\Phi_0}\eta)\epsilon$.

(c) Since $\overline{w}_T := \mathrm{prox}_{\eta f}(\widehat{w}_T - \eta\nabla\Phi_\gamma(\widehat{w}_T)))$, we have $\widehat{w}_T - \overline{w}_T - \eta\nabla\Phi_\gamma(\widehat{w}_T) \in \eta\partial f(\overline{w}_T)$. Using this inclusion and

$$\nabla\Phi_\gamma(\overline{w}_T) = \nabla F(\overline{w}_T)^T\nabla\phi_\gamma(F(\overline{w}_T)) = \nabla F(\overline{w}_T)^T K u_\gamma^*(F(\overline{w}_T)) = \nabla_w\mathcal{H}(\overline{w}_T, u_\gamma^*(\overline{w}_T)),$$

we can show that

$$\begin{aligned}
\bar{r}_w &:= \eta^{-1}(\widehat{w}_T - \overline{w}_T) + \nabla\Phi_\gamma(\overline{w}_T) - \nabla\Phi_\gamma(\widehat{w}_T) \in \nabla\Phi_\gamma(\overline{w}_T) + \partial f(\overline{w}_T)\\
&\equiv \nabla_w\mathcal{H}(\overline{w}_T, u_\gamma^*(\overline{w}_T)) + \partial f(\overline{w}_T).
\end{aligned}$$

Since $\nabla\Phi_\gamma$ is L_{Φ_γ}-Lipschitz continuous and $\mathcal{G}_\eta(\overline{w}_T) = \eta^{-1}(\widehat{w}_T - \overline{w}_T)$, we have

$$\|\bar{r}_w\| \leq \|\mathcal{G}_\eta(\overline{w}_T)\| + \|\nabla\Phi_\gamma(\overline{w}_T) - \nabla\Phi_\gamma(\widehat{w}_T)\| \leq (1+\eta L_{\Phi_\gamma})\|\mathcal{G}_\eta(\overline{w}_T)\|.$$

On the other hand, since $\bar{u}_T := u_\gamma^*(F(\overline{w}_T))$, using the optimality condition of (9), and noticing that $\mathcal{H}(w,u) = \langle F(w), Ku\rangle$, we have

$$\bar{r}_u := -\gamma\nabla b(\bar{u}_T) \in -K^TF(\bar{u}_T) + \partial h(\bar{u}_T) \equiv -\nabla_u\mathcal{H}(\overline{w}_T, \bar{u}_T) + \partial h(\bar{u}_T).$$

Since $\mathrm{dom}\,(h)$ is bounded by M_h by Assumption 2, we can show that $\|\bar{r}_u\| = \gamma\|\nabla b(\bar{u}_T)\| \leq \gamma D_b$, where $D_b := \sup\{\|\nabla b(u)\| : u \in \mathrm{dom}\,(h)\}$. Combining the above analysis and noticing that $\mathbb{E}[\|\mathcal{G}_\eta(\overline{w}_T)\|^2] \leq \epsilon^2$, we can show that

$$\bar{r}_w \in \nabla_w\mathcal{H}(\overline{w}_T, \bar{u}_T) + \partial f(\overline{w}_T) \quad \text{and} \quad \bar{r}_u \in -\nabla_u\mathcal{H}(\overline{w}_T, \bar{u}_T) + \partial h(\bar{u}_T).$$

where $\mathbb{E}[\|\bar{r}_w\|^2] \leq (1+\eta L_{\Phi_\gamma})^2\epsilon^2$ and $\mathbb{E}[\|\bar{r}_u\|^2] \leq \gamma^2 D_b^2$. This proves that $(\overline{w}_T, \bar{u}_T)$ is an $\hat{\epsilon}$-KKT of (1) with $\hat{\epsilon} := \max\{(1+\eta L_{\Phi_\gamma})\epsilon, \gamma D_b\}$. Clearly, we have $\hat{\epsilon} = \mathcal{O}(\max\{\epsilon,\gamma\})$. In particular, if we choose $\eta := \mathcal{O}(\epsilon)$ and $\gamma := \mathcal{O}(\epsilon)$, then $\hat{\epsilon} = \mathcal{O}(\epsilon)$. $\square$

B Convergence Analysis of Algorithm 1 – The NL Setting

We first prove some key estimates for the shuffling estimator of $\nabla\Phi_\gamma(w)$. Next, we establish the technical lemmas that will be used to prove Theorem 6. Finally, we prove Theorem 6 and Corollary 1.

B.1 Properties of shuffling estimators

We state the following properties of $\widetilde{\nabla}\Phi_\gamma(\cdot)$ defined by (24), which could be of independent interest.

Lemma 6 (Arbitrary permutation). *Assume that Assumption 3 holds. Then*

(a) *For any $i \in [n]$, the approximation $F_i^{(t)}$ defined by (21) satisfies*

$$\|F_i^{(t)} - F(w_0^{(t)})\|^2 \leq \tfrac{M_F^2}{n}\textstyle\sum_{j=1}^n\|w_{j-1}^{(t)} - w_0^{(t)}\|^2. \tag{38}$$

(b) *Let $\mathcal{T}_{[i]} := \|\frac{1}{i}\sum_{j=1}^i\widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)}) - \nabla\Phi_\gamma(w_0^{(t)})\|^2$ for $\widetilde{\nabla}\Phi_\gamma(w_{i-1}^{(t)})$ defined by (24). Then*

$$\begin{aligned}
\mathcal{T}_{[i]} &\leq \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\textstyle\sum_{j=1}^n\|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + \tfrac{2nC_2\sigma_J^2}{i}\\
\mathcal{T}_{[n]} &\leq \tfrac{1}{n}\left(C_1 + 2C_2L_F^2\right)\textstyle\sum_{j=1}^n\|w_{j-1}^{(t)} - w_0^{(t)}\|^2,
\end{aligned} \tag{39}$$

where $C_1 := \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2}$ and $C_2 := 2M_h^2\|K\|^2$.

Proof. (a) Since $F(w_0^{(t)}) = \frac{1}{n}\sum_{j=1}^n F_j(w_0^{(t)}) = \frac{1}{n}\sum_{j=1}^n F_{\pi^{(t)}(j)}(w_0^{(t)})$, using **Option 1** as (21), we have

$$
\begin{aligned}
\|F_i^{(t)} - F(w_0^{(t)})\|^2 &= \tfrac{1}{n^2}\| \sum_{j=1}^i F_{\pi^{(t)}(j)}(w_{j-1}^{(t)}) + \sum_{j=i+1}^n F_{\pi^{(t)}(j)}(w_0^{(t)}) - \sum_{j=1}^n F_{\pi^{(t)}(j)}(w_0^{(t)})\|^2 \\
&= \tfrac{1}{n^2}\| \sum_{j=1}^i \left[F_{\pi^{(t)}(j)}(w_{j-1}^{(t)}) - F_{\pi^{(t)}(j)}(w_0^{(t)})\right]\|^2 \\
&\leq \tfrac{i}{n^2}\sum_{j=1}^i \|F_{\pi^{(t)}(j)}(w_{j-1}^{(t)}) - F_{\pi^{(t)}(j)}(w_0^{(t)})\|^2 \\
&\leq \tfrac{i\cdot M_F^2}{n^2}\sum_{j=1}^i \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 \\
&\leq \tfrac{M_F^2}{n}\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2,
\end{aligned}
$$

which proves (38) due to $1 \leq i \leq n$.

Alternatively if we use the update (22) as in **Option 2**, then we have $F_i^{(t)} = F(w_0^{(t)})$ which also automatically satisfies (38).

(b) From the definition of $\nabla\Phi_\gamma(w_0^{(t)})$ in (11) and of $\widetilde{\nabla}\Phi_\gamma(w_{i-1}^{(t)})$ in (24), by Young's inequality in ① and ②, the Cauchy-Schwarz inequality in ②, and Lemma 3 in ③, we have

$$
\begin{aligned}
\mathcal{T}_{[i]} &:= \|\tfrac{1}{i}\sum_{j=1}^i \widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)}) - \nabla\Phi_\gamma(w_0^{(t)})\|^2 \\
&= \|\tfrac{1}{i}\sum_{j=1}^i \left[(\nabla F_j^{(t)})^T\nabla\phi_\gamma(F_j^{(t)}) - \nabla F(w_0^{(t)})^T\nabla\phi_\gamma(F(w_0^{(t)}))\right]\|^2 \\
&= \|\tfrac{1}{i}\sum_{j=1}^i \left[(\nabla F_j^{(t)})^T\nabla\phi_\gamma(F_j^{(t)}) - (\nabla F_j^{(t)})^T\nabla\phi_\gamma(F(w_0^{(t)})\right. \\
&\qquad \left. + (\nabla F_j^{(t)})^T\nabla\phi_\gamma(F(w_0^{(t)}) - \nabla F(w_0^{(t)})^T\nabla\phi_\gamma(F(w_0^{(t)}))\right]\|^2 \\
&\overset{①}{\leq} 2\|\tfrac{1}{i}\sum_{j=1}^i (\nabla F_j^{(t)})^T\left[\nabla\phi_\gamma(F_j^{(t)}) - \nabla\phi_\gamma(F(w_0^{(t)})\right]\|^2 \\
&\qquad + 2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_j^{(t)} - \nabla F(w_0^{(t)})\right]^T\nabla\phi_\gamma(F(w_0^{(t)})\|^2 \\
&\overset{②}{\leq} \tfrac{2}{i}\sum_{j=1}^i \|\nabla F_j^{(t)}\|^2\|\nabla\phi_\gamma(F_j^{(t)}) - \nabla\phi_\gamma(F(w_0^{(t)})\|^2 \\
&\qquad + 2\|\nabla\phi_\gamma(F(w_0^{(t)})\|^2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_j^{(t)} - \nabla F(w_0^{(t)})\right]\|^2 \\
&\overset{③}{\leq} \tfrac{2M_F^2\|K\|^4}{i(\mu_h+\gamma)^2}\sum_{j=1}^i \|F_j^{(t)} - F(w_0^{(t)})\|^2 + 2M_h^2\|K\|^2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_j^{(t)} - \nabla F(w_0^{(t)})\right]\|^2.
\end{aligned}
$$

Substituting (38) into this estimate and noting that $C_1 = \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2}$ and $C_2 = 2M_h^2\|K\|^2$ we obtain

$$
\begin{aligned}
\mathcal{T}_{[i]} &\leq \tfrac{2M_F^2\|K\|^4}{i(\mu_h+\gamma)^2}\sum_{j=1}^i \tfrac{M_F^2}{n}\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + C_2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_j^{(t)} - \nabla F(w_0^{(t)})\right]\|^2 \\
&\leq C_2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_j^{(t)} - \nabla F_{\pi^{(t)}(j)}(w_0^{(t)})\right] + \tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\right]\|^2 \\
&\qquad + \tfrac{C_1}{n}\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 \\
&\leq \tfrac{C_1}{n}\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_{\pi^{(t)}(j)}(w_{j-1}^{(t)}) - \nabla F_{\pi^{(t)}(j)}(w_0^{(t)})\right]\|^2 \\
&\qquad + 2C_2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\right]\|^2 \\
&\leq \tfrac{C_1}{n}\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\tfrac{1}{i}\sum_{j=1}^i \|\nabla F_{\pi^{(t)}(j)}(w_{j-1}^{(t)}) - \nabla F_{\pi^{(t)}(j)}(w_0^{(t)})\|^2 \\
&\qquad + 2C_2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\right]\|^2 \\
&\leq \tfrac{C_1}{n}\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\tfrac{1}{i}\sum_{j=1}^n L_{F_j}^2\|w_{j-1}^{(t)} - w_0^{(t)}\|^2 \\
&\qquad + 2C_2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\right]\|^2 \\
&\leq \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\|\tfrac{1}{i}\sum_{j=1}^i \left[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\right]\|^2.
\end{aligned}
\tag{40}
$$

For $i = n$, we have

$$
\begin{aligned}
\mathcal{T}_{[n]} &\leq \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{n}\right)\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\|\tfrac{1}{n}\sum_{j=1}^n \left[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\right]\|^2 \\
&\leq \tfrac{1}{n}\left(C_1 + 2C_2L_F^2\right)\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2.
\end{aligned}
$$

For any other index $i \in [n]$ and $i < n$, we can show that

$$
\begin{aligned}
\mathcal{T}_{[i]} &\leq \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\tfrac{1}{i}\sum_{j=1}^i \|\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\|^2 \\
&\leq \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\tfrac{n}{i}\tfrac{1}{n}\sum_{j=1}^n \|\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\|^2,
\end{aligned}
$$

which proves the desired estimate. $\square$

If $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are generated randomly and independently, the we have the following result.

Lemma 7 (Random permutation)**.** *Assume that Assumption 3 holds, and $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are two random permutations of $[n]$. We recall that $\mathcal{T}_{[i]} := \|\frac{1}{i}\sum_{j=1}^{i}\widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)}) - \nabla\Phi_\gamma(w_0^{(t)})\|^2$. Then*

$$\mathbb{E}[\mathcal{T}_{[i]}] \le \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\textstyle\sum_{j=1}^{n}\mathbb{E}\big[\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\big] + \tfrac{2C_2}{i}\sigma_J^2, \tag{41}$$

where $C_1 := \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2}$ and $C_2 := 2M_h^2\|K\|^2$.

Proof. In this proof, we will use [4][Lemma 1] for sampling without replacement at random. From (40) in Lemma 6 we have

$$\mathcal{T}_{[i]} \le \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\textstyle\sum_{j=1}^{n}\|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + 2C_2\|\frac{1}{i}\sum_{j=1}^{i}\big[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\big]\|^2.$$

For each epoch $t = 1, \cdots, T$, we denote by $\mathcal{F}_t := \sigma(w_0^{(1)}, \cdots, w_0^{(t)})$ as the σ-algebra generated by the iterates of our algorithm (*cf.* Algorithm 1) up to the beginning of the epoch t. We observe that the permutation $\pi^{(t)}$ used at time t is independent of the σ-algebra $\mathcal{F}_t$. We also denote by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ as the conditional expectation on the σ-algebra $\mathcal{F}_t$.

Taking the expectation conditioned on $\mathcal{F}_t$, we get

$$\begin{aligned}\mathbb{E}_t[\mathcal{T}_{[i]}] \le{}& \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\textstyle\sum_{j=1}^{n}\mathbb{E}\big[\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\big]\\ &+ 2C_2\mathbb{E}\Big[\|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\big[\nabla F_{\pi^{(t)}(j)}(w_0^{(t)}) - \nabla F(w_0^{(t)})\big]\|^2\Big].\end{aligned}$$

By [4][Lemma 1] and Assumption 2(c), we have

$$\mathbb{E}_t[\mathcal{T}_{[i]}] \le \left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\textstyle\sum_{j=1}^{n}\mathbb{E}\big[\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\big] + 2C_2\tfrac{n-i}{i(n-1)}\sigma_J^2.$$

Taking the total expectation and noting that $n - i \le n - 1$ as $i \ge 1$, we get the desired estimate. □

B.2 One-iteration analysis of Algorithm 1: Key lemmas

The update of $w_i^{(t)}$ in Algorithm 1 can be written as

$$w_i^{(t)} = w_0^{(t)} - \tfrac{\eta_t}{n}\textstyle\sum_{j=1}^{i}\widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)}) = \widetilde{w}_{t-1} - \tfrac{\eta_t}{n}\sum_{j=1}^{i}\widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)}), \tag{42}$$

for $i \in [n]$, and $\widetilde{w}_t := \mathrm{prox}_{\eta_t f}(w_n^{(t)})$.

For simplicity of our proof, we also denote by $C_1 := \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2}$ and $C_2 := 2M_h^2\|K\|^2$. Using the expression (42), we can prove the following two lemmas.

Lemma 8. *Let $\{w_i^{(t)}\}$ be generated by Algorithm 1. If $\left(2C_1 + 4C_2L_F^2\right)\eta_t^2 \le \frac{1}{2}$, then we have*

$$\Delta_t := \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \le 4\eta_t^2\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2 + 2C_2\sigma_J^2\big]. \tag{43}$$

If $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are two random permutations of $[n] := \{1, 2, \cdots, n\}$, then

$$\widetilde{\Delta}_t := \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbb{E}\big[\|w_i^{(t)} - w_0^{(t)}\|^2\big] \le 4\eta_t^2\big[\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big] + \tfrac{2C_2\sigma_J^2}{n}\big]. \tag{44}$$

Proof. Using (42) and then (39), we can first derive that

$$\begin{aligned}\|w_i^{(t)} - w_0^{(t)}\|^2 ={}& \tfrac{\eta_t^2\cdot i^2}{n^2}\|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)})\|^2\\ \le{}& \tfrac{2\eta_t^2\cdot i^2}{n^2}\|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\big[\widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)}) - \nabla\Phi_\gamma(w_0^{(t)})\big]\|^2 + \tfrac{2\eta_t^2\cdot i^2}{n^2}\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\\ \le{}& \tfrac{2\eta_t^2\cdot i^2}{n^2}\left(\tfrac{C_1}{n} + \tfrac{2C_2L_F^2}{i}\right)\textstyle\sum_{j=1}^{n}\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\\ &+ \tfrac{2\eta_t^2\cdot i^2}{n^2}\tfrac{2nC_2\sigma_J^2}{i} + \tfrac{2\eta_t^2\cdot i^2}{n^2}\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\\ \le{}& \eta_t^2\left(\tfrac{2C_1\cdot i^2}{n^3} + \tfrac{2C_2L_F^2\cdot i}{n^2}\right)\textstyle\sum_{j=1}^{n}\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\\ &+ \tfrac{4C_2\sigma_J^2\eta_t^2\cdot i}{n} + \tfrac{2\eta_t^2\cdot i^2}{n^2}\|\nabla\Phi_\gamma(w_0^{(t)})\|^2.\end{aligned} \tag{45}$$

Let us denote $\Delta_t := \frac{1}{n}\sum_{i=1}^n \|w_{i-1}^{(t)} - w_0^{(t)}\|^2$. Then, from (45), we have

$$\begin{aligned}
\Delta_t &:= \tfrac{1}{n}\textstyle\sum_{i=1}^n \|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\le \tfrac{1}{n}\textstyle\sum_{i=1}^n \Big[\eta_t^2\Big(\frac{2C_1\cdot i^2}{n^3} + \frac{2C_2L_F^2\cdot i}{n^2}\Big)\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 + \frac{4C_2\sigma_J^2\eta_t^2\cdot i}{n} + \frac{2\eta_t^2\cdot i^2}{n^2}\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\Big] \\
&\le \eta_t^2\Big(\frac{2C_1\cdot\sum_{i=1}^n i^2}{n^3} + \frac{2C_2L_F^2\cdot\sum_{i=1}^n i}{n^2}\Big)\tfrac{1}{n}\textstyle\sum_{j=1}^n \|w_{j-1}^{(t)} - w_0^{(t)}\|^2 \\
&\quad + \frac{4C_2\sigma_J^2\eta_t^2\cdot\sum_{i=1}^n i}{n^2} + \frac{2\eta_t^2\cdot\sum_{i=1}^n i^2}{n^3}\|\nabla\Phi_\gamma(w_0^{(t)})\|^2 \\
&\le \eta_t^2\big(2C_1 + 4C_2L_F^2\big)\Delta_t + 4C_2\sigma_J^2\eta_t^2 + 2\eta_t^2\|\nabla\Phi_\gamma(w_0^{(t)})\|^2.
\end{aligned}$$

Here, we have used $\sum_{i=1} i^2 = \frac{n(n+1)(2n+1)}{6} \le n^3$, $\sum_{i=1}^n i = \frac{n(n+1)}{2} \le n^2$ in the last inequality. Under the condition $\eta_t^2\big(2C_1 + 4C_2L_F^2\big) \le \frac{1}{2}$, we obtain (43) from the last inequality.

If $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are two random permutations of $[n] := \{1, 2, \cdots, n\}$ using similar argument with (42) and then, with (41) we have:

$$\begin{aligned}
\mathbb{E}\big[\|w_i^{(t)} - w_0^{(t)}\|^2\big] &= \frac{\eta_t^2\cdot i^2}{n^2}\mathbb{E}\big[\|\tfrac{1}{i}\textstyle\sum_{j=1}^i \widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)})\|^2\big] \\
&\le \frac{2\eta_t^2\cdot i^2}{n^2}\mathbb{E}\big[\|\tfrac{1}{i}\textstyle\sum_{j=1}^i [\widetilde{\nabla}\Phi_\gamma(w_{j-1}^{(t)}) - \nabla\Phi_\gamma(w_0^{(t)})]\|^2\big] \\
&\quad + \frac{2\eta_t^2\cdot i^2}{n^2}\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big] \\
&\le \frac{2\eta_t^2\cdot i^2}{n^2}\Big(\frac{C_1}{n} + \frac{2C_2L_F^2}{i}\Big)\textstyle\sum_{j=1}^n \mathbb{E}\big[\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\big] \\
&\quad + \frac{2\eta_t^2\cdot i^2}{n^2}\frac{2C_2\sigma_J^2}{i} + \frac{2\eta_t^2\cdot i^2}{n^2}\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big] \\
&\le \eta_t^2\Big(\frac{2C_1\cdot i^2}{n^3} + \frac{2C_2L_F^2\cdot i}{n^2}\Big)\textstyle\sum_{j=1}^n \mathbb{E}\big[\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\big] \\
&\quad + \frac{4C_2\sigma_J^2\eta_t^2\cdot i}{n^2} + \frac{2\eta_t^2\cdot i^2}{n^2}\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big].
\end{aligned} \tag{46}$$

Let us denote $\widetilde{\Delta}_t := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\|w_{i-1}^{(t)} - w_0^{(t)}\|^2\big]$. Then, from (45), we have

$$\begin{aligned}
\widetilde{\Delta}_t &:= \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}\big[\|w_i^{(t)} - w_0^{(t)}\|^2\big] \\
&\le \tfrac{1}{n}\textstyle\sum_{i=1}^n \Big[\eta_t^2\Big(\frac{2C_1\cdot i^2}{n^3} + \frac{2C_2L_F^2\cdot i}{n^2}\Big)\sum_{j=1}^n \mathbb{E}\big[\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\big] \\
&\quad + \frac{4C_2\sigma_J^2\eta_t^2\cdot i}{n^2} + \frac{2\eta_t^2\cdot i^2}{n^2}\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big]\Big] \\
&\le \eta_t^2\Big(\frac{2C_1\cdot\sum_{i=1}^n i^2}{n^3} + \frac{2C_2L_F^2\cdot\sum_{i=1}^n i}{n^2}\Big)\tfrac{1}{n}\textstyle\sum_{j=1}^n \mathbb{E}\big[\|w_{j-1}^{(t)} - w_0^{(t)}\|^2\big] \\
&\quad + \frac{4C_2\sigma_J^2\eta_t^2\cdot\sum_{i=1}^n i}{n^3} + \frac{2\eta_t^2\cdot\sum_{i=1}^n i^2}{n^3}\cdot\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big] \\
&\le \eta_t^2\big(2C_1 + 4C_2L_F^2\big)\widetilde{\Delta}_t + 2\eta_t^2\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big] + \frac{4C_2\sigma_J^2\eta_t^2}{n}.
\end{aligned}$$

Using similar arguments as before that $\sum_{i=1} i^2 = \frac{n(n+1)(2n+1)}{6} \le n^3$, $\sum_{i=1}^n i = \frac{n(n+1)}{2} \le n^2$ and $\eta_t^2\big(2C_1 + 4C_2L_F^2\big) \le \frac{1}{2}$, we obtain (44) from the last inequality. □

Lemma 9. *Let $\{(w_i^{(t)}, \widetilde{w}_t)\}$ be generated by Algorithm 1. Then, we have*

$$\begin{aligned}
\Psi_\gamma(\widetilde{w}_t) &\le \Psi_\gamma(\widetilde{w}_{t-1}) - \frac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \frac{(1-L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\
&\quad + \frac{L_{\Psi_\gamma}\cdot\eta_t}{2n}\textstyle\sum_{i=1}^n \|w_{i-1}^{(t)} - w_0^{(t)}\|^2,
\end{aligned} \tag{47}$$

where $L_{\Psi_\gamma} := \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2} + 4M_h^2\|K\|^2L_F^2$ and L_{Φ_γ} is given in Lemma 4.

Proof. The proof of this lemma is adopted from the proof of [5, Theorem 3] with some modification. First, we denote $\widehat{w}_t := \mathrm{prox}_{\eta_t f}\big(\widetilde{w}_{t-1} - \eta_t\nabla\Phi_\gamma(\widetilde{w}_{t-1})\big)$. Then, from (18), we have $\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1}) = \frac{1}{\eta_t}(\widetilde{w}_{t-1} - \widehat{w}_t)$. Moreover, we also have $\nabla f(\widehat{w}_t) := \eta_t^{-1}\big(\widetilde{w}_{t-1} - \widehat{w}_t\big) - \nabla\Phi_\gamma(\widetilde{w}_{t-1}) \in \partial f(\widehat{w}_t)$.

Next, by the convexity of f, we can easily show that

$$\begin{aligned}
f(\widehat{w}_t) &\le f(\widetilde{w}_{t-1}) + \langle\nabla f(\widehat{w}_t), \widehat{w}_t - \widetilde{w}_{t-1}\rangle \\
&= f(\widetilde{w}_{t-1}) - \langle\nabla\Phi_\gamma(\widetilde{w}_{t-1}), \widehat{w}_t - \widetilde{w}_{t-1}\rangle - \frac{1}{\eta_t}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2.
\end{aligned}$$

Next, by the L_{Φ_γ}-smoothness of Φ_γ from (34) of Lemma 4, we have

$$\Phi_\gamma(\widehat{w}_t) \le \Phi_\gamma(\widetilde{w}_{t-1}) + \langle \nabla\Phi_\gamma(\widetilde{w}_{t-1}), \widehat{w}_t - \widetilde{w}_{t-1}\rangle + \tfrac{L_{\Phi_\gamma}}{2}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2.$$

Adding the last two inequalities together and using $\Psi_\gamma(w) = f(w) + \Phi_\gamma(w)$ and $\widehat{w}_t - \widetilde{w}_{t-1} = -\eta_t\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})$, we have

$$\Psi_\gamma(\widehat{w}_t) \le \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{(2-L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2 = \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{\eta_t(2-L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2. \tag{48}$$

Now, let us denote $g_t := \frac{1}{n}\sum_{i=0}^n \widetilde{\nabla}\Phi_\gamma(w_i^{(t)})$. Then, from the update of $w_i^{(t)}$, we have

$$g_t = \tfrac{1}{\eta_t}(\widetilde{w}_{t-1} - w_n^{(t)}) = \tfrac{1}{\eta_t}(w_0^{(t)} - w_n^{(t)}).$$

Since $\widetilde{w}_t = \mathrm{prox}_{\eta_t f}(w_n^{(t)})$, we have $\nabla f(\widetilde{w}_t) := \eta_t^{-1}\big(w_n^{(t)} - \widetilde{w}_t\big) = -g_t - \eta_t^{-1}(\widetilde{w}_t - \widetilde{w}_{t-1}) \in \partial f(\widetilde{w}_t)$. Hence, by the convexity of f, we have

$$\begin{aligned}
f(\widetilde{w}_t) &\le f(\widehat{w}^t) + \langle \nabla f(\widetilde{w}_t), \widetilde{w}_t - \widehat{w}^t\rangle = f(\widehat{w}^t) - \langle g_t, \widetilde{w}_t - \widehat{w}^t\rangle - \tfrac{1}{\eta_t}\langle \widetilde{w}_t - \widetilde{w}_{t-1}, \widetilde{w}_t - \widehat{w}^t\rangle \\
&= f(\widehat{w}^t) - \langle g_t, \widetilde{w}_t - \widehat{w}^t\rangle + \tfrac{1}{2\eta_t}\big[\|\widehat{w}^t - \widetilde{w}_{t-1}\|^2 - \|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \|\widetilde{w}_t - \widehat{w}^t\|^2\big].
\end{aligned}$$

Again, by the L_{Φ_γ}-smoothness of Φ_γ from (34) of Lemma 4, we also have

$$\begin{aligned}
\Phi_\gamma(\widetilde{w}_t) &\le \Phi_\gamma(\widetilde{w}_{t-1}) + \langle \nabla\Phi_\gamma(\widetilde{w}_{t-1}), \widetilde{w}_t - \widetilde{w}_{t-1}\rangle + \tfrac{L_{\Phi_\gamma}}{2}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2, \\
\Phi_\gamma(\widetilde{w}_{t-1}) &\le \Phi_\gamma(\widehat{w}^t) + \langle \nabla\Phi_\gamma(\widetilde{w}_{t-1}), \widetilde{w}_{t-1} - \widehat{w}^t\rangle + \tfrac{L_{\Phi_\gamma}}{2}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2.
\end{aligned}$$

Adding the last three inequalities together, and using $\Psi_\gamma(w) = f(w) + \Phi_\gamma(w)$ and $\widehat{w}_t - \widetilde{w}_{t-1} = -\eta_t\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})$, we have

$$\begin{aligned}
\Psi_\gamma(\widetilde{w}_t) &\le \Psi_\gamma(\widehat{w}^t) + \langle \nabla\Phi_\gamma(\widetilde{w}_{t-1}) - g_t, \widetilde{w}_t - \widehat{w}^t\rangle - \tfrac{(1-L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\
&\quad + \tfrac{(1+L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{1}{2\eta_t}\|\widetilde{w}_t - \widehat{w}^t\|^2 \\
&\le \Psi_\gamma(\widehat{w}^t) + \tfrac{\eta_t}{2}\|\nabla\Phi_\gamma(\widetilde{w}_{t-1}) - g_t\|^2 - \tfrac{(1-L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\
&\quad + \tfrac{\eta_t(1+L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2,
\end{aligned} \tag{49}$$

where we have used Young's inequality in the last line as $\langle \nabla\Phi_\gamma(\widetilde{w}_{t-1}) - g_t, \widetilde{w}_t - \widehat{w}^t\rangle \le \frac{\eta_t}{2}\|\nabla\Phi_\gamma(\widetilde{w}_{t-1}) - g_t\|^2 + \frac{1}{2\eta_t}\|\widetilde{w}_t - \widehat{w}^t\|^2$.

Summing up (48) and (49), we get

$$\begin{aligned}
\Psi_\gamma(\widetilde{w}_t) &\le \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{(1-L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\
&\quad + \tfrac{\eta_t}{2}\|\nabla\Phi_\gamma(\widetilde{w}_{t-1}) - g_t\|^2.
\end{aligned} \tag{50}$$

Using (39) with $g_t = \frac{1}{n}\sum_{i=0}^n \widetilde{\nabla}\Phi_\gamma(w_i^{(t)})$, we arrive at

$$\begin{aligned}
\Psi_\gamma(\widetilde{w}_t) &\le \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{(1-L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \tfrac{\eta_t}{2}\cdot\mathcal{T}_{[n]} \\
&\overset{(39)}{\le} \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{(1-L_{\Phi_\gamma}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\
&\quad + \tfrac{\eta_t}{2}\tfrac{1}{n}\big(C_1 + 2C_2L_F^2\big)\textstyle\sum_{j=1}^n\|w_{j-1}^{(t)} - w_0^{(t)}\|^2,
\end{aligned}$$

which is (47), where $L_{\Psi_\gamma} := C_1 + 2C_2L_F^2 = \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2} + 4M_h^2\|K\|^2L_F^2$. $\square$

B.3 The proof of Theorem 6 and Corollary 1 for Algorithm 1

Let us recall that $C_1 := \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2}$, $C_2 := 2M_h^2\|K\|^2$, and $L_{\Phi_\gamma} := M_h\|K\|L_F + \frac{M_F^2\|K\|^2}{\mu_h+\gamma}$ from Lemma 4. To prove Theorem 6, we will need the following lemma.

Lemma 10. *Let* $\{w_i^{(t)}\}$ *be generated by Algorithm 1 using arbitrarily permutations* $\pi^{(t)} = \hat{\pi}^{(t)}$*, and* $\eta_t = \eta > 0$ *such that* $\left(2C_1 + 4C_2L_F^2\right)\eta^2 \le \frac{1}{2}$ *and* $4L_{\Phi_\gamma}\eta + 8L_\Psi\Lambda_0\eta^2 \le 1$*. Then*

$$\tfrac{1}{T+1}\textstyle\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \ \le\ \tfrac{4}{T\eta}\left[\Psi_\gamma(\widetilde{w}^0) - \Psi_\gamma^\star\right] + 8L_\Psi(2C_2\sigma_J^2 + \Lambda_1)\cdot\eta^2. \tag{51}$$

Alternatively, if $\pi^{(t)}$ *and* $\hat{\pi}^{(t)}$ *are random permutations and generated independently, then, with a similar condition on* η *as above, we have*

$$\tfrac{1}{T+1}\textstyle\sum_{t=0}^{T}\mathbb{E}\left[\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2\right] \ \le\ \tfrac{4}{T\eta}\left[\Psi_\gamma(\widetilde{w}^0) - \Psi_\gamma^\star\right] + 8L_\Psi(2C_2\tfrac{\sigma_J^2}{n} + \Lambda_1)\cdot\eta^2. \tag{52}$$

Proof. From (47), and note that $L_{\Phi_0}\eta_t \le 1$, we obtain

$$\Psi_\gamma(\widetilde{w}_t) \le \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \tfrac{L_\Psi\cdot\eta_t}{2n}\textstyle\sum_{i=1}^{n}\|w_{i-1}^{(t)} - w_0^{(t)}\|^2.$$

Using (43) with the condition $\left(2C_1 + 4C_2L_F^2\right)\eta_t^2 \le \frac{1}{2}$ and (20) of Assumption 5, and $w_0^{(t)} = \widetilde{w}_{t-1}$, we have

$$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \ &\le\ 4\eta_t^2\left[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2 + 2C_2\sigma_J^2\right] \\ &\overset{(20)}{\le}\ 4\eta_t^2\left[\Lambda_0\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + 2C_2\sigma_J^2 + \Lambda_1\right].\end{aligned}$$

Combining the two estimates, we obtain

$$\begin{aligned}\Psi_\gamma(\widetilde{w}_t) \ &\le\ \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \tfrac{L_\Psi\eta_t}{2}\cdot 4\eta_t^2\left[\Lambda_0\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + 2C_2\sigma_J^2 + \Lambda_1\right] \\ &=\ \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{\eta_t}{2}\left(1 - 2L_{\Phi_\gamma}\eta_t - 4L_\Psi\Lambda_0\eta_t^2\right)\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + 2L_\Psi(2C_2\sigma_J^2 + \Lambda_1)\cdot\eta_t^3 \\ &\le\ \Psi_\gamma(\widetilde{w}_{t-1}) - \tfrac{\eta_t}{4}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + 2L_\Psi(2C_2\sigma_J^2 + \Lambda_1)\cdot\eta_t^3,\end{aligned}$$

provided that $4L_{\Phi_\gamma}\eta_t + 8L_\Psi\Lambda_0\eta_t^2 \le 1$. Following the same proof as in [8, Theorem 3], we obtain our bound in (51).

For the randomized bound, we take expectation and obtain

$$\mathbb{E}\left[\Psi_\gamma(\widetilde{w}_t)\right] \le \mathbb{E}\left[\Psi_\gamma(\widetilde{w}_{t-1})\right] - \tfrac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\mathbb{E}\left[\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2\right] + \tfrac{L_\Psi\cdot\eta_t}{2n}\textstyle\sum_{i=1}^{n}\mathbb{E}\big[\|w_{i-1}^{(t)} - w_0^{(t)}\|^2\big].$$

Using (44) with similar argument as the deterministic case, we have

$$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbb{E}\big[\|w_{i-1}^{(t)} - w_0^{(t)}\|^2\big] \ &\le\ 4\eta_t^2\Big[\mathbb{E}\big[\|\nabla\Phi_\gamma(w_0^{(t)})\|^2\big] + 2C_2\tfrac{\sigma_J^2}{n}\Big] \\ &\overset{(20)}{\le}\ 4\eta_t^2\Big[\Lambda_0\mathbb{E}\big[\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2\big] + 2C_2\tfrac{\sigma_J^2}{n} + \Lambda_1\Big].\end{aligned}$$

Combining the last two estimates, we get

$$\begin{aligned}\mathbb{E}\left[\Psi_\gamma(\widetilde{w}_t)\right] &\le \mathbb{E}\left[\Psi_\gamma(\widetilde{w}_{t-1})\right] - \tfrac{\eta_t(1-2L_{\Phi_\gamma}\eta_t)}{2}\mathbb{E}\left[\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2\right] \\ &\quad + \tfrac{L_\Psi\eta_t}{2}\cdot 4\eta_t^2\Big[\Lambda_0\mathbb{E}\left[\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2\right] + 2C_2\tfrac{\sigma_J^2}{n} + \Lambda_1\Big] \\ &= \mathbb{E}\left[\Psi_\gamma(\widetilde{w}_{t-1})\right] - \tfrac{\eta_t}{2}\left(1 - 2L_{\Phi_\gamma}\eta_t - 4L_\Psi\Lambda_0\eta_t^2\right)\mathbb{E}\left[\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2\right] \\ &\quad + 2L_\Psi(2C_2\tfrac{\sigma_J^2}{n} + \Lambda_1)\cdot\eta_t^3 \\ &\le \mathbb{E}\left[\Psi_\gamma(\widetilde{w}_{t-1})\right] - \tfrac{\eta_t}{4}\mathbb{E}\left[\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2\right] + 2L_\Psi(2C_2\tfrac{\sigma_J^2}{n} + \Lambda_1)\cdot\eta_t^3,\end{aligned}$$

provided that $4L_{\Phi_\gamma}\eta_t + 8L_\Psi\Lambda_0\eta_t^2 \le 1$. Follow the same proof as in [8, Theorem 3], we can easily get (52). $\square$

The following theorem, Theorem 6 is the full version of Theorem 1 in the main text, where the learning rate η and the number of epochs T are given explicitly.

Theorem 6. *Suppose that Assumptions 1, 2, 3, and 5 holds for the setting (NL) of* (1) *and*

$$Q_\gamma := \tfrac{M_F^2\|K\|^2}{\mu_h+\gamma} + M_hL_F\|K\|. \tag{53}$$

Let $\{\widetilde{w}_t\}$ be generated by Algorithm 1 after T epochs using arbitrarily deterministic permutations $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ and a learning rate $\eta_t = \eta > 0$ such that

$$\eta := \frac{\epsilon}{\sqrt{2Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}} \quad \text{and} \quad T := \left\lfloor \frac{16\sqrt{Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^\star+\gamma B_{\phi_0}]}{\epsilon^3} \right\rfloor, \tag{54}$$

for a given sufficiently small tolerance $\epsilon > 0$ such that $\eta \le \frac{1}{8Q_\gamma}$. Then, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_{\eta_t}(\widetilde{w}_t)\|^2 \le \epsilon^2.$$

Alternatively, if $\{\widetilde{w}_t\}$ is generated by Algorithm 1 after T epochs using two random and independent permutations $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ and a learning rate $\eta_t = \eta > 0$ such that

$$\eta := \frac{\sqrt{n}\epsilon}{\sqrt{2Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}} \quad \text{and} \quad T := \left\lfloor \frac{16\sqrt{Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^\star+\gamma B_{\phi_0}]}{\sqrt{n}\epsilon^3} \right\rfloor, \tag{55}$$

for a given sufficiently small tolerance $\epsilon > 0$ such that $\eta \le \frac{1}{8Q_\gamma}$. Then, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\mathcal{G}_{\eta_t}(\widetilde{w}_t)\|^2] \le \epsilon^2.$$

Proof of Theorem 6. Recall that $C_1 := \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2}$ and $C_2 := 2M_h^2\|K\|^2$, $L_{\Phi_\gamma} := M_h\|K\|L_F + \frac{M_F^2\|K\|^2}{\mu_h+\gamma}$, and $L_\Psi := \frac{2M_F^4\|K\|^4}{(\mu_h+\gamma)^2} + 4M_h^2\|K\|^2L_F^2$. Let us denote by $Q_\gamma := \frac{M_F^2\|K\|^2}{\mu_h+\gamma} + M_hL_F\|K\|$ as in Theorem 6.

In this case, the first conditions $\left(2C_1 + 4C_2L_F^2\right)\eta^2 \le \frac{1}{2}$ and $4L_{\Phi_\gamma}\eta + 8L_\Psi\Lambda_0\eta^2 \le 1$ of Lemma 10 respectively reduce to

$$\begin{aligned}&\frac{M_F^4\|K\|^4+2(\mu_h+\gamma)^2M_h^2\|K\|^2L_F^2}{(\mu_h+\gamma)^2}\cdot\eta^2 \le \frac{1}{8} \quad \text{and}\\ &\frac{M_F^2\|K\|^2+(\mu_h+\gamma)M_h\|K\|L_F}{\mu_h+\gamma}\cdot\eta + \frac{4(M_F^4\|K\|^4+2(\mu_h+\gamma)^2M_h^2\|K\|^2L_F^2)}{(\mu_h+\gamma)^2}\cdot\eta^2 \le \frac{1}{4}.\end{aligned}$$

Since

$$\begin{aligned}2\big(M_F^2\|K\|^2 + (\mu_h+\gamma)M_h\|K\|L_F\big)^2 &= 2M_F^4\|K\|^4 + 2(\mu_h+\gamma)^2M_h^2\|K\|^2L_F^2\\ &\quad + 4(\mu_h+\gamma)M_h\|K\|^3L_FM_F^2\\ &\ge M_F^4\|K\|^4 + 2(\mu_h+\gamma)^2M_h^2\|K\|^2L_F^2,\end{aligned}$$

the last two conditions hold if $0 < \eta \le \frac{\mu_h+\gamma}{8(M_F^2\|K\|^2+(\mu_h+\gamma)M_hL_F\|K\|)} = \frac{1}{8Q_\gamma}$. Moreover, we also have $L_\Psi \le 2Q_\gamma$.

Now, from (51), to guarantee $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \le \epsilon^2$, we impose

$$\frac{4}{T\eta}\left[\Psi_\gamma(\widetilde{w}^0) - \Psi_\gamma^\star\right] + 8L_\Psi(2C_2\sigma_J^2 + \Lambda_1)\cdot\eta^2 \le \epsilon^2.$$

Since $0 < \eta \le \frac{1}{8Q_\gamma}$ and $L_\Psi \le 2Q_\gamma$, we can choose $\eta := \frac{1}{2}\min\left\{\frac{1}{4Q_\gamma}, \frac{\epsilon}{\sqrt{Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}}\right\}$. Hence, the last inequality holds if

$$T \ge 16\cdot\max\left\{\frac{\sqrt{Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}}{\epsilon^3}, \frac{4Q_\gamma}{\epsilon^2}\right\}\cdot\left[\Psi_\gamma(\widetilde{w}^0) - \Psi_\gamma^\star\right].$$

By Lemma 3(c), we can easily show that $\Psi_\gamma(w) \le \Psi_0(w) \le \Psi_\gamma(w) + \gamma B_{\phi_0}$ for any w, where $B_{\phi_0} := \sup\{b(u) : u \in \mathrm{dom}\,(h)\}$. Hence, we have $\Psi_\gamma(\widetilde{w}^0) - \Psi_\gamma^\star \le \Psi_0(\widetilde{w}_0) - \Psi_0^\star + \gamma B_{\phi_0}$. Using this condition, we obtain

$$T := \left\lfloor 16\cdot\max\left\{\frac{\sqrt{Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}}{\epsilon^3}, \frac{4Q_\gamma}{\epsilon^2}\right\}\cdot[\Psi_0(\widetilde{w}_0) - \Psi_0^\star + \gamma B_{\phi_0}]\right\rfloor.$$

If we choose ϵ sufficiently small such that the $0 < \epsilon \le \frac{\sqrt{Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}}{4Q_\gamma}$, then

$$\eta := \frac{\epsilon}{\sqrt{2Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}} \quad \text{and} \quad T := \left\lfloor \frac{16\sqrt{Q_\gamma(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^\star+\gamma B_{\phi_0}]}{\epsilon^3} \right\rfloor,$$

as shown in (54) of Theorem 6.

If a random shuffling strategy is used, then to guarantee $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\mathcal{G}_{\eta}(\widetilde{w}_t)\|^2] \leq \epsilon^2$, from (52), we can impose the following condition

$$\frac{4}{T\eta}\left[\Psi_{\gamma}(\widetilde{w}^0) - \Psi_{\gamma}^{\star}\right] + 8L_{\Psi}(2C_2\frac{\sigma_J^2}{n} + \Lambda_1)\cdot\eta^2 \leq \epsilon^2.$$

Reasoning the same way as above, we can choose $\eta := \frac{1}{2}\min\left\{\frac{1}{4Q_{\gamma}}, \frac{\sqrt{n}\epsilon}{\sqrt{Q_{\gamma}(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}}\right\}$. This leads to the choice of T as

$$T := \left\lfloor 16\cdot\max\left\{\frac{\sqrt{Q_{\gamma}(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}}{\sqrt{n}\epsilon^3}, \frac{4Q_{\gamma}}{\epsilon^2}\right\}\cdot\left[\Psi_0(\widetilde{w}_0) - \Psi_0^{\star} + \gamma B_{\phi_0}\right]\right\rfloor.$$

If we choose ϵ sufficiently small such that the $0 < \epsilon \leq \frac{\sqrt{Q_{\gamma}(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}}{4Q_{\gamma}\sqrt{n}}$, then

$$\eta := \frac{\sqrt{n}\epsilon}{\sqrt{2Q_{\gamma}(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}} \quad\text{and}\quad T := \left\lfloor\frac{16\sqrt{Q_{\gamma}(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^{\star}+\gamma B_{\phi_0}]}{\sqrt{n}\epsilon^3}\right\rfloor,$$

as shown in (55) of Theorem 6. □

Proof of Corollary 1. (a) If h is μ_h-strongly convex with $\mu_h > 0$, then we can set $\gamma = 0$, i.e. without using smoothing technique. Then, we have $Q_{\gamma} := \frac{M_F^2\|K\|^2}{\mu_h+\gamma} + M_hL_F\|K\|$ reduces to $Q_0 := \frac{M_F^2\|K\|^2}{\mu_h} + M_hL_F\|K\|$.

If arbitrary permutations $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are used, then T from (54) reduces to

$$T := \left\lfloor\frac{16\sqrt{Q_0(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^{\star}]}{\epsilon^3}\right\rfloor.$$

Note that, each epoch $t \in [T]$ requires either $2n$ (for **Option 1**) or n (for **Option 2**) evaluations of F_i and n evaluations of ∇F_i. Hence, Algorithm 1 requires $\mathcal{O}(n\epsilon^{-3})$ evaluations of F_i and $\mathcal{O}(n\epsilon^{-3})$ evaluations of ∇F_i to achieve an ϵ-stationary point of (3).

Alternatively, if $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are random and independent permutations, then T from (55) reduces to

$$T := \left\lfloor\frac{16\sqrt{Q_0(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^{\star}]}{\sqrt{n}\epsilon^3}\right\rfloor.$$

Clearly, if $\Lambda_1 = \frac{\Gamma}{n}$ for some constant $\Gamma > 0$, then plugging this Λ_1 into the right-hand side of T above, we can conclude that Algorithm 1 requires $\mathcal{O}(\sqrt{n}\epsilon^{-3})$ evaluations of F_i and $\mathcal{O}(\sqrt{n}\epsilon^{-3})$ evaluations of ∇F_i to achieve an ϵ-stationary point of (3).

(b) If h is only merely convex, i.e. $\mu_h = 0$, then we have $Q_{\gamma} = \frac{M_F^2\|K\|^2}{\gamma} + M_hL_F\|K\| = \mathcal{O}(\gamma^{-1})$. Moreover, to obtain an ϵ-stationary point of (3) from a stationary point of its smoothed problem (10), with a similar proof as of Lemma 2, we need to choose $\gamma := \epsilon$. In this case, we get $Q_{\epsilon} = \mathcal{O}(\epsilon^{-1})$.

If arbitrary permutations $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are used, then T from (54) reduces to

$$T := \left\lfloor\frac{16\sqrt{Q_{\epsilon}(4M_h^2\|K\|^2\sigma_J^2+\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^{\star}+\epsilon B_{\phi_0}]}{\epsilon^3}\right\rfloor = \mathcal{O}\left(\frac{1}{\epsilon^{7/2}}\right).$$

Hence, Algorithm 1 requires $\mathcal{O}(n\epsilon^{-7/2})$ evaluations of F_i and $\mathcal{O}(n\epsilon^{-7/2})$ evaluations of ∇F_i to achieve an ϵ-stationary point of (3).

Alternatively, if $\pi^{(t)}$ and $\hat{\pi}^{(t)}$ are random and independent permutations, then T from (55) reduces to

$$T := \left\lfloor\frac{16\sqrt{Q_{\epsilon}(4M_h^2\|K\|^2\sigma_J^2+n\Lambda_1)}\cdot[\Psi_0(\widetilde{w}_0)-\Psi_0^{\star}+\epsilon B_{\phi_0}]}{\sqrt{n}\epsilon^3}\right\rfloor.$$

Clearly, if $\Lambda_1 = \frac{\Gamma}{n}$ for some constant $\Gamma > 0$, then plugging this Λ_1 into the right-hand side of T above, we can conclude that Algorithm 1 requires $\mathcal{O}(\sqrt{n}\epsilon^{-7/2})$ evaluations of F_i and $\mathcal{O}(\sqrt{n}\epsilon^{-7/2})$ evaluations of ∇F_i to achieve an ϵ-stationary point of (3). □

Remark 1. *We note that since each epoch t of Algorithm 1 requires one evaluation of $\mathrm{prox}_{\eta_t f}$, the total number of $\mathrm{prox}_{\eta_t f}$ evaluations is T.*

C Convergence Analysis of Algorithm 2 – The NC Setting

In this section, we present the full convergence analysis of Algorithm 2 for both the **semi-shuffling** and the **full-shuffling** variants.

For our notational convenience, we introduce the following function:

$$\psi(w, u) := -\mathcal{H}(w, u) + h(u). \tag{56}$$

By Assumption 4, $\psi(w, \cdot)$ is μ_ψ-strongly convex with the strong convexity parameter $\mu_\psi := \mu_h + \mu_H > 0$ for any w such that $(w, u) \in \mathrm{dom}\,(\mathcal{L})$. Moreover, the Lipschitz constant κ of $u_0^*(\cdot)$ in Lemma 1 becomes $\kappa := \frac{L_u}{\mu_h+\mu_H} = \frac{L_u}{\mu_\psi} > 0$.

Furthermore, Φ_0 and Ψ_0 defined by (2) and (3), respectively can be expressed as

$$\begin{aligned} \Phi_0(w) &:= \max_{u\in\mathbb{R}^q}\big\{\mathcal{H}(w,u) - h(u)\big\} = -\min_{u\in\mathbb{R}^q}\psi(w,u), \\ \Psi_0(w) &:= f(w) + \Phi_0(w) = f(w) + \mathcal{H}(w, u_0^*(w)) - h(u_0^*(w)), \end{aligned} \tag{57}$$

where $u_0^*(w) := \arg\min_{u\in\mathbb{R}^q} \psi(w,u)$ is computed by (2).

C.1 One-epoch analysis: Key lemmas

We separate the technical lemmas for two variants: the *semi-shuffling variant* using (26), and the *full-shuffling variant* using (27) into two subsections, respectively.

(a) Key bound for the gradient-ascent scheme (26). If we apply (26) to approximate $u_0^*(\widetilde{w}_{t-1})$, then we have the following result.

Lemma 11. *Suppose that Assumption 4 holds. Let $\{\widehat{u}_s^{(t)}\}$ be updated by (26) such that $0 < \hat{\eta}_t \le \frac{2}{L_u+\mu_H}$. Then, we have*

$$\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \le \tfrac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1 - \tfrac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\big)^S\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2. \tag{58}$$

Proof. The proof of Lemma 11 is certainly classical and not new. It can be found in the literature, including [6]. However, it may be inconvenient to find a unified proof for the strong convexity of $\mathcal{H}_i$ and h altogether. Therefore, we present it here for completeness.

For simplicity of our presentation, we denote $\varphi(u) := -\mathcal{H}(\widetilde{w}_{t-1}, u) = -\frac{1}{n}\sum_{i=1}^n \mathcal{H}_i(\widetilde{w}_{t-1}, u)$ and $u_t^* := u_0^*(\widetilde{w}_{t-1})$ computed by (2).

By Assumption 4, φ is μ_H-strongly convex and L_u-smooth. The scheme (26) is exactly a proximal gradient method to solve $\min_u\{Q(u) := \varphi(u) + h(u)\}$, where h is also μ_h-strongly convex. Moreover, by the definition of φ and of u_t^*, and (26), it is obvious to show that

$$\begin{cases} u_t^* &= \mathrm{prox}_{\hat{\eta}_t h}\big(u_t^* - \hat{\eta}_t\nabla\varphi(u_t^*)\big), \\ \widehat{u}_s^{(t)} &= \mathrm{prox}_{\hat{\eta}_t h}\big(\widehat{u}_{s-1}^{(t)} - \hat{\eta}_t\nabla\varphi(\widehat{u}_{s-1}^{(t)})\big). \end{cases}$$

Hence, by (32) from Fact $[F_1]$, we have

$$\begin{aligned} \|\widehat{u}_s^{(t)} - u_t^*\|^2 &= \|\mathrm{prox}_{\hat{\eta}_t h}\big(\widehat{u}_{s-1}^{(t)} - \hat{\eta}_t\nabla\varphi(\widehat{u}_{s-1}^{(t)})\big) - \mathrm{prox}_{\hat{\eta}_t h}\big(u_t^* - \hat{\eta}_t\nabla\varphi(u_t^*)\big)\|^2 \\ &\le \tfrac{1}{1+2\mu_h\hat{\eta}_t}\|\widehat{u}_{s-1}^{(t)} - u_t^* - \hat{\eta}_t[\nabla\varphi(\widehat{u}_{s-1}^{(t)}) - \nabla\varphi(u_t^*)]\|^2. \end{aligned}$$

Expanding the right-hand side of the last estimate, we get

$$\begin{aligned} \|\widehat{u}_{s-1}^{(t)} - u_t^* - \hat{\eta}_t[\nabla\varphi(\widehat{u}_{s-1}^{(t)}) - \nabla\varphi(u_t^*)]\|^2 &= \|\widehat{u}_{s-1}^{(t)} - u_t^*\|^2 + \hat{\eta}_t^2\|\nabla\varphi(\widehat{u}_{s-1}^{(t)}) - \nabla\varphi(u_t^*)\|^2 \\ &\quad - 2\hat{\eta}_t\langle\nabla\varphi(\widehat{u}_{s-1}^{(t)}) - \nabla\varphi(u_t^*), \widehat{u}_{s-1}^{(t)} - u_t^*\rangle. \end{aligned}$$

Using [6, Theorem 2.1.12], we can show that

$$\langle\nabla\varphi(\widehat{u}_{s-1}^{(t)}) - \nabla\varphi(u_t^*), \widehat{u}_{s-1}^{(t)} - u_t^*\rangle \ge \tfrac{L_u\mu_H}{L_u+\mu_H}\|\widehat{u}_{s-1}^{(t)} - u_t^*\|^2 + \tfrac{1}{L_u+\mu_H}\|\nabla\varphi(\widehat{u}_{s-1}^{(t)}) - \nabla\varphi(u_t^*)\|^2.$$

Combining the last three inequalities, we obtain

$$\begin{aligned}\|\widehat{u}_s^{(t)} - u_t^*\|^2 \ \leq \ & \tfrac{1}{1+2\mu_h\hat{\eta}_t}\big(1 - \tfrac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\big)\|\widehat{u}_{s-1}^{(t)} - u_t^*\|^2 \\ & - \tfrac{\hat{\eta}_t}{1+2\mu_h\hat{\eta}_t}\big(\tfrac{2}{L_u+\mu_H} - \hat{\eta}_t\big)\|\nabla\varphi(\widehat{u}_{s-1}^{(t)}) - \nabla\varphi(u_t^*)\|^2.\end{aligned}$$

Therefore, if $0 < \hat{\eta}_t \leq \frac{2}{L_u+\mu_H}$, then the last inequality reduces to

$$\|\widehat{u}_s^{(t)} - u_t^*\|^2 \ \leq \ \tfrac{1}{1+2\mu_h\hat{\eta}_t}\big(1 - \tfrac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\big)\|\widehat{u}_{s-1}^{(t)} - u_t^*\|^2.$$

By induction, and noting that $\widehat{u}_0^{(t)} := \widetilde{u}_{t-1}$ and $\widetilde{u}_t := \widehat{u}_S^{(t)}$, this inequality implies (58). □

(b) Key bound for the shuffling gradient-ascent scheme (27). Alternatively, if the *full-shuffling variant* (27) is used in Algorithm 2, then we can bound $\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2$ for (27) as follows.

First, let us define $u_0^{s*} := u_0^*(\widetilde{w}_{t-1})$ and for all $i \in [n]$:

$$u_i^{s*} \ := \ u_0^*(\widetilde{w}_{t-1}) + \tfrac{\hat{\eta}_t}{n}\textstyle\sum_{j=1}^i \nabla_u\mathcal{H}_{\pi^{(s,t)}(j)}(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1})). \tag{59}$$

Here, $\nabla_u\mathcal{H}_i$ is the partial derivative (or the gradient) of $\mathcal{H}_i$ w.r.t. u.

Next, we prove the following lemma.

Lemma 12. *Suppose that Assumption 4 holds, and u_i^{s*} is defined by (59) for all $i = 0, \cdots, n$. Then*

$$\begin{aligned}\|u_i^{s*} - u_0^*(\widetilde{w}_{t-1})\|^2 \ &\leq \ \tfrac{2\hat{\eta}_t^2\cdot i}{n}\cdot\big(\Theta_u\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big) + \tfrac{2\hat{\eta}_t^2\cdot i^2}{n^2}\cdot\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 \\ &\leq \ 2\hat{\eta}_t^2\big[(\Theta_u+1)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big].\end{aligned} \tag{60}$$

Proof. For simplicity, we denote $u_t^* := u_0^*(\widetilde{w}_{t-1})$. For $i = 0$, we obviously have $\|u_0^{s*} - u_t^*\|^2 = 0$, showing that (60) trivially holds.

Next, for $i \in [n]$, using u_i^{s*} from (59) and Young's inequality twice in ① and ②, we can derive that

$$\begin{aligned}\|u_i^{s*} - u_t^*\|^2 &= \tfrac{\hat{\eta}_t^2}{n^2}\|\textstyle\sum_{j=1}^i \nabla_u\mathcal{H}_{\pi^{(s)}(j)}(\widetilde{w}_{t-1}, u_t^*)\|^2 \\ &\overset{①}{\leq} \tfrac{2\hat{\eta}_t^2}{n^2}\cdot i^2\cdot\|\tfrac{1}{i}\textstyle\sum_{j=1}^i\big[\nabla_u\mathcal{H}_{\pi^{(s)}(j)}(\widetilde{w}_{t-1}, u_t^*) - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, u_t^*)\big]\|^2 \\ &\quad + \tfrac{2\hat{\eta}_t^2}{n^2}\cdot i^2\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1}, u_t^*)\|^2 \\ &\overset{②}{\leq} \tfrac{2i\hat{\eta}_t^2}{n^2}\textstyle\sum_{j=1}^i\big\|\nabla_u\mathcal{H}_{\pi^{(s)}(j)}(\widetilde{w}_{t-1}, u_t^*) - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, u_t^*)\|^2 + \tfrac{2i^2\hat{\eta}_t^2}{n^2}\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1}, u_t^*)\|^2 \\ &\overset{(5)}{\leq} \tfrac{2i\hat{\eta}_t^2}{n^2}\textstyle\sum_{j=1}^n\big\|\nabla_u\mathcal{H}_{\pi^{(s)}(j)}(\widetilde{w}_{t-1}, u_t^*) - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, u_t^*)\|^2 + \tfrac{2i^2\hat{\eta}_t^2}{n^2}\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2.\end{aligned}$$

By (13) from Assumption 4 and (5), we have

$$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{j=1}^n\big\|\nabla_u\mathcal{H}_{\pi^{(s)}(j)}(\widetilde{w}_{t-1}, u_t^*) - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, u_t^*)\|^2 \ &\overset{(13)}{\leq} \ \Theta_u\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1}, u_t^*)\|^2 + \sigma_u^2 \\ &\overset{(5)}{=} \ \Theta_u\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2.\end{aligned}$$

Combining the last two inequalities, and noting that $0 \leq i \leq n$, we obtain (60). □

Finally, we can prove the necessary bound for $\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2$. For simplicity of our proof, let us denote $g_{i-1}^{s,t}(\cdot) := -\mathcal{H}_{\pi^{(s)}(i)}(\widetilde{w}_{t-1}, \cdot)$ and again $u_t^* := u_0^*(\widetilde{w}_{t-1})$. By Assumption 4(a) and (b), it is clear that $g_{i-1}^{s,t}(\cdot)$ is μ_H-strongly convex and L_u-smooth. Let us consider the following the Bregman distance constructed from $g_{i-1}^{s,t}$:

$$D_{i-1}^{s,t}(u, \hat{u}) = g_{i-1}^{s,t}(u) - g_{i-1}^{s,t}(\hat{u}) - \langle\nabla_u g_{i-1}^{s,t}(\hat{u}), u - \hat{u}\rangle. \tag{61}$$

The following lemma is adapted from Theorems 2 and 3 in [5] with some modification.

Lemma 13. *Suppose that Assumption 4 holds. Let u_i^{s*} be defined by (59), $\{u_i^{(s,t)}\}$ be updated by (27) at the s-th epoch for all $i \in [n]$, and $D_{i-1}^{s,t}$ be defined by (61). Then, it holds that*

$$\begin{aligned}\|u_i^{(s,t)} - u_i^{s*}\|^2 \ \le\ & \big(1 - \tfrac{\mu_H \hat{\eta}_t}{n}\big)\|u_{i-1}^{(s,t)} - u_{i-1}^{s*}\|^2 + \tfrac{2L_u\hat{\eta}_t^3}{n}\big[(\Theta_u+1)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big] \\ & - \tfrac{2\hat{\eta}_t}{n}\big(1 - \tfrac{L_u\hat{\eta}_t}{n}\big) D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_0^*(\widetilde{w}_{t-1})).\end{aligned} \tag{62}$$

Consequently, at each epoch s, the following bound holds:

$$\begin{aligned}\|\widehat{u}_s^{(t)} - u_0^*(\widetilde{w}_{t-1})\|^2 \ \le\ & \tfrac{1}{1+2\mu_h\hat{\eta}_t}\big(1 - \tfrac{\mu_H\hat{\eta}_t}{n}\big)^n \|\widehat{u}_{s-1}^{(t)} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ & + \tfrac{2L_u\cdot\hat{\eta}_t^3}{n(1+2\mu_h\hat{\eta}_t)}\big[\textstyle\sum_{j=0}^{n-1}\big(1 - \tfrac{\mu_H\hat{\eta}_t}{n}\big)^j\big]\big[(\Theta_u+1)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big].\end{aligned} \tag{63}$$

If we update (27) by S epochs starting from $\widehat{u}_0^{(t)} := \widetilde{u}_{t-1}$ and output $\widetilde{u}_t := \widehat{u}_S^{(t)}$, then

$$\begin{aligned}\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \ \le\ & \tfrac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1 - \tfrac{\mu_H\hat{\eta}_t}{n}\big)^{nS}\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ & + \tfrac{2L_u}{n}C_S\hat{\eta}_t^3 \cdot \big[(\Theta_u+1)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big],\end{aligned} \tag{64}$$

where $C_S := \big[\sum_{j=0}^{n-1}\frac{1}{(1+2\mu_h\hat{\eta}_t)}\big(1 - \frac{\mu_H\hat{\eta}_t}{n}\big)^j\big]\sum_{s=0}^{S-1}\frac{1}{(1+2\mu_h\hat{\eta}_t)^s}\big(1 - \frac{\mu_H\hat{\eta}_t}{n}\big)^{ns}$.

Proof. By (27), using the definition of $g_{i-1}^{s,t}(\cdot)$ above, and u_i^{s*} defined by (59), we have

$$\begin{aligned} u_i^{(s,t)} &= u_0^{(s,t)} + \tfrac{\hat{\eta}_t}{n}\textstyle\sum_{j=1}^i \nabla_u \mathcal{H}_{\pi^{(s)}(j)}(\widetilde{w}_{t-1}, u_{j-1}^{(s,t)}) = u_0^{(s,t)} - \tfrac{\hat{\eta}_t}{n}\sum_{j=1}^i \nabla_u g_{j-1}^{s,t}(u_{j-1}^{(s,t)}) \\ &= u_{i-1}^{(s,t)} - \tfrac{\hat{\eta}_t}{n}\nabla_u g_{i-1}^{s,t}(u_{i-1}^{(s,t)}) \\ u_i^{s*} &= u_0^*(\widetilde{w}_{t-1}) - \tfrac{\hat{\eta}_t}{n}\textstyle\sum_{j=1}^i \nabla_u g_{j-1}^{s,t}(u_t^*) = u_{i-1}^{s*} - \tfrac{\hat{\eta}_t}{n}\nabla_u g_{i-1}^{s,t}(u_t^*).\end{aligned}$$

Using these expressions, for any $i \in [n]$, we can show that

$$\begin{aligned}\|u_i^{(s,t)} - u_i^{s*}\|^2 \ =\ & \|u_{i-1}^{(s,t)} - u_{i-1}^{s*}\|^2 - \tfrac{2\hat{\eta}_t}{n}\langle \nabla_u g_{i-1}^{s,t}(u_{i-1}^{(s,t)}) - \nabla_u g_{i-1}^{s,t}(u_t^*), u_{i-1}^{(s,t)} - u_{i-1}^{s*}\rangle \\ & + \tfrac{\hat{\eta}_t^2}{n^2}\|\nabla_u g_{i-1}^{s,t}(u_{i-1}^{(s,t)}) - \nabla_u g_{i-1}^{s,t}(u_t^*)\|^2.\end{aligned} \tag{65}$$

By the L-smoothness condition (12) of $\mathcal{H}_i$ from Assumption 4, we have

$$\|\nabla_u g_{i-1}^{s,t}(u_{i-1}^{(s,t)}) - \nabla_u g_{i-1}^{s,t}(u_t^*)\|^2 \ \le\ 2L_u D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_t^*). \tag{66}$$

By the well-known three-point identity, see, e.g., [2], we have

$$\begin{aligned}\langle \nabla_u g_{i-1}^{s,t}(u_{i-1}^{(s,t)}) - \nabla_u g_{i-1}^{s,t}(u_t^*), u_{i-1}^{(s,t)} - u_{i-1}^{s*}\rangle \ =\ & D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_{i-1}^{s*}) + D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_t^*) \\ & - D_{i-1}^{s,t}(u_{i-1}^{s*}, u_t^*).\end{aligned}$$

Substituting this inequality and (66) into (65), we can show that

$$\begin{aligned}\|u_i^{(s.t)} - u_i^{s*}\|^2 \ \le\ & \|u_{i-1}^{(s,t)} - u_{i-1}^{s*}\|^2 - \tfrac{2\hat{\eta}_t}{n}D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_{i-1}^{s*}) + \tfrac{2\hat{\eta}_t}{n}D_{i-1}^{s,t}(u_{i-1}^{s*}, u_t^*) \\ & - \tfrac{2\hat{\eta}_t}{n}\big(1 - \tfrac{L_u\hat{\eta}_t}{n}\big)D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_t^*).\end{aligned}$$

On the one hand, by the μ_H-strong-convexity of $g_{i-1}^{s,t}$, we have $D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_{i-1}^{s*}) \ge \frac{\mu_H}{2}\|u_{i-1}^{(s,t)} - u_{i-1}^{s*}\|^2$. On the other hand, by the L_u-smoothness of $g_{i-1}^{s,t}$, we also have $D_{i-1}^{s,t}(u_{i-1}^{s*}, u_t^*) \le \frac{L_u}{2}\|u_{i-1}^{s*} - u_t^*\|^2$. Using these bounds into the last inequality, we can show that

$$\begin{aligned}\|u_i^{(s,t)} - u_i^{s*}\|^2 \ \le\ & \big(1 - \tfrac{\mu_H\hat{\eta}_t}{n}\big)\|u_{i-1}^{(s,t)} - u_{i-1}^{s*}\|^2 + \tfrac{L_u\hat{\eta}_t}{n}\|u_{i-1}^{s*} - u_t^*\|^2 \\ & - \tfrac{2\hat{\eta}_t}{n}\big(1 - \tfrac{L_u\hat{\eta}_t}{n}\big)D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_t^*).\end{aligned}$$

Combining this inequality and (60), we obtain (62).

Next, since $1 - \frac{L_u\hat{\eta}_t}{n} \ge 0$ and $D_{i-1}^{s,t}(u_{i-1}^{(s,t)}, u_t^*) \ge 0$, we obtain from (62) that

$$\|u_i^{(s,t)} - u_i^{s*}\|^2 \ \le\ \big(1 - \tfrac{\mu_H\hat{\eta}_t}{n}\big)\|u_{i-1}^{(s,t)} - u_{i-1}^{s*}\|^2 + 2n^2L_u\big[(\Theta_u+1)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big]\cdot\hat{\eta}_t^3.$$

By induction, rolling this inequality from $i = 1$ to n, we have

$$\begin{aligned}\|u_n^{(s,t)} - u_n^{s*}\|^2 \le & \left(1 - \tfrac{\mu_H \hat{\eta}_t}{n}\right)^n \|u_0^{(s,t)} - u_0^{s*}\|^2 \\ & + \tfrac{2L_u}{n}\big[(\Theta_u + 1)\|\nabla \Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big] \cdot \hat{\eta}_t^3 \cdot \textstyle\sum_{j=0}^{n-1}\left(1 - \tfrac{\mu_H \hat{\eta}_t}{n}\right)^j.\end{aligned} \tag{67}$$

Next, from (59) and (33), it is not hard to show that $u_t^* = u_0^*(\widetilde{w}_{t-1}) = \mathrm{prox}_{\hat{\eta}_t h}(u_n^{s*})$. Furthermore, by the second line of (27), we also have $\widehat{u}_s^{(t)} = \mathrm{prox}_{\hat{\eta}_t h}(u_n^{(s,t)})$. Since h is μ_h-strongly convex, by (32) from Fact $[F_1]$, we can show that

$$\|\widehat{u}_s^{(t)} - u_0^*(\widetilde{w}_{t-1})\|^2 = \|\mathrm{prox}_{\hat{\eta}_t h}(u_n^{(s,t)}) - \mathrm{prox}_{\hat{\eta}_t h}(u_n^{s*})\|^2 \le \tfrac{1}{1+2\mu_h \hat{\eta}_t}\|u_n^{(s,t)} - u_n^{s*}\|^2.$$

Using this inequality, $u_0^{(s,t)} = \widehat{u}_{s-1}^{(t)}$, and $u_0^{s*} = u_t^* = u_0^*(\widetilde{w}_{t-1})$, it follows from (67) that

$$\begin{aligned}\|\widehat{u}_s^{(t)} - u_0^*(\widetilde{w}_{t-1})\|^2 \le & \tfrac{1}{1+2\mu_h\hat{\eta}_t}\left(1 - \tfrac{\mu_H \hat{\eta}_t}{n}\right)^n \|\widehat{u}_{s-1}^{(t)} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ & + \tfrac{2L_u \hat{\eta}_t^3}{n(1+2\mu_h \hat{\eta}_t)} \cdot \big[\textstyle\sum_{j=0}^{n-1}\left(1 - \tfrac{\mu_H \hat{\eta}_t}{n}\right)^j\big] \cdot \big[(\Theta_u + 1)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big],\end{aligned}$$

which proves (63).

Next, rolling (63) from $s = 1$ to S, we have

$$\begin{aligned}\|\widehat{u}_s^{(t)} - u_0^*(\widetilde{w}_{t-1})\|^2 \le & \tfrac{1}{(1+2\mu_h\hat{\eta}_t)^S}\left(1 - \tfrac{\mu_H \hat{\eta}_t}{n}\right)^{nS}\|\widehat{u}_0^{(t)} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ & + \tfrac{2L_u}{n} C_S \hat{\eta}_t^3 \cdot \big[(\Theta_u+1)\|\nabla \Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_u^2\big],\end{aligned}$$

where $C_S := \big[\sum_{j=0}^{n-1} \frac{1}{(1+2\mu_h\hat{\eta}_t)}\left(1 - \frac{\mu_H\hat{\eta}_t}{n}\right)^j\big]\sum_{s=0}^{S-1}\frac{1}{(1+2\mu_h\hat{\eta}_t)^s}\left(1 - \frac{\mu_H \hat{\eta}_t}{n}\right)^{ns}$. Substituting $\widehat{u}_0^{(t)} := \widetilde{u}_{t-1}$ and $\widetilde{u}_t := \widehat{u}_S^{(t)}$ into the last inequality, it proves (64). □

(c) Key bounds for the shuffling gradient descent scheme (28). We define the following quantity:

$$g_t := \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \nabla_w \mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t). \tag{68}$$

From the update of $w_i^{(t)}$ in (28), for any $i \in [n]$, we have

$$w_i^{(t)} = w_0^{(t)} - \tfrac{\eta_t}{n}\textstyle\sum_{j=1}^{i} \nabla_w \mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t). \tag{69}$$

Then, it is obvious that $w_n^{(t)} = w_0^{(t)} - \eta_t g_t$.

First, we bound $\Delta_t := \frac{1}{n}\sum_{i=0}^{n-1}\|w_i^{(t)} - w_0^{(t)}\|^2$ for (28) to handle the upper-level problem (3).

Lemma 14. *Suppose that Assumption 4 holds. Let $\{w_i^{(t)}\}$ be generated by (28) such that $w_0^{(t)} := \widetilde{w}_{t-1}$. Then, if we choose $\eta_t > 0$ such that $1 - 3L_w^2\eta_t^2 \ge 0$, then*

$$\begin{aligned}\Delta_t := \tfrac{1}{n}\textstyle\sum_{i=0}^{n-1}\|w_i^{(t)} - w_0^{(t)}\|^2 \le & 2(3\Theta_w + 1)\eta_t^2\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + 6\eta_t^2\sigma_w^2 \\ & + 4L_u^2\eta_t^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2.\end{aligned} \tag{70}$$

Let g_t be defined by (68) and Φ_0 be defined by (2). Then, we have

$$\begin{aligned}\|g_t - \nabla_w\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 & \le \tfrac{L_w^2}{n}\textstyle\sum_{i=0}^{n-1}\|w_i^{(t)} - w_0^{(t)}\|^2 \equiv L_w^2\Delta_t, \\ \|\nabla_w\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_t) - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 & \le L_u^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2, \\ \|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 & \le L_w^2\Delta_t + L_u^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2.\end{aligned} \tag{71}$$

Proof. Utilizing (69) and Young's inequality in ① and ② below, we can show that

$$\begin{aligned}\|w_i^{(t)} - w_0^{(t)}\|^2 & \overset{(69)}{=} \tfrac{i^2\cdot\eta_t^2}{n^2}\big\|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)},\widetilde{u}_t)\big\|^2 \\ & \overset{①}{\le} \tfrac{3i^2\cdot\eta_t^2}{n^2}\big\|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\big[\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)},\widetilde{u}_t) - \nabla\Phi_0(\widetilde{w}_{t-1})\big]\big\|^2 + \tfrac{3i^2\cdot\eta_t^2}{n^2}\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 \\ & \quad + \tfrac{3i^2\cdot\eta_t^2}{n^2}\big\|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\big[\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)},\widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)},\widetilde{u}_t)\big]\big\|^2 \\ & \overset{②}{\le} \tfrac{3i^2\cdot\eta_t^2}{n^2}\big\|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\big[\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)},\widetilde{u}_t) - \nabla\Phi_0(\widetilde{w}_{t-1})\big]\big\|^2 + \tfrac{3i^2\cdot\eta_t^2}{n^2}\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 \\ & \quad + \tfrac{3i\cdot\eta_t^2}{n^2}\textstyle\sum_{j=1}^{i}\big\|\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)},\widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)},\widetilde{u}_t)\big\|^2.\end{aligned}$$

Let us denote $\Delta_t := \frac{1}{n}\sum_{j=0}^{n-1}\|w_j^{(t)} - w_0^{(t)}\|^2 = \frac{1}{n}\sum_{j=0}^{n-1}\|w_j^{(t)} - \widetilde{w}_{t-1}\|^2$. Then, by (12) of Assumption 4, we have

$$\frac{1}{n}\sum_{j=1}^{i}\big\|\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)}, \widetilde{u}_t)\big\|^2 \overset{(12)}{\leq} \frac{L_w^2}{n}\sum_{j=1}^{i}\|w_{j-1}^{(t)} - w_0^{(t)}\|^2 \leq L_w^2\Delta_t.$$

Next, by Young's inequality again in ①, $w_0^{(t)} = \widetilde{w}_{t-1}$, and (12) and (13) from Assumption 4, and the fact that $\nabla_w\mathcal{H}(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1})) = \nabla\Phi_0(\widetilde{w}_{t-1})$ from (5), we can show that

$$\begin{aligned}
\mathcal{T}_{[2]} &:= \|\tfrac{1}{i}\textstyle\sum_{j=1}^{i}\big[\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)}, \widetilde{u}_t) - \nabla\Phi_0(\widetilde{w}_{t-1})\big]\big\|^2 \\
&\overset{①}{\leq} \tfrac{2}{i}\textstyle\sum_{j=1}^{i}\big\|\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)}, \widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)}, u_0^*(w_0^{(t)}))\big\|^2 \\
&\quad + \tfrac{2}{i}\textstyle\sum_{j=1}^{i}\big\|\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)}, u_0^*(w_0^{(t)})) - \nabla_w\mathcal{H}(w_0^{(t)}, u_0^*(w_0^{(t)}))\big\|^2 \\
&\overset{(12)}{\leq} \tfrac{2}{i}\textstyle\sum_{j=1}^{n}\|\nabla_w\mathcal{H}_i(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1})) - \nabla_w\mathcal{H}(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\|^2 \\
&\quad + \tfrac{2L_u^2}{i}\textstyle\sum_{j=1}^{i}\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \\
&\overset{(13),(5)}{\leq} 2L_u^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{2n}{i}\big[\Theta_w\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_w^2\big].
\end{aligned}$$

Combining three inequalities above, we arrive at

$$\begin{aligned}
\|w_i^{(t)} - w_0^{(t)}\|^2 &\leq \tfrac{6i^2\cdot L_u^2\eta_t^2}{n^2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{6i\cdot\eta_t^2}{n}\big[\Theta_w\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \sigma_w^2\big] \\
&\quad + \tfrac{3i^2\cdot\eta_t^2}{n^2}\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \tfrac{3i\cdot L_w^2\eta_t^2}{n}\Delta_t \\
&= \tfrac{6i^2\cdot L_u^2\eta_t^2}{n^2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{3i\eta_t^2}{n^2}(2n\Theta_w + i)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 \\
&\quad + \tfrac{6i\cdot\eta_t^2}{n}\sigma_w^2 + \tfrac{3i\cdot L_w^2\eta_t^2}{n}\Delta_t.
\end{aligned}$$

Averaging this inequality from $i = 0$ to $n-1$, we get

$$\begin{aligned}
\Delta_t &:= \tfrac{1}{n}\textstyle\sum_{i=0}^{n-1}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\leq \tfrac{1}{n}\textstyle\sum_{i=0}^{n-1}\big[\tfrac{6i^2\cdot L_u^2\eta_t^2}{n^2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{3i\eta_t^2}{n^2}(2n\Theta_w + i)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2\big] \\
&\quad + \tfrac{1}{n}\textstyle\sum_{i=0}^{n-1}\big[\tfrac{6i\cdot\eta_t^2}{n}\sigma_w^2 + \tfrac{3i\cdot L_w^2\eta_t^2}{n}\Delta_t\big] \\
&\leq 2L_u^2\eta_t^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + (3\Theta_w + 1)\eta_t^2\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + 3\eta_t^2\sigma_w^2 + \tfrac{3L_w^2\eta_t^2}{2}\Delta_t.
\end{aligned}$$

Here, we have used the facts that $\sum_{i=0}^{n-1} i = \frac{n(n-1)}{2} \leq \frac{n^2}{2}$ and $\sum_{i=0}^{n-1} i^2 = \frac{n(n-1)(2n-1)}{6} \leq \frac{n^3}{3}$. Rearranging the last inequality, we obtain (70).

Finally, to prove (71), we proceed as follows. Using (68) and (12) from Assumption 4, we have

$$\begin{aligned}
\|g_t - \nabla_w\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 &\overset{(68)}{=} \big\|\tfrac{1}{n}\textstyle\sum_{j=1}^{n}\big[\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(\widetilde{w}_{t-1}, \widetilde{u}_t)\big]\big\|^2 \\
&\leq \tfrac{1}{n}\textstyle\sum_{j=1}^{n}\big\|\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_0^{(t)}, \widetilde{u}_t)\big\|^2 \\
&\overset{(12)}{\leq} \tfrac{L_w^2}{n}\textstyle\sum_{j=1}^{n}\|w_{j-1}^{(t)} - w_0^{(t)}\|^2,
\end{aligned}$$

which proves the first line of (71).

We also note that $\nabla\Phi_0(\widetilde{w}_{t-1}) = \sum_{j=1}^{n}\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))$ due to (5). Using this expression, and (12) from Assumption 4, we can show that

$$\begin{aligned}
\|\nabla_w\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t) - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 &= \big\|\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\big[\nabla_w\mathcal{H}_i(\widetilde{w}_{t-1}, \widetilde{u}_t) - \nabla_w\mathcal{H}_i(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\big]\big\|^2 \\
&\leq \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\big\|\nabla_w\mathcal{H}_i(\widetilde{w}_{t-1}, \widetilde{u}_t) - \nabla_w\mathcal{H}_i(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\big\|^2 \\
&\overset{(12)}{\leq} \tfrac{L_u^2}{n}\textstyle\sum_{i=1}^{n}\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 = L_u^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2,
\end{aligned}$$

which proves the second line of (71).

Similarly, combining (68), (5), and (12) from Assumption 4, we can show that

$$\begin{aligned}
\|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 &\overset{(68),(5)}{=} \big\|\tfrac{1}{n}\textstyle\sum_{j=1}^{n}\big[\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\big]\big\|^2 \\
&\leq \tfrac{1}{n}\textstyle\sum_{j=1}^{n}\big\|\nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t) - \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\big\|^2 \\
&\overset{(12)}{\leq} \tfrac{1}{n}\textstyle\sum_{j=1}^{n}\big[L_w^2\|w_{j-1}^{(t)} - \widetilde{w}_{t-1}\|^2 + L_u^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2\big],
\end{aligned}$$

which proves the third line of (71). □

Lemma 15. *Suppose that Assumption 4 holds. Let $\{w_i^{(t)}\}$ be generated by (28), g_t be defined by (68), Ψ be defined by (3), and $\mathcal{G}_\eta$ be defined by (18). Then, we have*

$$\begin{aligned}\Psi_0(\widetilde{w}_t) \le\ & \Psi_0(\widetilde{w}_{t-1}) - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\ & + \tfrac{\eta_t}{2}\|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2.\end{aligned} \tag{72}$$

Proof. Let us denote $\widehat{w}_t := \mathrm{prox}_{\eta_t f}\big(\widetilde{w}_{t-1} - \eta_t\nabla\Phi_0(\widetilde{w}_{t-1})\big)$. Then, from (18), we can easily show that $\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1}) = \frac{1}{\eta_t}(\widetilde{w}_{t-1} - \widehat{w}_t)$. Therefore, we have $\nabla f(\widehat{w}_t) := \eta_t^{-1}\big(\widetilde{w}_{t-1} - \widehat{w}_t\big) - \nabla\Phi_0(\widetilde{w}_{t-1}) \in \partial f(\widehat{w}_t)$. By the convexity of f, we have

$$\begin{aligned} f(\widehat{w}_t) &\le f(\widetilde{w}_{t-1}) + \langle \nabla f(\widehat{w}_t), \widehat{w}_t - \widetilde{w}_{t-1}\rangle \\ &= f(\widetilde{w}_{t-1}) - \langle \nabla\Phi_0(\widetilde{w}_{t-1}), \widehat{w}_t - \widetilde{w}_{t-1}\rangle - \tfrac{1}{\eta_t}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2.\end{aligned}$$

Next, by the L_{Φ_0}-smoothness of Φ from (36), we have

$$\Phi_0(\widehat{w}_t) \le \Phi_0(\widetilde{w}_{t-1}) + \langle \nabla\Phi_0(\widetilde{w}_{t-1}), \widehat{w}_t - \widetilde{w}_{t-1}\rangle + \tfrac{L_{\Phi_0}}{2}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2.$$

Adding the last two inequalities together and using $\Psi_0(w) = f(w) + \Phi_0(w)$ from (3) and $\widehat{w}_t - \widetilde{w}_{t-1} = -\eta_t\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})$, we can derive

$$\Psi_0(\widehat{w}_t) \le \Psi_0(\widetilde{w}_{t-1}) - \tfrac{(2-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2 = \Psi_0(\widetilde{w}_{t-1}) - \tfrac{\eta_t(2-L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2. \tag{73}$$

Now, from (69), we have

$$g_t := \tfrac{1}{\eta_t}(\widetilde{w}_{t-1} - w_n^{(t)}) = \tfrac{1}{\eta_t}(w_0^{(t)} - w_n^{(t)}) = \tfrac{1}{n}\textstyle\sum_{j=1}^n \nabla_w\mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \tilde{u}_t). \tag{74}$$

Since $\widetilde{w}_t = \mathrm{prox}_{\eta_t f}(w_n^{(t)})$ from the second line of (28), we get $\nabla f(\widetilde{w}_t) := \eta_t^{-1}\big(w_n^{(t)} - \widetilde{w}_t\big) = -g_t - \eta_t^{-1}(\widetilde{w}_t - \widetilde{w}_{t-1}) \in \partial f(\widetilde{w}_t)$. Hence, again by the convexity of f, we can deduce that

$$\begin{aligned} f(\widetilde{w}_t) &\le f(\widehat{w}^t) + \langle \nabla f(\widetilde{w}_t), \widetilde{w}_t - \widehat{w}^t\rangle = f(\widehat{w}^t) - \langle g_t, \widetilde{w}_t - \widehat{w}^t\rangle - \tfrac{1}{\eta_t}\langle \widetilde{w}_t - \widetilde{w}_{t-1}, \widetilde{w}_t - \widehat{w}^t\rangle \\ &= f(\widehat{w}^t) - \langle g_t, \widetilde{w}_t - \widehat{w}^t\rangle - \tfrac{1}{2\eta_t}\big[\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 + \|\widetilde{w}_t - \widehat{w}^t\|^2 - \|\widehat{w}^t - \widetilde{w}_{t-1}\|^2\big].\end{aligned}$$

Again, by the L_{Φ_0}-smoothness of Φ from (36), we also have

$$\begin{aligned} \Phi_0(\widetilde{w}_t) &\le \Phi_0(\widetilde{w}_{t-1}) + \langle \nabla\Phi_0(\widetilde{w}_{t-1}), \widetilde{w}_t - \widetilde{w}_{t-1}\rangle + \tfrac{L_{\Phi_0}}{2}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2, \\ \Phi_0(\widetilde{w}_{t-1}) &\le \Phi_0(\widehat{w}^t) + \langle \nabla\Phi_0(\widetilde{w}_{t-1}), \widetilde{w}_{t-1} - \widehat{w}^t\rangle + \tfrac{L_{\Phi_0}}{2}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2.\end{aligned}$$

Adding the last three inequalities together, and using $\Psi_0(w) = f(w) + \Phi_0(w)$ from (3) and $\widehat{w}_t - \widetilde{w}_{t-1} = -\eta_t\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})$, we can prove that

$$\begin{aligned}\Psi_0(\widetilde{w}_t) &\le \Psi_0(\widehat{w}^t) + \langle \nabla\Phi_0(\widetilde{w}_{t-1}) - g_t, \widetilde{w}_t - \widehat{w}^t\rangle - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ &\quad + \tfrac{(1+L_{\Phi_0}\eta_t)}{2\eta_t}\|\widehat{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{1}{2\eta_t}\|\widetilde{w}_t - \widehat{w}^t\|^2 \\ &\overset{①}{\le} \Psi_0(\widehat{w}^t) + \tfrac{\eta_t}{2}\|\nabla\Phi_0(\widetilde{w}_{t-1}) - g_t\|^2 - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ &\quad + \tfrac{\eta_t(1+L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2,\end{aligned} \tag{75}$$

where we have used Young's inequality in the last line ① as $\langle \nabla\Phi_0(\widetilde{w}_{t-1}) - g_t, \widetilde{w}_t - \widehat{w}^t\rangle \le \frac{\eta_t}{2}\|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \frac{1}{2\eta_t}\|\widetilde{w}_t - \widehat{w}^t\|^2$.

Finally, summing up (73) and (75) we arrive at

$$\begin{aligned}\Psi_0(\widetilde{w}_t) \le\ & \Psi_0(\widetilde{w}_{t-1}) - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\ & + \tfrac{\eta_t}{2}\|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2,\end{aligned}$$

which proves (72). □

C.2 Convergence of the semi-shuffling variant of Algorithm 2

We now prove the convergence of the semi-shuffling variant of Algorithm 2 using (26).

Lemma 16. *Suppose that Assumptions 4 and 5 hold for (1). Let Ψ be defined by (3) and $\mathcal{G}_\eta$ be defined by (18). Let $\{(\widetilde{w}_t, \widetilde{u}_t)\}$ be generated by the* ***semi-shuffling variant*** *of Algorithm 2 using (26). For a fixed $\omega > 0$, suppose that we choose η_t and $\hat{\eta}_t$ such that $1 - 3L_w^2\eta_t^2 \geq 0$ and $0 < \hat{\eta}_t \leq \frac{2}{L_u + \mu_H}$, and the following conditions hold:*

$$\begin{cases} 2L_{\Phi_0}\eta_t + 2\omega L_u^2\kappa^2\eta_t^2 \leq 1, \\ \frac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1 - \frac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\big)^S\big(1 + \omega + \omega^2L_u^2\kappa^2\eta_t^2 + 2L_w^2\eta_t^2\big) \leq \omega. \end{cases} \tag{76}$$

Then, the following bound holds:

$$\begin{aligned} \Psi_0(\widetilde{w}_t) \ + \ \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 &\leq \Psi_0(\widetilde{w}_{t-1}) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &- \tfrac{\eta_t B_t}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \big[3L_w^2\sigma_w^2 + L_w^2(3\Theta_w+1)\Lambda_1\big]\cdot\eta_t^3, \end{aligned} \tag{77}$$

where $B_t := 1 - 2L_{\Phi_0}\eta_t - 2L_w^2(3\Theta_w+1)\Lambda_0\eta_t^2$.

Proof. First, combining (72) and the last line of (71), we can derive

$$\begin{aligned} \Psi_0(\widetilde{w}_t) \ \leq \ \Psi_0(\widetilde{w}_{t-1}) &- \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ &+ \tfrac{L_w^2\eta_t}{2n}\textstyle\sum_{j=1}^n \|w_{j-1}^{(t)} - \widetilde{w}_{t-1}\|^2 + \tfrac{L_u^2\eta_t}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2. \end{aligned} \tag{78}$$

Next, substituting (70) into (78), we can show that

$$\begin{aligned} \Psi_0(\widetilde{w}_t) \ \leq \ \Psi_0(\widetilde{w}_{t-1}) &- \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 + 3L_w^2\eta_t^3\sigma_w^2 \\ &+ \tfrac{L_u^2\eta_t}{2}\big(1 + 4L_w^2\eta_t^2\big)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + L_w^2(3\Theta_w+1)\eta_t^3\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2. \end{aligned}$$

By (35) and Young's inequality in ①, for any $s_t > 0$, we have

$$\begin{aligned} \|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 &\overset{①}{\leq} (1+s_t)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{(1+s_t)}{s_t}\|u_0^*(\widetilde{w}_t) - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\overset{(35)}{\leq} (1+s_t)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{(1+s_t)\kappa^2}{s_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2. \end{aligned}$$

Multiplying this inequality by $\frac{\omega L_u^2\eta_t}{2}$ for some $\omega > 0$ and adding the result to the last estimate yields

$$\begin{aligned} \mathcal{T}_{[1]} \ &:= \ \Psi_0(\widetilde{w}_t) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 \\ &\leq \ \Psi_0(\widetilde{w}_{t-1}) + \tfrac{L_u^2\eta_t}{2}\big[1 + \omega(1+s_t) + 4L_w^2\eta_t^2\big]\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \big[\tfrac{1-L_{\Phi_0}\eta_t}{2\eta_t} - \tfrac{\omega L_u^2\kappa^2\eta_t(1+s_t)}{2s_t}\big]\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ &\quad + L_w^2(3\Theta_w+1)\eta_t^3\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + 3L_w^2\eta_t^3\sigma_w^2 \\ &\overset{(58)}{\leq} \ \Psi_0(\widetilde{w}_{t-1}) + \tfrac{L_u^2\eta_t}{2(1+2\mu_h\hat{\eta}_t)^S}\big(1 - \tfrac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\big)^S\big[1 + \omega(1+s_t) + 4L_w^2\eta_t^2\big]\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \big[\tfrac{1-L_{\Phi_0}\eta_t}{2\eta_t} - \tfrac{\omega L_u^2\kappa^2\eta_t(1+s_t)}{2s_t}\big]\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ &\quad + L_w^2(3\Theta_w+1)\eta_t^3\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + 3L_w^2\eta_t^3\sigma_w^2. \end{aligned}$$

We need to choose the parameters η_t, $\hat{\eta}_t$, and s_t such that

$$\begin{cases} \frac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1 - \frac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\big)^S\big[1 + \omega(1+s_t) + 4L_w^2\eta_t^2\big] \leq \omega, \\ \frac{1-L_{\Phi_0}\eta_t}{\eta_t} - \frac{\omega L_u^2\kappa^2\eta_t(1+s_t)}{s_t} \geq 0. \end{cases}$$

The second condition leads to $\frac{1-L_{\Phi_0}\eta_t - \omega L_u^2\kappa^2\eta_t^2}{\omega L_u^2\kappa^2\eta_t^2} \geq \frac{1}{s_t}$, or equivalently $0 < s_t \leq \frac{\omega L_u^2\kappa^2\eta_t^2}{1-L_{\Phi_0}\eta_t - \omega L_u^2\kappa^2\eta_t^2}$. If $2L_{\Phi_0}\eta_t + 2\omega L_u^2\kappa^2\eta_t^2 \leq 1$ as stated in the first line of (76), then we can choose $s_t := 2\omega L_u^2\kappa^2\eta_t^2$. In this case, the second condition is satisfied, while the first condition becomes

$$\tfrac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1 - \tfrac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\big)^S\big(1 + \omega + 2\omega^2L_u^2\kappa^2\eta_t^2 + 4L_w^2\eta_t^2\big) \leq \omega,$$

which is exactly the second condition of (76).

By (20) from Assumption 5, we have

$$\begin{aligned} \mathcal{T}_{[1]} &:= \Psi_0(\widetilde{w}_t) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 \\ &\leq \Psi_0(\widetilde{w}_{t-1}) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\ &\quad + L_w^2(3\Theta_w+1)\Lambda_0\eta_t^3\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \big[3L_w^2\sigma_w^2 + L_w^2(3\Theta_w+1)\Lambda_1\big]\eta_t^3. \end{aligned}$$

Rearranging this inequality, we prove (77). □

The following theorem, Theorem 7, is the full version of Theorem 2 in the main text, where the learning rates η_t and $\hat{\eta}_t$, and the numbers of epochs S and T are given explicitly.

Theorem 7. *Suppose that Assumptions 1, 2, 4, and 5 hold for (1). Let Ψ_0 be defined by (3), and $\mathcal{G}_\eta$ be defined by (18). Let C_0 and C_w be two constants given as follows:*

$$C_0 := 2\Lambda_0 L_w^2(3\Theta_w+1) \quad \text{and} \quad C_w := L_w^2(3\Theta_w+1)\Lambda_1 + 3L_w^2\sigma_w^2. \tag{79}$$

*Let $\{(\widetilde{w}_t, \widetilde{u}_t)\}$ be generated by Algorithm 2 using the **gradient ascent scheme** (26), and fixed learning rates $\eta_t := \eta > 0$ and $\hat{\eta}_t := \hat{\eta} \in \big(0, \frac{2}{L_u+\mu_H}\big]$ such that for a fixed $\omega > 0$:*

$$S := \Big\lfloor \tfrac{M_\omega(\eta)}{2\hat{\eta}}\big(\mu_h + \tfrac{4\mu_H L_u}{L_u+\mu_H}\big)^{-1}\Big\rfloor \quad \text{and} \quad 0 < \eta \leq \min\Big\{\tfrac{1}{2\sqrt{C_0}}, \tfrac{1}{4L_{\Phi_0}}, \tfrac{1}{2\omega L_u\kappa}\Big\}, \tag{80}$$

where $M_\omega(\eta) := \frac{1}{\omega} + \big(\omega L_u^2\kappa^2 + \frac{2L_w^2}{\omega}\big)\eta^2$. Then, the following estimate holds:

$$\tfrac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \tfrac{4\big[2(\Psi_0(\widetilde{w}_0)-\Psi_0^\star)+\omega L_u^2\eta\|\widetilde{u}^0-u_0^*(\widetilde{w}_0)\|^2\big]}{\eta(T+1)} + 8C_w\eta^2. \tag{81}$$

For a given $\epsilon > 0$, if we choose $\eta := \frac{s\epsilon}{4\sqrt{C_w}}$ for a fixed $s \in (0,1)$ satisfying (76), $\hat{\eta} \in \big(0, \frac{2}{L_u+\mu_h}\big]$, and $T := \mathcal{O}\big(\frac{1}{\epsilon^3}\big)$, then $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \epsilon^2$.

For a given $\hat{\eta} \in \big(0, \frac{2}{L_u+\mu_h}\big]$, we denote $B_0 := \hat{\eta}\big(\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big)$. If we choose $\omega := \frac{1}{B_0}$ and

$$0 < \eta \leq \min\left\{\tfrac{1}{2\sqrt{C_0}}, \tfrac{1}{4L_{\Phi_0}}, \tfrac{B_0}{\sqrt{L_u^2\kappa^2 + 2B_0^2L_w^2}}\right\}, \tag{82}$$

*then we have $S = 1$, i.e. we only need to perform one iteration of the **gradient ascent scheme** (26). Consequently, Algorithm 2 requires $\mathcal{O}\big(\frac{n}{\epsilon^3}\big)$ evaluations of $\nabla_w\mathcal{H}_i$ and of $\nabla_u\mathcal{H}_i$, and $\mathcal{O}(\epsilon^{-3})$ evaluations of $\mathrm{prox}_{\eta_t f}$ and of $\mathrm{prox}_{\hat{\eta}_t h}$ to achieve an ϵ-stationary point $\widehat{w}_T$ of (1) computed by (19).*

Proof of Theorem 7. Let us choose $\eta_t := \eta$ such that η satisfies (80). Then, it is obvious to verify that $1 - 3L_w^2\eta_t^2 \geq 0$ and $2L_{\Phi_0}\eta_t + 2\omega L_u^2\kappa^2\eta_t^2 \leq 1$. Moreover, we have $\eta_t = \eta \leq \frac{1}{4L_{\Phi_0}}$, $\eta_t = \eta \leq \frac{1}{2L_w} \leq \frac{1}{\sqrt{3}L_w}$, and $\eta_t = \eta \leq \frac{1}{2\omega L_u\kappa}$. Using these bounds, we can further lower bound $B_t := 1 - 2L_{\Phi_0}\eta_t - 2L_w^2(3\Theta_w+1)\Lambda_0\eta_t^2$ from Lemma 16 as

$$B_t \geq \tfrac{1}{2} - 2\Lambda_0 L_w^2(3\Theta_w+1)\eta^2.$$

Now, we need to choose $0 < \eta \leq \frac{1}{2\sqrt{2\Lambda_0 L_w^2(3\Theta_w+1)}} = \frac{1}{2\sqrt{C_0}}$ so that $B_t \geq \frac{1}{4}$, where C_0 is given in (79). Moreover, the second condition of (76) holds if

$$\begin{aligned} \tfrac{1}{(1+\mu_h\hat{\eta})^S}\big(1 - \tfrac{2L_u\mu_H\hat{\eta}}{L_u+\mu_H}\big)^S &\leq \tfrac{\omega}{1+\omega+(\omega^2L_u^2\kappa^2+2L_w^2)\eta^2}, \\ \Leftrightarrow\ S\ln(1+\mu_h\hat{\eta}) - S\ln\big(1 - \tfrac{2L_u\mu_H\hat{\eta}}{L_u+\mu_H}\big) &\geq \ln\big(1 + \tfrac{1}{\omega} + (\omega L_u^2\kappa^2 + \tfrac{2L_w^2}{\omega})\eta^2\big). \end{aligned}$$

Using the elementary facts $-\ln(1-\tau) \geq \tau$ and $\tau \geq \ln(1+\tau) \geq \frac{e}{2}$ for all $\tau \in (0, 1/2]$, we can show that the last inequality holds if

$$S\hat{\eta}\big(\tfrac{\mu_h}{2} + \tfrac{2L_u\mu_H}{L_u+\mu_H}\big) \geq \tfrac{1}{\omega} + (\omega L_u^2\kappa^2 + \tfrac{2L_w^2}{\omega})\eta^2.$$

Simplifying this condition, we get

$$S \geq \frac{M_\omega(\eta)}{2\hat{\eta}}\big(\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big)^{-1}, \quad \text{where} \quad M_\omega(\eta) := \frac{1}{\omega} + \big(\omega L_u^2\kappa^2 + \frac{2L_w^2}{\omega}\big)\eta^2.$$

Clearly, this leads to the choice of S as in (80).

Let us define $C_w := 3L_w^2\sigma_w^2 + L_w^2(3\Theta_w + 1)\Lambda_1$ as in (79). Then, (77) reduces to

$$\begin{aligned}\Psi_0(\widetilde{w}_t) + \tfrac{\omega L_u^2\eta}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 \ &\leq \ \Psi_0(\widetilde{w}_{t-1}) + \tfrac{\omega L_u^2\eta}{2}\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{\eta}{8}\|\mathcal{G}_\eta(\widetilde{w}_{t-1})\|^2 + C_w\cdot\eta^3.\end{aligned}$$

Subtracting $\Psi_0^\star$ from both sides of this inequality, and averaging the result from $t = 0$ to $t = T$, and noting that $\Psi_0(\widetilde{w}_T) - \Psi_0^\star \geq 0$, we obtain (81).

Without loss of generality, let us choose $\omega := 1$. We also choose $\hat{\eta} \in \big(0, \frac{2}{L_u+\mu_h}\big]$. Then, we have $M_\omega = 1 + (L_u^2\kappa^2 + 2L_w^2)\eta^2 \leq 2$. Moreover, from (80), we also have

$$S = \Big\lfloor \frac{M_\omega(\eta)}{2\hat{\eta}}\big[\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big]^{-1}\Big\rfloor \leq \bar{S} := \Big\lfloor \frac{1}{\hat{\eta}}\big[\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big]^{-1}\Big\rfloor = \mathcal{O}(1).$$

To achieve an ϵ-stationary point of (3), from (81) we need to impose the following condition:

$$\frac{2[\Psi_0(\widetilde{w}_0) - \Psi_0^\star]}{\eta(T+1)} + \frac{L_u^2\|\widetilde{u}^0 - u_0^*(\widetilde{w}_0)\|^2}{T+1} + 4C_w\eta^2 \leq \frac{\epsilon^2}{4}.$$

If we choose $\eta := \frac{s\epsilon}{4\sqrt{C_w}}$ for some $s \in (0,1)$ satisfying (80), then the last inequality leads to

$$T \geq \bar{T} := \Big\lfloor \frac{32\sqrt{C_w}[\Psi_0(\widetilde{w}_0) - \Psi_0^\star]}{s(1-s^2)\epsilon^3} + \frac{4L_u^2\|\widetilde{u}^0 - u_0^*(\widetilde{w}_0)\|^2}{(1-s^2)\epsilon^2}\Big\rfloor = \mathcal{O}\big(\frac{1}{\epsilon^3}\big).$$

Therefore, we can choose $T := \bar{T} = \mathcal{O}(\epsilon^{-3})$. Since each iteration t, we run S epochs of the shuffling scheme (26), the total number of evaluations of $\nabla_u\mathcal{H}_i$ is $\mathcal{T}_u := T \times S \times n$. However, since $1 \leq S \leq \bar{S} = \mathcal{O}(1)$, we get $\mathcal{T}_u := \mathcal{O}(n\epsilon^{-3})$. The total number of evaluations of $\nabla_w\mathcal{H}_i$ is $\mathcal{T}_w := Tn = \mathcal{O}(n\epsilon^{-3})$ as stated.

Since each epoch t, Algorithm 2 requires one evaluation of $\mathrm{prox}_{\eta_t f}$, and S evaluations of $\mathrm{prox}_{\hat{\eta}_t h}$, but since $S = \mathcal{O}(1)$, the total number of $\mathrm{prox}_{\eta_t f}$ evaluations is $T = \mathcal{O}(\epsilon^{-3})$, while the total number of $\mathrm{prox}_{\hat{\eta}_t h}$ evaluations is $TS = \mathcal{O}(\epsilon^{-3})$. Overall, Algorithm 2 needs $\mathcal{O}(\epsilon^{-3})$ evaluations of both $\mathrm{prox}_{\eta_t f}$ and $\mathrm{prox}_{\hat{\eta}_t h}$.

Finally, to perform only one iteration the **gradient ascent scheme** (26) at each epoch t, we need to choose ω such that

$$\frac{M_\omega(\eta)}{2\hat{\eta}}\big(\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big)^{-1} \leq S = 1.$$

This condition leads to

$$M_\omega(\eta) = \frac{1}{\omega} + \big(\omega L_u^2\kappa^2 + \frac{2L_w^2}{\omega}\big)\eta^2 \leq 2\hat{\eta}\big(\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big).$$

For a given $\hat{\eta} \in \big(0, \frac{2}{L_u+\mu_h}\big]$, let us choose $\frac{1}{\omega} := \hat{\eta}\big(\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big) := B_0$. Then, the last condition becomes $(L_u^2\kappa^2 + 2B_0^2L_w^2)\eta^2 \leq B_0^2$, or equivalently $\eta \leq \frac{B_0}{\sqrt{L_u^2\kappa^2 + 2B_0^2L_w^2}}$. Combining this condition and (80), we get (82), i.e.:

$$0 < \eta \leq \min\left\{\frac{1}{2\sqrt{C_0}}, \frac{1}{4L_{\Phi_0}}, \frac{B_0}{\sqrt{L_u^2\kappa^2 + 2B_0^2L_w^2}}\right\}, \quad \text{where} \quad B_0 := \hat{\eta}\big(\mu_h + \frac{4\mu_H L_u}{L_u+\mu_H}\big).$$

Thus we have $S = 1$, i.e. we need to perform only one iteration of (26) per epoch t. □

C.3 Convergence of the full-shuffling variant of Algorithm 2 – The case $S > 1$

We can combine the results above to obtain the following lemma.

Lemma 17. *Suppose that Assumptions 4 and 5 hold for (1), Ψ be defined by (3), and $\mathcal{G}_\eta$ be defined by (18). Let $\{(\widetilde{w}_t, \widetilde{u}_t)\}$ be generated by the full-shuffling variant of Algorithm 2 using (27). For a fixed $\omega > 0$, assume that η_t and $\hat{\eta}_t$ are chosen such that $1 - 3L_w^2\eta_t^2 \geq 0$ and*

$$\left\{\begin{array}{l} 2L_{\Phi_0}\eta_t + 2\omega L_u^2\kappa^2\eta_t^2 \leq 1, \\ \frac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1 - \frac{\mu_H\hat{\eta}_t}{n}\big)^{nS}\big(1 + \omega + 2\omega^2L_u^2\kappa^2\eta_t^2 + 4L_w^2\eta_t^2\big) \leq \omega. \end{array}\right. \tag{83}$$

Then, the following bound holds:

$$\begin{aligned}\Psi_0(\widetilde{w}_t) \ + \ \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 &\leq \Psi_0(\widetilde{w}_{t-1}) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &- \tfrac{\eta_t B_t}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \big[3L_w^2\sigma_w^2 + L_w^2(3\Theta_w+1)\Lambda_1\big]\cdot\eta_t^3 \\ &+ \tfrac{L_u^3}{n}(1+\omega+2\omega^2L_u^2\kappa^2\eta_t^2+4L_w^2\eta_t^2)\big[\Lambda_1(\Theta_u+1)+\sigma_u^2\big]\cdot C_S\eta_t\hat{\eta}_t^3,\end{aligned} \tag{84}$$

where $B_t := 1 - 2L_{\Phi_0}\eta_t - 2L_w^2(3\Theta_w+1)\Lambda_0\eta_t^2 - \frac{2L_u^3}{n}(1+\omega+2\omega^2L_u^2\kappa^2\eta_t^2+4L_w^2\eta_t^2)\hat{\eta}_t^3\cdot C_S\cdot\Lambda_0(\Theta_u+1)$ *for* C_S *given in Lemma 13.*

Proof. First, combining (78) and (70), we get

$$\begin{aligned}\Psi_0(\widetilde{w}_t) \ &\leq \ \Psi_0(\widetilde{w}_{t-1}) - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 + 3L_w^2\sigma_w^2\eta_t^3 \\ &+ \tfrac{L_u^2\eta_t}{2}\big(1+4L_w^2\eta_t^2\big)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + L_w^2(3\Theta_w+1)\eta_t^3\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2.\end{aligned}$$

By (35) and Young's inequality in ①, for any $s_t > 0$, we have

$$\begin{aligned}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 \ &\overset{①}{\leq} \ (1+s_t)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{(1+s_t)}{s_t}\|u_0^*(\widetilde{w}_t) - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\overset{(35)}{\leq} \ (1+s_t)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + \tfrac{(1+s_t)\kappa^2}{s_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2.\end{aligned}$$

Multiplying this inequality by $\frac{\omega L_u^2\eta_t}{2}$ for some $\omega > 0$ and adding the result to the last estimate, we can show that

$$\begin{aligned}\mathcal{T}_{[1]} \ &:= \ \Psi_0(\widetilde{w}_t) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 \\ &\leq \ \Psi_0(\widetilde{w}_{t-1}) + \tfrac{L_u^2\eta_t}{2}\big[1+\omega(1+s_t)+4L_w^2\eta_t^2\big)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \big[\tfrac{1-L_{\Phi_0}\eta_t}{2\eta_t} - \tfrac{\omega L_u^2\kappa^2\eta_t(1+s_t)}{2s_t}\big]\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ &\quad + L_w^2(3\Theta_w+1)\eta_t^3\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + 3L_w^2\eta_t^3\sigma_w^2 \\ &\overset{(64)}{\leq} \ \Psi_0(\widetilde{w}_{t-1}) + \tfrac{L_u^2\eta_t}{2(1+2\mu_h\hat{\eta}_t)^S}\big(1-\tfrac{\mu_H\hat{\eta}_t}{n}\big)^{nS}\big[1+\omega(1+s_t)+4L_w^2\eta_t^2\big]\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \big[\tfrac{1-L_{\Phi_0}\eta_t}{2\eta_t} - \tfrac{\omega L_u^2\kappa^2\eta_t(1+s_t)}{2s_t}\big]\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ &\quad + L_w^2(3\Theta_w+1)\eta_t^3\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + 3L_w^2\eta_t^3\sigma_w^2 \\ &\quad + \tfrac{L_u^3}{n}\big[1+\omega(1+s_t)+4L_w^2\eta_t^2\big]\eta_t\hat{\eta}_t^3\cdot C_S\cdot\big[(\Theta_u+1)\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2+\sigma_u^2\big].\end{aligned}$$

We need to choose the parameters η_t, $\hat{\eta}_t$, and s_t such that

$$\begin{aligned}&\tfrac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1-\tfrac{\mu_H\hat{\eta}_t}{n}\big)^{Sn}\big[1+\omega(1+s_t)+4L_w^2\eta_t^2\big] \leq \omega \\ &\tfrac{1-L_{\Phi_0}\eta_t}{2\eta_t} - \tfrac{\omega L_u^2\kappa^2\eta_t(1+s_t)}{2s_t} \geq 0.\end{aligned}$$

The second one leads to $\frac{1-L_{\Phi_0}\eta_t-\omega L_u^2\kappa^2\eta_t^2}{\omega L_u^2\kappa^2\eta_t^2} \geq \frac{1}{s_t}$, or equivalently $0 < s_t \leq \frac{\omega L_u^2\kappa^2\eta_t^2}{1-L_{\Phi_0}\eta_t-\omega L_u^2\kappa^2\eta_t^2}$. If $2L_{\Phi_0}\eta_t + 2\omega L_u^2\kappa^2\eta_t^2 \leq 1$ as stated in the first line of (83), then we can choose $s_t := 2\omega L_u^2\kappa^2\eta_t^2$. In this case, the second condition above holds, and the first condition becomes

$$\tfrac{1}{(1+2\mu_h\hat{\eta}_t)^S}\big(1-\tfrac{\mu_H\hat{\eta}_t}{n}\big)^{nS}\big(1+\omega+2\omega^2L_u^2\kappa^2\eta_t^2+4L_w^2\eta_t^2\big) \leq \omega,$$

which is exactly the second line of (83).

By (20) from Assumption 5, we have

$$\begin{aligned}\mathcal{T}_{[1]} \ &:= \ \Psi_0(\widetilde{w}_t) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_t - u^*(\widetilde{w}_t)\|^2 \\ &\leq \ \Psi_0(\widetilde{w}_{t-1}) + \tfrac{\omega L_u^2\eta_t}{2}\|\widetilde{u}_{t-1} - u^*(\widetilde{w}_{t-1})\|^2 - \tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\ &\quad + L_w^2(3\Theta_w+1)\Lambda_0\eta_t^3\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \big[3L_w^2\sigma_w^2 + L_w^2(3\Theta_w+1)\Lambda_1\big]\eta_t^3 \\ &\quad + \tfrac{L_u^3}{n}(1+\omega+2\omega^2L_u^2\kappa^2\eta_t^2+4L_w^2\eta_t^2)\eta_t\hat{\eta}_t^3\cdot C_S\cdot\Lambda_0(\Theta_u+1)\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\ &\quad + \tfrac{L_u^3}{n}(1+\omega+2\omega^2L_u^2\kappa^2\eta_t^2+4L_w^2\eta_t^2)\eta_t\hat{\eta}_t^3\cdot C_S\big[\Lambda_1(\Theta_u+1)+\sigma_u^2\big].\end{aligned}$$

Rearranging this inequality, we prove (84). □

The following theorem, Theorem 8, is the full version of Theorem 3 in the main text, where the learning rates η_t and $\hat{\eta}_t$, and the numbers of epochs S and T are given explicitly.

Theorem 8 (Strong convexity of $\mathcal{H}_i$)**.** *Suppose that Assumptions 1, 2, 4, and 5 hold for (1), and $\mathcal{H}_i$ is μ_H-strongly concave with $\mu_H > 0$ for all $i \in [n]$, but h is only merely convex. Let Ψ_0 be defined by (3), and $\mathcal{G}_\eta$ be defined by (18). We define C_w and C_u respectively as*

$$C_w := L_w^2\big[(3\Theta_w + 1)\Lambda_1 + 3\sigma_w^2\big] \quad \textit{and} \quad C_u := \tfrac{7L_u^3}{2\mu_H}\big[\Lambda_1(\Theta_u + 1) + \sigma_u^2\big]. \tag{85}$$

Let $\{(\widetilde{w}_t, \widetilde{u}_t)\}$ be generated by Algorithm 2 using S epochs of ***shuffling routine*** *(27), and fixed learning rates $\eta_t := \eta > 0$ and $\hat{\eta}_t := \hat{\eta}$ such that*

$$S := \big\lfloor \tfrac{\ln(7/2)}{\mu_H\hat{\eta}} \big\rfloor, \qquad 0 < \eta \leq \bar{\eta}, \quad \textit{and} \quad 0 < \hat{\eta} \leq \bar{\bar{\eta}}, \tag{86}$$

where

$$\bar{\eta} := \min\Big\{\tfrac{1}{4L_w\sqrt{2\Lambda_0(\Theta_w+1)}}, \tfrac{1}{4L_{\Phi_0}}, \tfrac{1}{2L_w}, \tfrac{1}{2L_u\kappa}\Big\} \quad \textit{and} \quad \bar{\bar{\eta}} := \tfrac{\sqrt{\mu_H}}{2\sqrt{14\Lambda_0 L_u^3(\Theta_u+1)}}. \tag{87}$$

Then, the following bounds hold:

$$\tfrac{1}{T+1}\textstyle\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \tfrac{4[2\Psi_0(\widetilde{w}_0) - 2\Psi_0^\star + L_u^2\eta\|\widetilde{u}^0 - u^*(\widetilde{w}_0)\|^2]}{\eta(T+1)} + 8(C_w + C_u)\eta^2. \tag{88}$$

For a given $\epsilon > 0$, if we choose both $\eta := \mathcal{O}(\epsilon)$ and $\hat{\eta} := \mathcal{O}(\epsilon)$ satisfying (83) and $T := \mathcal{O}\big(\frac{1}{\epsilon^3}\big)$, then $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \epsilon^2$.

Consequently, Algorithm 2 requires $\mathcal{O}\big(\frac{n}{\epsilon^4}\big)$ evaluations of $\nabla_u\mathcal{H}_i$ and $\mathcal{O}\big(\frac{n}{\epsilon^3}\big)$ evaluations of $\nabla_w\mathcal{H}_i$ to achieve an ϵ-stationary point $\widehat{w}_T$ of (1) computed by (19). This algorithm also requires $\mathcal{O}(\epsilon^{-3})$ evaluations of $\mathrm{prox}_{\eta_t f}$ and $\mathcal{O}(\epsilon^{-4})$ evaluations of $\mathrm{prox}_{\hat{\eta}_t h}$.

Proof of Theorem 8. Since $\mu_H > 0$ and $\mu_h = 0$, for C_S given in Lemma 13, it reduces to

$$\begin{aligned} C_S &:= \Big[\textstyle\sum_{j=0}^{n-1}\frac{1}{(1+2\mu_h\hat{\eta}_t)}\big(1 - \frac{\mu_H\hat{\eta}_t}{n}\big)^j\Big]\cdot\Big[\sum_{s=0}^{S-1}\frac{1}{(1+2\mu_h\hat{\eta}_t)^s}\big(1 - \frac{\mu_H\hat{\eta}_t}{n}\big)^{ns}\Big] \\ &= \textstyle\sum_{s=0}^{S-1}\sum_{j=0}^{n-1}\big(1 - \frac{\mu_H\hat{\eta}_t}{n}\big)^{ns+j} \\ &\leq \tfrac{n}{\mu_H\hat{\eta}_t}. \end{aligned}$$

In this case, we can lower bound B_t from Lemma 17 as

$$\begin{aligned} B_t &:= 1 - 2L_{\Phi_0}\eta_t - 2L_w^2(3\Theta_w + 1)\Lambda_0\eta_t^2 \\ &\quad - \tfrac{2L_u^3}{n}(1 + \omega + 2\omega^2L_u^2\kappa^2\eta_t^2 + 4L_w^2\eta_t^2)\hat{\eta}_t^3\cdot C_S\cdot\Lambda_0(\Theta_u + 1) \\ &\geq 1 - 2L_{\Phi_0}\eta_t - 2L_w^2(3\Theta_w + 1)\Lambda_0\eta_t^2 - \tfrac{2L_u^3}{\mu_H}(1 + \omega + 2\omega^2L_u^2\kappa^2\eta_t^2 + 4L_w^2\eta_t^2)\Lambda_0(\Theta_u + 1)\cdot\hat{\eta}_t^2. \end{aligned}$$

Since $\eta_t := \eta \in (0, \bar{\eta}]$ for $\bar{\eta}$ satisfying (87), we have $\eta \leq \frac{1}{4L_{\Phi_0}}$, $\eta \leq \frac{1}{2L_w}$, and $\eta \leq \frac{1}{2L_u\kappa}$. Moreover, we choose $\omega := 1$ and $\hat{\eta}_t := \hat{\eta} \in (0, \bar{\bar{\eta}}]$. Hence, we can further lower bound B_t as

$$\begin{aligned} B_t &\geq \tfrac{1}{2} - 2\Lambda_0L_w^2(3\Theta_w + 1)\eta^2 - \tfrac{2\Lambda_0L_u^3(1+\omega+\omega M_\omega\eta^2)}{\mu_H}(\Theta_u + 1)\hat{\eta}^2 \\ &= \tfrac{1}{2} - 2\Lambda_0L_w^2(3\Theta_w + 1)\eta^2 - \tfrac{2\Lambda_0L_u^3(2+M_1\eta^2)}{\mu_H}(\Theta_u + 1)\hat{\eta}^2, \end{aligned} \tag{89}$$

where $M_\omega := 2\omega L_u^2\kappa^2 + \frac{4L_w^2}{\omega} = M_1 = 2L_u^2\kappa^2 + 4L_w^2$.

We can see that the second condition of (83) holds if

$$\begin{aligned} &\big(1 - \tfrac{\mu_H\hat{\eta}}{n}\big)^{nS}\big(2 + M_1\eta^2\big) \leq 1, \\ \Leftrightarrow\ &-nS\ln\big(1 - \tfrac{\mu_H\hat{\eta}}{n}\big) \geq \ln\big(2 + M_1\eta^2\big). \end{aligned}$$

Since $\eta \leq \frac{1}{4L_{\Phi_0}}$ and $\eta \leq \frac{1}{2L_w}$, we have $M_1\eta^2 = (2L_u^2\kappa^2 + 4L_w^2)\eta^2 \leq \frac{3}{2}$. Using this relation, and $-\ln(1-\tau) \geq \tau$ for $\tau \in (0, 1)$, the last condition holds if

$$\mu_H\hat{\eta}S \geq \ln(7/2) \quad \Leftrightarrow \quad S \geq \tfrac{\ln(7/2)}{\mu_H\hat{\eta}}.$$

Hence, we can choose $S := \lfloor \frac{\ln(7/2)}{\mu_H \hat{\eta}} \rfloor$ as stated in (86).

The condition (89) holds if

$$B_t \ \geq \ \tfrac{1}{2} - 2\Lambda_0 L_w^2(3\Theta_w + 1)\eta^2 - \tfrac{7\Lambda_0 L_u^3(\Theta_u+1)}{\mu_H}\hat{\eta}^2.$$

This condition shows that if we choose

$$0 < \eta \leq \tfrac{1}{4L_w\sqrt{2\Lambda_0(\Theta_w+1)}} \quad \text{and} \quad 0 < \hat{\eta} \leq \bar{\hat{\eta}} := \tfrac{\sqrt{\mu_H}}{2\sqrt{14\Lambda_0 L_u^3(\Theta_u+1)}},$$

then, from (89), we have $B_t \geq \frac{1}{4}$. Due to (86), both conditions here are satisfied.

Next, let us define C_w and C_u as in (85), respectively, i.e.:

$$C_w \ := \ L_w^2\big[(3\Theta_w + 1)\Lambda_1 + 3\sigma_w^2\big], \quad \text{and} \quad C_u := \tfrac{7L_u^3}{\mu_H}\big[\Lambda_1(\Theta_u + 1) + \sigma_u^2\big].$$

In this case, (84) reduces to

$$\begin{aligned} \Psi_0(\widetilde{w}_t) + \tfrac{L_u^2\eta}{2}\|\widetilde{u}_t - u_0^*(\widetilde{w}_t)\|^2 \ &\leq \ \Psi_0(\widetilde{w}_{t-1}) + \tfrac{L_u^2\eta}{2}\|\widetilde{u}_{t-1} - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{\eta}{8}\|\mathcal{G}_\eta(\widetilde{w}_{t-1})\|^2 + C_w\eta^3 + C_u\hat{\eta}^2\eta. \end{aligned}$$

Subtracting $\Psi_0^\star$ from both sides of this inequality, and averaging the result from $t = 0$ to T, and noting that $\Psi_0(\widetilde{w}_T) - \Psi_0^\star \geq 0$, we obtain (88).

To achieve an ϵ-stationary point of (3), from (88), we need to impose the following condition:

$$\tfrac{2[\Psi_0(\widetilde{w}_0) - \Psi_0^\star]}{\eta(T+1)} + \tfrac{L_u^2\|\widetilde{u}^0 - u_0^*(\widetilde{w}_0)\|^2}{T+1} + 2C_w\eta^2 + 2C_u\hat{\eta}^2 \leq \tfrac{\epsilon^2}{4}.$$

Since other terms are constant, if we choose $\eta := \mathcal{O}(\epsilon)$ and $\hat{\eta} := \mathcal{O}(\epsilon)$ such that they still satisfies (86), then we can choose $T := \mathcal{O}(\epsilon^{-3})$ to guarantee the last condition. Since each iteration t, we run S epochs of the shuffling scheme (27), the total number of evaluations of $\nabla_u\mathcal{H}_i$ is $\mathcal{T}_u := T \times S \times n$. However, we have $S = \lfloor \frac{\ln(7/2)}{\mu_H\hat{\eta}} \rfloor = \mathcal{O}(\epsilon^{-1})$, we get $\mathcal{T}_u := \mathcal{O}(n\epsilon^{-4})$. The total number of evaluations of $\nabla_w\mathcal{H}_i$ is $\mathcal{T}_w := Tn = \mathcal{O}(n\epsilon^{-3})$ as stated.

Finally, since each epoch t, Algorithm 2 requires one evaluation of $\mathrm{prox}_{\eta_t f}$, and S evaluations of $\mathrm{prox}_{\hat{\eta}_t h}$, but since $S = \mathcal{O}(\epsilon^{-1})$, the total number of $\mathrm{prox}_{\eta_t f}$ evaluations is $T = \mathcal{O}(\epsilon^{-3})$, while the total number of $\mathrm{prox}_{\hat{\eta}_t h}$ evaluations is $TS = \mathcal{O}(\epsilon^{-4})$. Overall, Algorithm 2 needs $\mathcal{O}(\epsilon^{-3})$ evaluations of $\mathrm{prox}_{\eta_t f}$ and $\mathcal{O}(\epsilon^{-4})$ evaluations of $\mathrm{prox}_{\hat{\eta}_t h}$. □

Proof of Theorem 4. Since $\mu_h > 0$ and $\mu_H = 0$, for C_S given by Lemma 13, it reduces to

$$\begin{aligned} C_S \ &:= \ \big[\textstyle\sum_{j=0}^{n-1}\tfrac{1}{(1+2\mu_h\hat{\eta}_t)}\big(1 - \tfrac{\mu_H\hat{\eta}_t}{n}\big)^j\big] \cdot \big[\textstyle\sum_{s=0}^{S-1}\tfrac{1}{(1+2\mu_h\hat{\eta}_t)^s}\big(1 - \tfrac{\mu_H\hat{\eta}_t}{n}\big)^{ns}\big] \\ &= \ \textstyle\sum_{s=0}^{S-1}\tfrac{1}{(1+2\mu_h\hat{\eta}_t)^{s+1}} \leq \tfrac{1}{2\mu_h\hat{\eta}_t}. \end{aligned}$$

In this case, we can lower bound B_t from Lemma 17 as in Theorem 3, i.e.:

$$B_t \ \geq \ 1 - 2L_{\Phi_0}\eta_t - 2L_w^2(3\Theta_w + 1)\Lambda_0\eta_t^2 - \tfrac{2L_u^3}{n\mu_H}(1 + L_u^2\kappa^2\eta_t^2 + 2L_w^2\eta_t^2)\Lambda_0(\Theta_u + 1) \cdot \hat{\eta}_t^2.$$

We need to choose η as in Theorem 3. Since $\eta_t = \eta \in (0, \bar{\eta}]$, we have $\eta \leq \frac{1}{4L_{\Phi_0}}$, $\eta \leq \frac{1}{2L_w}$, and $\eta \leq \frac{1}{2L_u\kappa}$. Alternatively, we also have $\hat{\eta}_t := \hat{\eta} \in (0, \bar{\hat{\eta}}]$. Therefore, we can further lower bound B_t as $B_t \geq \frac{1}{4}$ as in Theorem 3.

Now, the second condition of (83) holds if

$$\big(1 + 2\mu_h\hat{\eta}\big)^S \geq 2 + (2L_u^2\kappa^2 + 4L_w^2)\eta^2.$$

Using the fact that $\ln(1 + e) \geq \frac{e}{2}$ for $e \in (0, 1/2)$ and $(2L_u^2\kappa^2 + 4L_w^2)\eta^2 \leq \frac{3}{2}$, the last inequality holds if $\mu_h\hat{\eta}S \geq \ln(7/2)$. Hence, we can choose $S := \big\lfloor \frac{\ln(7/2)}{\mu_h\hat{\eta}} \big\rfloor$ as stated in Theorem 4. The remaining proof follows from Theorem 3. □

C.4 Convergence of the full-shuffling variant of Algorithm 2 – The case $S = 1$

In this subsection, we analyze the convergence of Algorithm 2 using only one epoch of the ***shuffling gradient ascent*** scheme (27). In this case, by dropping the superscript s, the scheme (27) can be simplified as follows:

$$\begin{cases} u_0^{(t)} := \widetilde{u}_{t-1}, \\ \text{For } i = 1, 2, \cdots, n, \text{ update} \\ \qquad u_i^{(t)} := u_{i-1}^{(t)} + \frac{\hat{\eta}_t}{n} \nabla_u \mathcal{H}_{\pi^{(t)}(i)}(\widetilde{w}_{t-1}, u_{i-1}^{(t)}), \\ \widetilde{u}_t := \mathrm{prox}_{\hat{\eta}_t h}(u_n^{(t)}), \end{cases} \tag{90}$$

where $\hat{\eta}_t > 0$ is a given learning rate.

We divide our analysis into different tasks as follows.

C.4.1 Potential function and a technical lemma

One key step of our analysis is to construct an appropriate potential function. We exploit the ideas from [3] to construct this function as follows.

For Ψ_0 defined by (3) and $\mathcal{L}$ given in (1), we consider the following **potential function**:

$$\mathcal{V}_\lambda(w, u) := \lambda\big[\Psi_0(w) - \Psi_0^\star\big] + \Psi_0(w) - \mathcal{L}(w, u), \tag{91}$$

where $\lambda > 0$ is a given parameter determined later. Since $\Psi_0(w) \geq \Psi_0^\star := \inf_w \Psi_0(w)$ and $\Psi_0(w) = \sup_u \mathcal{L}(w, u) \geq \mathcal{L}(w, u)$, it is obvious to see that $\mathcal{V}_\lambda(w, u) \geq 0$ for all $(w, u) \in \mathrm{dom}\,(\mathcal{L})$.

Similar to (18), we consider the following gradient mappings for both (2) and (3), respectively:

$$\begin{array}{lll} \hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1}) := \frac{1}{\hat{\eta}_t}\big(\widetilde{u}_{t-1} - \widehat{u}_t\big), & \text{where} & \widehat{u}_t := \mathrm{prox}_{\hat{\eta}_t h}\big(\widetilde{u}_{t-1} + \hat{\eta}_t \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\big), \\ \mathcal{G}_{\eta_t}(\widetilde{w}_{t-1}) := \frac{1}{\eta_t}\big(\widetilde{w}_{t-1} - \widehat{w}_t\big), & \text{where} & \widehat{w}_t := \mathrm{prox}_{\eta_t f}\big(\widetilde{w}_{t-1} - \eta_t \nabla \Phi_0(\widetilde{w}_{t-1})\big). \end{array} \tag{92}$$

We need the following result.

Lemma 18. *Let $\widehat{u}_t$ and $\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})$ be defined by (92), and ψ be defined by (56). Then, we have*

$$2\mu_\psi \|\widehat{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \;\leq\; \psi(\widetilde{w}_{t-1}, \widehat{u}_t) - \psi(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1})). \tag{93}$$

If $0 < \hat{\eta}_t \leq \frac{2}{L_u + \mu_H}$, then we have

$$2\mu_\psi[\psi(\widetilde{w}_{t-1}, \widehat{u}_t) - \psi(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))] \leq \left(1 - \frac{2L_u\mu_H\hat{\eta}_t}{L_u + \mu_H}\right) \|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2. \tag{94}$$

Proof. Since $\psi(\widetilde{w}_{t-1}, \cdot)$ is μ_ψ-strongly convex and $u_0^*(\widetilde{w}_{t-1}) := \operatorname*{argmin}_u \psi(\widetilde{w}_{t-1}, u)$, for $\widehat{u}_t$ given in (92), we easily obtain (93).

Next, using again the μ_ψ-strong convexity of $\psi(\widetilde{w}_{t-1}, \cdot)$ and $u_0^*(\widetilde{w}_{t-1}) := \operatorname*{argmin}_u \psi(\widetilde{w}_{t-1}, u)$, for $\widehat{u}_t$ given in (92), by [6, Theorem 2.1.10], we have

$$2\mu_\psi[\psi(\widetilde{w}_{t-1}, \widehat{u}_t) - \psi(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))] \leq \|\nabla_u \psi(\widetilde{w}_{t-1}, \widehat{u}_t)\|^2. \tag{95}$$

where $\nabla_u \psi(\widetilde{w}_{t-1}, \widehat{u}_t) = -\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t) + \nabla h(\widehat{u}_t) \in \partial\psi(\widehat{w}_t) := -\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t) + \partial h(\widehat{u}_t)$.

Now, for $\widehat{u}_t$ defined by (92), we have $\frac{1}{\hat{\eta}_t}(\widetilde{u}_{t-1} - \widehat{u}_t) + \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) \in \partial h(\widehat{u}_t)$, leading to

$$\nabla_u \psi(\widetilde{w}_{t-1}, \widehat{u}_t) := \frac{1}{\hat{\eta}_t}(\widetilde{u}_{t-1} - \widehat{u}_t) + \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t) \in \partial\psi(\widehat{u}_t). \tag{96}$$

Since $-\mathcal{H}(\widetilde{w}_{t-1}, \cdot)$ is L_u-smooth and μ_H-strongly convex, by [6, Theorem 2.1.12], we have

$$\begin{array}{l} -\langle \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t), \widetilde{u}_{t-1} - \widehat{u}_t \rangle \geq \frac{L_u\mu_H}{L_u + \mu_H}\|\widetilde{u}_{t-1} - \widehat{u}_t\|^2 \\ \qquad + \frac{1}{L_u + \mu_H}\|\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t)\|^2. \end{array} \tag{97}$$

Utilizing (96) and (97), we can show that

$$
\begin{aligned}
\|\nabla_u \psi(\widetilde{w}_{t-1}, \widehat{u}_t)\|^2 &= \tfrac{1}{\hat{\eta}_t^2}\|\widetilde{u}_{t-1} - \widehat{u}_t\|^2 + \tfrac{2}{\hat{\eta}_t}\langle \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t), \widetilde{u}_{t-1} - \widehat{u}_t\rangle \\
&\quad + \|\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t)\|^2 \\
&\le \tfrac{1}{\hat{\eta}_t^2}\left(1 - \tfrac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\right)\|\widetilde{u}_{t-1} - \widehat{u}_t\|^2 \\
&\quad - \left(\tfrac{2}{\hat{\eta}_t(L_u+\mu_H)} - 1\right)\|\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t)\|^2.
\end{aligned}
$$

Substituting this inequality into (95) and noting that $0 < \hat{\eta}_t \le \frac{2}{L_u+\mu_H}$ and $\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1}) := \frac{1}{\hat{\eta}_t}(\widetilde{u}_{t-1} - \widehat{u}_t)$, we obtain (94). $\square$

C.4.2 A key bound for the shuffling gradient descent scheme (28)

The following lemma bounds the difference $\mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_t) - \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t)$.

Lemma 19. *Suppose that Assumptions 4 and 7 hold. Let $\mathcal{L}$ be defined by (1) and g_t be defined by (68). Then, we have*

$$
\mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_t) \le \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t) + \tfrac{\eta_t}{2}\|g_t - \nabla_w \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 + \tfrac{3+(L_f+L_w)\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2. \tag{98}
$$

Proof. From (68) and (69), we have

$$
g_t := \tfrac{1}{n}\textstyle\sum_{j=1}^n \nabla_w \mathcal{H}_{\hat{\pi}^{(t)}(j)}(w_{j-1}^{(t)}, \widetilde{u}_t) \overset{(69)}{=} \tfrac{1}{\eta_t}(\widetilde{w}_{t-1} - w_n^{(t)}) = \tfrac{1}{\eta_t}(w_0^{(t)} - w_n^{(t)}).
$$

Since $\widetilde{w}_t = \mathrm{prox}_{\eta_t f}(w_n^{(t)})$ from the second line of (28), we have $f'(\widetilde{w}_t) := \eta_t^{-1}(w_n^{(t)} - \widetilde{w}_t) = -g_t - \eta_t^{-1}(\widetilde{w}_t - \widetilde{w}_{t-1}) \in \partial f(\widetilde{w}_t)$. Hence, by (30) from Assumption 7, we have

$$
\begin{aligned}
f(\widetilde{w}_{t-1}) &\le f(\widetilde{w}_t) + \langle f'(\widetilde{w}_t), \widetilde{w}_{t-1} - \widetilde{w}_t\rangle + \tfrac{L_f}{2}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\
&= f(\widetilde{w}_t) + \langle g_t, \widetilde{w}_t - \widetilde{w}_{t-1}\rangle + \tfrac{2+L_f\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2.
\end{aligned}
$$

Next, by the L-smoothness of $\mathcal{H}$ from (12) of Assumption 4, we also have

$$
\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t) \le \mathcal{H}(\widetilde{w}_t, \widetilde{u}_t) - \langle \nabla_w \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t), \widetilde{w}_t - \widetilde{w}_{t-1}\rangle + \tfrac{L_w}{2}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2.
$$

Summing up the last two inequalities and using $\mathcal{L}(w, \widetilde{u}_t) = f(w) + \mathcal{H}(w, \widetilde{u}_t) - h(\widetilde{u}_t)$ from (1) and Young's inequality in ①, we can show that

$$
\begin{aligned}
\mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_t) &\le \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t) + \langle g_t - \nabla_w \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t), \widetilde{w}_t - \widetilde{w}_{t-1}\rangle + \tfrac{2+(L_f+L_w)\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\
&\overset{①}{\le} \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t) + \tfrac{\eta_t}{2}\|g_t - \nabla_w \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 + \tfrac{3+(L_f+L_w)\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2,
\end{aligned}
$$

which proves (98). $\square$

C.4.3 Key bounds for the shuffling gradient ascent scheme (27)

We also derive necessary bounds to analyze Algorithm 2 using the simplified version (90) of the ***shuffling gradient ascent*** scheme (27). Let us define

$$
v_t := \tfrac{1}{n}\textstyle\sum_{j=1}^n \nabla_u \mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1}, u_{j-1}^{(t)}). \tag{99}
$$

We establish the following two lemmas.

Lemma 20. *Suppose that Assumption 4 holds for (1). Let $\{u_i^{(t)}\}_{i=1}^n$ be generated by the simplified version (90) of (27). Then, if we choose $\hat{\eta}_t > 0$ such that $1 - 3L_u^2\hat{\eta}_t^2 \ge 0$, then*

$$
\hat{\Delta}_t := \tfrac{1}{n}\textstyle\sum_{i=0}^{n-1}\|u_i^{(t)} - u_0^{(t)}\|^2 \le (3\Theta_u + 2)\hat{\eta}_t^2 \cdot \|\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2 + 3\hat{\eta}_t^2\sigma_u^2. \tag{100}
$$

Let v_t be defined by (99). Then, we also have

$$
\|v_t - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2 \le \tfrac{L_u^2}{n}\textstyle\sum_{i=1}^n \|u_{i-1}^{(t)} - \widetilde{u}_{t-1}\|^2 = L_u^2\hat{\Delta}_t. \tag{101}
$$

Proof. First, for simplicity of notation, we denote $\nabla_u\mathcal{H}_{t-1} := \nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})$. Then, from the update of $u_i^{(t)}$ in (90), for any $i\in[n]$, we have

$$u_i^{(t)} = u_0^{(t)} + \frac{\hat{\eta}_t}{n}\sum_{j=1}^{i}\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1}, u_{j-1}^{(t)}). \tag{102}$$

Using this expression and Young's inequality in ① and ② below, we can show that

$$\begin{array}{lcl}
\|u_i^{(t)} - u_0^{(t)}\|^2 & \overset{(102)}{=} & \frac{i^2\cdot\hat{\eta}_t^2}{n^2}\|\frac{1}{i}\sum_{j=1}^{i}\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_{j-1}^{(t)})\|^2\\
& \overset{①}{\leq} & \frac{3i^2\cdot\hat{\eta}_t^2}{n^2}\big\|\frac{1}{i}\sum_{j=1}^{i}\big[\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_0^{(t)}) - \nabla_u\mathcal{H}_{t-1}\big]\big\|^2 + \frac{3i^2\cdot\hat{\eta}_t^2}{n^2}\|\nabla_u\mathcal{H}_{t-1}\|^2\\
& & + \frac{3i^2\cdot\hat{\eta}_t^2}{n^2}\big\|\frac{1}{i}\sum_{j=1}^{i}\big[\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_{j-1}^{(t)}) - \nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_0^{(t)})\big]\big\|^2\\
& \overset{②}{\leq} & \frac{3i^2\cdot\hat{\eta}_t^2}{n^2}\big\|\frac{1}{i}\sum_{j=1}^{i}\big[\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_0^{(t)}) - \nabla_u\mathcal{H}_{t-1}\big]\big\|^2 + \frac{3i^2\cdot\hat{\eta}_t^2}{n^2}\|\nabla_u\mathcal{H}_{t-1}\|^2\\
& & + \frac{3i\cdot\hat{\eta}_t^2}{n^2}\sum_{j=1}^{i}\big\|\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_{j-1}^{(t)}) - \nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_0^{(t)})\big\|^2.
\end{array}$$

Now, we denote $\hat{\Delta}_t := \frac{1}{n}\sum_{j=0}^{n-1}\|u_j^{(t)} - u_0^{(t)}\|^2 = \frac{1}{n}\sum_{j=0}^{n-1}\|u_j^{(t)} - \widetilde{u}_{t-1}\|^2$. Then, by (12) from Assumption 4, we have

$$\begin{array}{lcl}
\frac{1}{n}\sum_{j=1}^{i}\big\|\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_{j-1}^{(t)}) - \nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_0^{(t)})\big\|^2 & \overset{(12)}{\leq} & \frac{L_u^2}{n}\sum_{j=1}^{i}\|u_{j-1}^{(t)} - u_0^{(t)}\|^2\\
& \leq & L_u^2\hat{\Delta}_t.
\end{array}$$

Next, by Young's inequality again in ①, $u_0^{(t)} = \widetilde{u}_{t-1}$, and (14) from Assumption 4, and the fact that $\nabla_u\mathcal{H}_{t-1} := \nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})$, we can show that

$$\begin{array}{lcl}
\mathcal{T}_{[2]} & := & \|\frac{1}{i}\sum_{j=1}^{i}\big[\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},u_0^{(t)}) - \nabla_u\mathcal{H}_{t-1}\big]\|^2\\
& \overset{①}{\leq} & \frac{1}{i}\sum_{j=1}^{i}\big\|\nabla_u\mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1},\widetilde{u}_{t-1}) - \nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1}))\big\|^2\\
& \overset{(14)}{\leq} & \frac{n}{i}\big[\Theta_u\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2 + \sigma_u^2\big].
\end{array}$$

Combining the last three inequalities above, we can show that

$$\begin{array}{lcl}
\|u_i^{(t)} - u_0^{(t)}\|^2 & \leq & \frac{3i\cdot\hat{\eta}_t^2}{n}\big[\Theta_u\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2 + \sigma_u^2\big]\\
& & + \frac{3i^2\cdot\hat{\eta}_t^2}{n^2}\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2 + \frac{3i\cdot L_u^2\hat{\eta}_t^2}{n}\hat{\Delta}_t\\
& = & \frac{3i\hat{\eta}_t^2}{n^2}(n\Theta_u + i)\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2 + \frac{3i\cdot\hat{\eta}_t^2}{n}\sigma_u^2 + \frac{3i\cdot L_u^2\hat{\eta}_t^2}{n}\hat{\Delta}_t.
\end{array}$$

Averaging this inequality from $i = 0$ to $i = n-1$, we get

$$\begin{array}{lcl}
\hat{\Delta}_t & := & \frac{1}{n}\sum_{i=0}^{n-1}\|u_i^{(t)} - u_0^{(t)}\|^2\\
& \leq & \frac{1}{n}\sum_{i=0}^{n-1}\big[\frac{3i\hat{\eta}_t^2}{n^2}(n\Theta_u + i)\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2 + \frac{3i\cdot\hat{\eta}_t^2}{n}\sigma_u^2 + \frac{3i\cdot L_u^2\hat{\eta}_t^2}{n}\hat{\Delta}_t\big]\\
& \leq & \frac{(3\Theta_u+2)\hat{\eta}_t^2}{2}\cdot\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2 + \frac{3\hat{\eta}_t^2}{2}\sigma_u^2 + \frac{3L_u^2\hat{\eta}_t^2}{2}\hat{\Delta}_t.
\end{array}$$

Here, we have used the facts that $\sum_{i=0}^{n-1} i = \frac{n(n-1)}{2} \leq \frac{n^2}{2}$ and $\sum_{i=0}^{n-1} i^2 = \frac{n(n-1)(2n-1)}{6} \leq \frac{n^3}{3}$. Rearranging the last inequality and noting that $1 - \frac{3L_u^2\hat{\eta}_t^2}{2} \geq \frac{1}{2}$, we obtain (100).

Finally, using (99) and (12) from Assumption 4, we have

$$\begin{array}{lcl}
\|v_t - \nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2 & \overset{(99)}{=} & \big\|\frac{1}{n}\sum_{i=1}^{n}\big[\nabla_u\mathcal{H}_{\pi^{(t)}(i)}(\widetilde{w}_{t-1},u_{i-1}^{(t)}) - \nabla_u\mathcal{H}_{\pi^{(t)}(i)}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\big]\big\|^2\\
& \leq & \frac{1}{n}\sum_{i=1}^{n}\big\|\nabla_u\mathcal{H}_{\pi^{(t)}(i)}(\widetilde{w}_{t-1},u_{i-1}^{(t)}) - \nabla_u\mathcal{H}_{\pi^{(t)}(i)}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\big\|^2\\
& \overset{(12)}{\leq} & \frac{L_u^2}{n}\sum_{i=1}^{n}\|u_{i-1}^{(t)} - \widetilde{u}_{t-1}\|^2,
\end{array}$$

which proves (101). □

Lemma 21. *Suppose that Assumption 4 holds for (1). Let $\{u_i^{(t)}\}_{i=1}^n$ be generated by the simplified version (90) of (27), ψ be defined by (56), v_t be defined by (99), and $\hat{\mathcal{G}}_{\hat{\eta}}$ be defined as in (92). Then*

$$\begin{array}{lcl}
\psi(\widetilde{w}_{t-1},\widetilde{u}_t) & \leq & \psi(\widetilde{w}_{t-1},\widetilde{u}_{t-1}) - \frac{\hat{\eta}_t[1+(\mu_h+\mu_H)\hat{\eta}_t - L_u\hat{\eta}_t]}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2\\
& & - \frac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 + \frac{\hat{\eta}_t}{2(1+\mu_h\hat{\eta}_t)}\|v_t - \nabla_u\mathcal{H}(\widetilde{w}_{t-1},\widetilde{u}_{t-1})\|^2.
\end{array} \tag{103}$$

Proof. Let us denote $\widehat{u}_t := \mathrm{prox}_{\hat{\eta}_t h}\big(\widetilde{u}_{t-1} + \hat{\eta}_t \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\big)$ as in (92). Then, from (92), we have $\hat{\eta}_t \hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1}) = \widetilde{u}_{t-1} - \widehat{u}_t$. Moreover, we can show that $\nabla h(\widehat{u}_t) := \frac{1}{\hat{\eta}_t}\big(\widetilde{u}_{t-1} - \widehat{u}_t\big) + \nabla \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) \in \partial h(\widehat{u}_t)$.

By the μ_h-strong convexity of h, we have

$$\begin{aligned} h(\widehat{u}_t) &\leq h(\widetilde{u}_{t-1}) + \langle \nabla h(\widehat{u}_t), \widehat{u}_t - \widetilde{u}_{t-1}\rangle - \tfrac{\mu_h}{2}\|\widehat{u}_t - \widetilde{u}_{t-1}\|^2 \\ &= h(\widetilde{w}_{t-1}) + \langle \nabla \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}), \widehat{u}_t - \widetilde{u}_{t-1}\rangle - \tfrac{2+\mu_h\hat{\eta}_t}{2\hat{\eta}_t}\|\widehat{u}_t - \widetilde{u}_{t-1}\|^2. \end{aligned}$$

Next, by the L_u-smoothness of $\mathcal{H}(\widetilde{w}_{t-1}, \cdot)$ from Assumption 4, we have

$$-\mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t) \leq -\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \langle \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}), \widehat{u}_t - \widetilde{u}_{t-1}\rangle + \tfrac{L_u}{2}\|\widehat{u}_t - \widetilde{u}_{t-1}\|^2.$$

Summing up the last two inequalities and using both $\psi(\widetilde{w}_{t-1}, u) := -\mathcal{H}(\widetilde{w}_{t-1}, u) + h(u)$ from (56) and $\widehat{u}_t - \widetilde{u}_{t-1} = -\hat{\eta}_t \hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})$, we can show that

$$\begin{aligned} \psi(\widetilde{w}_{t-1}, \widehat{u}_t) &\leq \psi(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \tfrac{(2+\mu_h\hat{\eta}_t - L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widehat{u}_t - \widetilde{u}_{t-1}\|^2 \\ &= \psi(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \tfrac{\hat{\eta}_t(2+\mu_h\hat{\eta}_t - L_u\hat{\eta}_t)}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2. \end{aligned} \tag{104}$$

Next, from (99) and (90), one can derive that

$$v_t \overset{(99)}{:=} \tfrac{1}{n}\sum_{j=1}^n \nabla_u \mathcal{H}_{\pi^{(t)}(j)}(\widetilde{w}_{t-1}, u_{j-1}^{(t)}) \overset{(90)}{=} \tfrac{1}{\hat{\eta}_t}(u_n^{(t)} - \widetilde{u}_{t-1}) = \tfrac{1}{\hat{\eta}_t}(u_n^{(t)} - u_0^{(t)}). \tag{105}$$

Since $\widetilde{u}_t = \mathrm{prox}_{\hat{\eta}_t h}(u_n^{(t)})$ from (90), we have $\nabla h(\widetilde{u}_t) := \frac{1}{\hat{\eta}_t}\big(u_n^{(t)} - \widetilde{u}_t\big) = v_t - \frac{1}{\hat{\eta}_t}(\widetilde{u}_t - \widetilde{u}_{t-1}) \in \partial h(\widetilde{u}_t)$. Hence, again by the μ_h-strong convexity of h, we have

$$\begin{aligned} h(\widetilde{u}_t) &\leq h(\widehat{u}_t) + \langle \nabla h(\widetilde{u}_t), \widetilde{u}_t - \widehat{u}_t\rangle - \tfrac{\mu_h}{2}\|\widetilde{u}_t - \widehat{u}_t\|^2 \\ &= h(\widehat{u}_t) + \langle v_t, \widetilde{u}_t - \widehat{u}_t\rangle - \tfrac{1}{\hat{\eta}_t}\langle \widetilde{u}_t - \widetilde{u}_{t-1}, \widetilde{u}_t - \widehat{u}_t\rangle - \tfrac{\mu_h}{2}\|\widetilde{u}_t - \widehat{u}_t\|^2 \\ &= h(\widehat{u}_t) + \langle v_t, \widetilde{u}_t - \widehat{u}_t\rangle - \tfrac{1}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 - \tfrac{(1+\mu_h\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widehat{u}_t\|^2 + \tfrac{1}{2\hat{\eta}_t}\|\widehat{u}_t - \widetilde{u}_{t-1}\|^2. \end{aligned}$$

Again, by the L_u-smoothness and μ_H-strong concavity of $\mathcal{H}(\widetilde{w}_{t-1}, \cdot)$ from Assumption 4, we have

$$\begin{aligned} -\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t) &\leq -\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \langle \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}), \widetilde{u}_t - \widetilde{u}_{t-1}\rangle + \tfrac{L_u}{2}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2, \\ -\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) &\leq -\mathcal{H}(\widetilde{w}_{t-1}, \widehat{u}_t) - \langle \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}), \widetilde{u}_{t-1} - \widehat{u}_t\rangle - \tfrac{\mu_H}{2}\|\widehat{u}_t - \widetilde{u}_{t-1}\|^2. \end{aligned}$$

Adding the last three inequalities together, and using $\psi(\widetilde{w}_{t-1}, u) = h(u) - \mathcal{H}(\widetilde{w}_{t-1}, u)$ from (56) and $\widehat{u}_t - \widetilde{u}_{t-1} = -\hat{\eta}_t \hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})$, we have

$$\begin{aligned} \psi(\widetilde{w}_{t-1}, \widetilde{u}_t) &\leq \psi(\widetilde{w}_{t-1}, \widehat{u}_t) - \langle \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - v_t, \widetilde{u}_t - \widehat{u}_t\rangle - \tfrac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 \\ &\quad + \tfrac{1-\mu_H\hat{\eta}_t}{2\hat{\eta}_t}\|\widehat{u}_t - \widetilde{u}_{t-1}\|^2 - \tfrac{(1+\mu_h\hat{\eta}_t)}{2\eta_t}\|\widetilde{u}_t - \widehat{u}_t\|^2 \\ &\overset{①}{\leq} \varphi_t(\widehat{u}_t) + \tfrac{\hat{\eta}_t}{2(1+\mu_h\hat{\eta}_t)}\|\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - v_t\|^2 - \tfrac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 \\ &\quad + \tfrac{\hat{\eta}_t(1-\mu_H\hat{\eta}_t)}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2, \end{aligned} \tag{106}$$

where we have used Young's inequality in the last line ① as $\langle \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - v_t, \widetilde{u}_t - \widehat{u}_t\rangle \leq \frac{\hat{\eta}_t}{2(1+\mu_h\hat{\eta}_t)}\|\nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - v_t\|^2 + \frac{1+\mu_h\hat{\eta}_t}{2\hat{\eta}_t}\|\widetilde{u}_t - \widehat{u}_t\|^2$.

Finally, summing up (104) and (106), we get

$$\begin{aligned} \psi(\widetilde{w}_{t-1}, \widetilde{u}_t) &\leq \psi(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \tfrac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 - \tfrac{\hat{\eta}_t[1+(\mu_h+\mu_H)\hat{\eta}_t - L_u\hat{\eta}_t]}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \\ &\quad + \tfrac{\hat{\eta}_t}{2(1+\mu_h\hat{\eta}_t)}\|v_t - \nabla_u \mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2, \end{aligned}$$

which proves (103). $\square$

C.4.4 Convergence analysis of the full-shuffling variant of Algorithm 2 – The case $S = 1$

To analyze the convergence of the full-shuffling variant of Algorithm 2, we need the following lemma.

Lemma 22. *Let $\mathcal{V}_\lambda$ be defined by (91), $\mathcal{V}_t := \mathcal{V}_\lambda(\widetilde{w}_t, \widetilde{u}_t)$, and g_t and v_t be defined by (68) and (99), respectively. Suppose that f satisfies (30) of Assumption 7. Then, the following bound holds:*

$$\begin{aligned} \mathcal{V}_t - \mathcal{V}_{t-1} \leq & -\tfrac{\lambda-2-[(1+\lambda)L_{\Phi_0}+L_w+L_f]\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 \\ & - \tfrac{(\lambda+1)\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \tfrac{\hat{\eta}_t(1+(\mu_h+\mu_H)\hat{\eta}_t - L_u\hat{\eta}_t)}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \\ & + \tfrac{(\lambda+1)\eta_t}{2}\|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \tfrac{\eta_t}{2}\|g_t - \nabla_w\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 \\ & + \tfrac{\hat{\eta}_t}{2(1+\mu_h\hat{\eta}_t)}\|v_t - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2. \end{aligned} \tag{107}$$

Proof. From (91), if we denote $\mathcal{V}_t := \mathcal{V}_\lambda(\widetilde{w}_t, \widetilde{u}_t)$, then we have

$$\begin{aligned} \mathcal{V}_t - \mathcal{V}_{t-1} &= (\lambda+1)\big[\Psi_0(\widetilde{w}_t) - \Psi_0(\widetilde{w}_{t-1})\big] + \mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t) \\ &= (\lambda+1)\big[\Psi_0(\widetilde{w}_t) - \Psi_0(\widetilde{w}_{t-1})\big] + \mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) - \mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_t) \\ &\quad + \mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_t) - \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t) \\ &= (\lambda+1)\big[\Psi_0(\widetilde{w}_t) - \Psi_0(\widetilde{w}_{t-1})\big] + \mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_t) - \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t) \\ &\quad + \psi(\widetilde{w}_{t-1}, \widetilde{u}_t) - \psi(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}). \end{aligned} \tag{108}$$

Next, from (72), we have

$$\begin{aligned} \Psi_0(\widetilde{w}_t) - \Psi_0(\widetilde{w}_{t-1}) \leq & -\tfrac{\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \tfrac{(1-L_{\Phi_0}\eta_t)}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 \\ & + \tfrac{\eta_t}{2}\|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2. \end{aligned} \tag{109}$$

From (98), we also have

$$\mathcal{L}(\widetilde{w}_{t-1}, \widetilde{u}_t) - \mathcal{L}(\widetilde{w}_t, \widetilde{u}_t) \leq \tfrac{\eta_t}{2}\|g_t - \nabla_w\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 + \tfrac{3+(L_f+L_w)\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2. \tag{110}$$

From (103), we can rewrite it as

$$\begin{aligned} \psi(\widetilde{w}_{t-1}, \widetilde{u}_t) - \psi(\widetilde{w}_{t-1}, \widetilde{u}_{t-1}) \leq & -\tfrac{\hat{\eta}_t[1+(\mu_h+\mu_H)\hat{\eta}_t - L_u\hat{\eta}_t]}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \\ & + \tfrac{\hat{\eta}_t}{2(1+\mu_h\hat{\eta}_t)}\|v_t - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2 \\ & - \tfrac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2. \end{aligned} \tag{111}$$

Substituting (109), (110), and (111) into (108), we can derive that

$$\begin{aligned} \mathcal{V}_t - \mathcal{V}_{t-1} \leq & -\tfrac{\lambda-2-[(1+\lambda)L_{\Phi_0}+L_w+L_f]\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 \\ & - \tfrac{(\lambda+1)\eta_t(1-2L_{\Phi_0}\eta_t)}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 - \tfrac{\hat{\eta}_t(1+(\mu_h+\mu_H)\hat{\eta}_t - L_u\hat{\eta}_t)}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \\ & + \tfrac{(\lambda+1)\eta_t}{2}\|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 + \tfrac{\eta_t}{2}\|g_t - \nabla_w\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 \\ & + \tfrac{\hat{\eta}_t}{2(1+\mu_h\hat{\eta}_t)}\|v_t - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2, \end{aligned}$$

which proves (107). $\square$

Next, we further upper bound (107) from Lemma 22 as follows.

Lemma 23. *Under the same condition as in Lemma 22, $1 - 3L_w^2\eta_t^2 \geq 0$, $1 - 3L_u^2\hat{\eta}_t^2 \geq 0$, and $\hat{\eta}_t \leq \frac{2}{L_u+\mu_H}$, we have*

$$\begin{aligned} \mathcal{V}_t - \mathcal{V}_{t-1} \leq & -\tfrac{C_0}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{1-(L_u+3C_1)\hat{\eta}_t}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 - \tfrac{C_3\eta_t}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\ & - \big(\mu_\psi C_2\hat{\eta}_t - \tfrac{3C_1\eta_t}{4\mu_\psi}\big)\big[\psi(\widetilde{w}_{t-1}, \hat{u}_t) - \psi(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\big] + C_4\eta_t^3 + C_5\hat{\eta}_t^3, \end{aligned} \tag{112}$$

where C_i for $i = 0, 1, \cdots, 5$ are respectively given as follows:

$$\begin{cases} C_0 := \lambda - 2 - [(1+\lambda)L_{\Phi_0} + L_w + L_f]\eta_t, \\ C_1 := L_u^2[\lambda + 1 + 4(\lambda+2)L_w^2\eta_t^2], \\ C_2 := 1 - (L_u - \mu_\psi)\hat{\eta}_t - \frac{\hat{\Lambda}_0 L_u^2(3\Theta_u+2)\hat{\eta}_t^2}{1+\mu_h\hat{\eta}_t} - 3C_1\eta_t\hat{\eta}_t, \\ C_3 := (\lambda+1)(1-2L_{\Phi_0}\eta_t) - 2\Lambda_0(\lambda+2)L_w^2(3\Theta_w+1)\eta_t^2, \\ C_4 := (\lambda+2)L_w^2\big[3\sigma_w^2 + \Lambda_1(3\Theta_w+1)\big], \\ C_5 := \frac{\hat{\Lambda}_1 L_u^2(3\Theta_u+2)+3L_u^2\sigma_u^2}{2(1+\mu_h\hat{\eta}_t)}. \end{cases} \tag{113}$$

Proof. First, since $1 - 3L_w^2\eta_t^2 \geq 0$, combining the last line of (71) and (70) of Lemma 14, we can show that

$$\begin{aligned} \|g_t - \nabla\Phi_0(\widetilde{w}_{t-1})\|^2 &\overset{(71)}{\leq} \tfrac{L_w^2}{n}\textstyle\sum_{j=1}^n \|w_{i-1}^{(t)} - w_0^{(t)}\|^2 + L_u^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\overset{(70)}{\leq} L_u^2\big(4L_w^2\eta_t^2+1\big)\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + 6L_w^2\sigma_w^2\eta_t^2 \\ &\quad + 2L_w^2(3\Theta_w+1)\eta_t^2\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2. \end{aligned} \tag{114}$$

Similarly, from the first line of (71) and (70), we also have

$$\begin{aligned} \|g_t - \nabla_w\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_t)\|^2 &\leq 4L_w^2L_u^2\eta_t^2\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 + 6L_w^2\sigma_w^2\eta_t^2 \\ &\quad + 2L_w^2(3\Theta_w+1)\eta_t^2\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2. \end{aligned} \tag{115}$$

Next, since $1 - 3L_u^2\hat{\eta}_t \geq 0$, combining (100) and (101) of Lemma 20, and (29) from Assumption 6, we have

$$\begin{aligned} \|v_t - \nabla_u\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2 &\leq L_u^2(3\Theta_u+2)\hat{\eta}_t^2\|\nabla_u\mathcal{H}(\widetilde{w}_{t-1}, \widetilde{u}_{t-1})\|^2 + 3L_u^2\sigma_u^2\hat{\eta}_t^2 \\ &\leq \hat{\Lambda}_0 L_u^2(3\Theta_u+2)\hat{\eta}_t^2\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \\ &\quad + \big(\hat{\Lambda}_1 L_u^2(3\Theta_u+2) + 3L_u^2\sigma_u^2\big)\hat{\eta}_t^2. \end{aligned} \tag{116}$$

Substituting (114), (115), and (116) into (107), and noting that $\mu_\psi := \mu_h + \mu_H > 0$, we obtain

$$\begin{aligned} \mathcal{V}_t - \mathcal{V}_{t-1} &\leq -\tfrac{\lambda-2-[(1+\lambda)L_{\Phi_0}+L_w+L_f]\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{(1-L_u\hat{\eta}_t)}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 \\ &\quad - \tfrac{(\lambda+1)(1-2L_{\Phi_0}\eta_t)\eta_t}{2}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + (\lambda+2)L_w^2(3\Theta_w+1)\eta_t^3\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{\hat{\eta}_t}{2}\big[1 - (L_u - \mu_\psi)\hat{\eta}_t - \tfrac{\hat{\Lambda}_0 L_u^2(3\Theta_u+2)\hat{\eta}_t^2}{1+\mu_h\hat{\eta}_t}\big]\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \\ &\quad + \tfrac{L_u^2\eta_t}{2}\big[4(\lambda+2)L_w^2\eta_t^2 + \lambda + 1\big]\|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\quad + 3(\lambda+2)L_w^2\sigma_w^2\eta_t^3 + \tfrac{[\hat{\Lambda}_1 L_u^2(3\Theta_u+2)+3L_u^2\sigma_u^2]\hat{\eta}_t^3}{2(1+\mu_h\hat{\eta}_t)}. \end{aligned} \tag{117}$$

Next, by Young's inequality, (92), and (93), we can show that

$$\begin{aligned} \|\widetilde{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 &\leq 3\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 + 3\|\widetilde{u}_{t-1} - \widehat{u}_t\|^2 + 3\|\widehat{u}_t - u_0^*(\widetilde{w}_{t-1})\|^2 \\ &\overset{(92),(93)}{\leq} 3\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 + 3\hat{\eta}_t^2\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \\ &\quad + \tfrac{3}{2\mu_\psi}\big[\psi(\widetilde{w}_{t-1}, \widehat{u}_t) - \psi(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\big]. \end{aligned}$$

Substituting the last inequality and $\|\nabla\Phi_0(\widetilde{w}_{t-1})\|^2 \leq \Lambda_0\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + \Lambda_1$ from (20) of Assumption 5 into (117), we can derive that

$$\begin{aligned} \mathcal{V}_t - \mathcal{V}_{t-1} &\leq -\tfrac{\lambda-2-[(1+\lambda)L_{\Phi_0}+L_w+L_f]\eta_t}{2\eta_t}\|\widetilde{w}_t - \widetilde{w}_{t-1}\|^2 - \tfrac{1-(L_u+3C_1\eta_t)\hat{\eta}_t}{2\hat{\eta}_t}\|\widetilde{u}_t - \widetilde{u}_{t-1}\|^2 \\ &\quad - \tfrac{\eta_t}{2}\big[(\lambda+1)(1-2L_{\Phi_0}\eta_t) - 2\Lambda_0(\lambda+2)L_w^2(3\Theta_w+1)\eta_t^2\big]\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 \\ &\quad - \tfrac{C_2\hat{\eta}_t}{2}\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 + \tfrac{3C_1\eta_t}{4\mu_\psi}\big[\psi(\widetilde{w}_{t-1}, \widehat{u}_t) - \psi(\widetilde{w}_{t-1}, u_0^*(\widetilde{w}_{t-1}))\big] \\ &\quad + \big[3(\lambda+2)L_w^2\sigma_w^2 + \Lambda_1(\lambda+2)L_w^2(3\Theta_w+1)\big]\eta_t^3 \\ &\quad + \tfrac{\big[\hat{\Lambda}_1 L_u^2(3\Theta_u+2)+3L_u^2\sigma_u^2\big]\hat{\eta}_t^3}{2(1+\mu_h\hat{\eta}_t)}, \end{aligned} \tag{118}$$

where $C_1 := L_u^2\big[4(\lambda+2)L_w^2\eta_t^2+\lambda+1\big]$ and $C_2 := 1-(L_u-\mu_h-\mu_H)\hat{\eta}_t-\frac{\hat{\Lambda}_0 L_u^2(3\Theta_u+2)\hat{\eta}_t^2}{1+\mu_h\hat{\eta}_t}-3C_1\eta_t\hat{\eta}_t$.

Finally, from (94), since $\hat{\eta}_t \leq \frac{2}{L_u+\mu_H}$, we also have

$$-\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \leq -\left(1-\tfrac{2L_u\mu_H\hat{\eta}_t}{L_u+\mu_H}\right)\|\hat{\mathcal{G}}_{\hat{\eta}_t}(\widetilde{u}_{t-1})\|^2 \leq -2\mu_\psi[\psi(\widetilde{w}_{t-1},\widehat{u}_t)-\psi(\widetilde{w}_{t-1},u_0^*(\widetilde{w}_{t-1}))].$$

Substituting this inequality into (118), we arrive at

$$\begin{aligned}
\mathcal{V}_t-\mathcal{V}_{t-1} \leq & -\tfrac{\lambda-2-[(1+\lambda)L_{\Phi_0}+L_w+L_f]\eta_t}{2\eta_t}\|\widetilde{w}_t-\widetilde{w}_{t-1}\|^2-\tfrac{1-(L_u+3C_1\eta_t)\hat{\eta}_t}{2\hat{\eta}_t}\|\widetilde{u}_t-\widetilde{u}_{t-1}\|^2\\
& -\tfrac{\eta_t}{2}\big[(\lambda+1)(1-2L_{\Phi_0}\eta_t)-2\Lambda_0(\lambda+2)L_w^2(3\Theta_w+1)\eta_t^2\big]\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2\\
& -\big(\mu_\psi C_2\hat{\eta}_t-\tfrac{3C_1\eta_t}{4\mu_\psi}\big)\big[\psi(\widetilde{w}_{t-1},\widehat{u}_t)-\psi(\widetilde{w}_{t-1},u_0^*(\widetilde{w}_{t-1}))\big]\\
& +(\lambda+2)L_w^2\big[3\sigma_w^2+\Lambda_1(3\Theta_w+1)\big]\eta_t^3+\tfrac{\big[\hat{\Lambda}_1L_u^2(3\Theta_u+2)+3L_u^2\sigma_u^2\big]\hat{\eta}_t^3}{2(1+\mu_h\hat{\eta}_t)},
\end{aligned}$$

which is exactly (112). $\square$

Now, we are ready to prove the convergence of Algorithm 2 using only **one epoch** (i.e. $S=1$) of the **shuffling routine** (27). The following theorem is the full version of Theorem 5 in the main text.

Theorem 9. *Suppose that Assumptions 1, 2, 4, 5, 6, and 7 hold for (1) under the (NC) setting. Let Ψ_0 be defined by (3), and $\mathcal{G}_\eta$ be defined by (18). Let us denote C_w and C_u respectively by*

$$C_w := 5L_w^2\big[\Lambda_1(3\Theta_w+1)+3\sigma_w^2\big] \quad \textit{and} \quad C_u := \tfrac{L_u^2}{2}\big[\hat{\Lambda}_1(3\Theta_u+2)+3\sigma_u^2\big]. \tag{119}$$

*Let $\{(\widetilde{w}_t,\widetilde{u}_t)\}$ be generated by Algorithm 2 using only **one epoch** (i.e. $S=1$) of the **shuffling routine** (27), and fixed learning rates $\eta_t := \eta \in (0,\bar{\eta}]$ and $\hat{\eta}_t := \hat{\eta} := 15\kappa^2\eta$, where*

$$\bar{\eta} := \min\left\{\tfrac{1}{60\kappa^2L_u},\ \tfrac{1}{\sqrt{10\Lambda_0L_w^2(3\Theta_w+1)}},\ \tfrac{2\sqrt{L_u}}{\kappa\sqrt{15(4L_u^2+\mu_\psi^2)}},\ \tfrac{\sqrt{L_u}}{15\kappa\sqrt{2L_u^3\hat{\Lambda}_0(3\Theta_u+2)+\mu_\psi^2}},\ \tfrac{1}{4L_{\Phi_0}+L_w+L_f}\right\}.$$

Then, the following bound holds:

$$\tfrac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \tfrac{24[\Psi_0(\widetilde{w}_0)-\Psi_0^\star]+8[\Psi_0(\widetilde{w}_0)-\mathcal{L}(\widetilde{w}_0,\widetilde{u}_0)]}{\eta(T+1)}+8C_w\eta^2+\tfrac{8C_u\hat{\eta}^3}{\eta} \tag{120}$$

For a given $\epsilon>0$, if $\eta := \mathcal{O}(\epsilon)\in(0,\bar{\eta}]$ and $T := \mathcal{O}\big(\frac{1}{\epsilon^3}\big)$, then $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \epsilon^2$.

Consequently, Algorithm 2 requires $\mathcal{O}\big(\frac{n}{\epsilon^3}\big)$ evaluations of both $\nabla_w\mathcal{H}_i$ and $\nabla_u\mathcal{H}_i$, and $\mathcal{O}(\epsilon^{-3})$ evaluations of $\mathrm{prox}_{\eta_t f}$ and $\mathrm{prox}_{\hat{\eta}_t h}$ to achieve an ϵ-stationary point $\widehat{w}_T$ of (1) computed by (19).

Proof. Let us choose $\lambda := 3$, $\eta_t := \eta > 0$ and $\hat{\eta}_t := \hat{\eta} > 0$. First, we need to guarantee that $1-3L_w^2\eta^2 \geq 0$ and $1-3L_u^2\hat{\eta}_t \geq 0$ in Theorem 23. Suppose that $\eta \leq \frac{1}{4L_{\Phi_0}+L_w+L_f}$. Then, since $L_{\Phi_0} = (1+\kappa)L_w \geq L_w$, we have $\eta \leq \frac{1}{5L_w}$, which obviously guarantees that $1-3L_w^2\eta^2 \geq 0$.

Moreover, for C_i for $i=0,\cdots,5$ defined by (113), we can show that

$$\begin{cases}
C_0 := 1-(4L_{\Phi_0}+L_w+L_f)\eta \geq 0,\\
C_1 := L_u^2(4+20L_w^2\eta^2) \leq 5L_u^2,\\
C_2 := 1-(L_u-\mu_\psi)\hat{\eta}-\frac{\hat{\Lambda}_0L_u^2(3\Theta_u+2)\hat{\eta}^2}{1+\mu_h\hat{\eta}}-6C_1\eta\hat{\eta}^2 \geq 1-L_u\hat{\eta}-\hat{\Lambda}_0L_u^2(3\Theta_u+2)\hat{\eta}^2-30L_u^2\eta\hat{\eta}^2,\\
C_3 := 4-8L_{\Phi_0}\eta-10\Lambda_0L_w^2(3\Theta_w+1)\eta^2 \geq 2\big[1-5\Lambda_0L_w^2(3\Theta_w+1)\eta^2\big],\\
C_4 := 5L_w^2\big[\Lambda_1(3\Theta_w+1)+3\sigma_w^2\big] = C_w,\\
C_5 := \frac{\hat{\Lambda}_1L_u^2(3\Theta_u+2)+3L_u^2\sigma_u^2}{2(1+\mu_h\hat{\eta})} \leq C_u.
\end{cases}$$

Now, suppose that

$$\begin{cases}
L_u\hat{\eta} \leq \frac{1}{4}, \quad \hat{\Lambda}_0L_u^2(3\Theta_u+2)\hat{\eta}^2+30L_u^2\eta\hat{\eta}^2 \leq \frac{1}{2}, \quad (L_u+15L_u^2\eta)\hat{\eta} \leq 1,\\
4C_2\mu_\psi^2\hat{\eta} \geq 3C_1\eta, \quad \text{and} \quad 5\Lambda_0L_w^2(3\Theta_w+1)\eta^2 \leq \frac{1}{2},
\end{cases} \tag{121}$$

then we can easily show that $C_2 \geq \frac{1}{4}$, $C_3 \geq 1$, $1-(L_u+3C_1\eta)\hat{\eta} \geq 0$, and $\mu_\psi C_2\hat{\eta}-\frac{3C_1\eta}{4\mu_\psi} \geq 0$.

In this case, (112) reduces to

$$\mathcal{V}_t \leq \mathcal{V}_{t-1} - \tfrac{\eta}{8}\|\mathcal{G}_{\eta_t}(\widetilde{w}_{t-1})\|^2 + C_w\eta^3 + C_u\hat{\eta}^3. \tag{122}$$

By induction, we obtain (120) from (122) and $\mathcal{V}_0 := 3[\Psi_0(\widetilde{w}_0) - \Psi_0^\star] + \Psi_0(\widetilde{w}_0) - \mathcal{L}(\widetilde{w}_0, \widetilde{u}_0)$.

From (121), let us choose $\hat{\eta} = \frac{15L_u^2}{\mu_\psi^2}\eta = 15\kappa^2\eta$ with $\kappa := \frac{L_u}{\mu_\psi}$. Then, we can verify the five conditions of (121) as follows.

- We have $4C_2\mu_\psi^2\hat{\eta} \geq \mu_\psi^2\hat{\eta} = 15L_u^2\eta \geq 3C_1\eta$, which satisfies the fourth condition of (121).
- If $\eta \leq \frac{1}{60L_u\kappa^2}$, then the condition $L_u\hat{\eta} \leq \frac{1}{4}$ in (121) holds. This condition also guarantees $1 - 3L_u^2\hat{\eta}^2 \geq 0$.
- If $\eta \leq \frac{1}{\sqrt{10\Lambda_0 L_w^2(3\Theta_w+1)}}$, then the last condition $5\Lambda_0 L_w^2(3\Theta_w + 1)\eta^2 \leq \frac{1}{2}$ of (121) holds.
- If $\eta \leq \frac{2\sqrt{L_u}}{\kappa\sqrt{15(4L_u^2+\mu_\psi^2)}}$, then the condition $(L_u + 15L_u^2\eta)\hat{\eta} \leq 1$ of (121) holds.
- Finally, if $\eta \leq \frac{\sqrt{L_u}}{15\kappa\sqrt{2L_u^3\hat{\Lambda}_0(3\Theta_u+2)+\mu_\psi^2}}$, then the second condition $\hat{\Lambda}_0 L_u^2(3\Theta_u + 2)\hat{\eta}^2 + 30L_u^2\eta\hat{\eta}^2 \leq \frac{1}{2}$ of (121) also holds.

Overall, we can conclude that if we choose $\eta \in (0, \bar{\eta}]$ as in Theorem 9, where

$$\bar{\eta} := \min\left\{\frac{1}{60\kappa^2 L_u}, \frac{1}{\sqrt{10\Lambda_0 L_w^2(3\Theta_w+1)}}, \frac{2\sqrt{L_u}}{\kappa\sqrt{15(4L_u^2+\mu_\psi^2)}}, \frac{\sqrt{L_u}}{15\kappa\sqrt{2L_u^3\hat{\Lambda}_0(3\Theta_u+2)+\mu_\psi^2}}, \frac{1}{4L_{\Phi_0}+L_w+L_f}\right\},$$

then all the conditions in (121) are satisfied. In addition, since $\mu_H \leq L_u$, we have $L_u + \mu_H \leq 2L_u$. Thus the condition $\eta \leq \frac{1}{60\kappa^2 L_u}$ implies $\hat{\eta} \leq \frac{2}{L_u+\mu_H}$ due to $\hat{\eta} = 15\kappa^2\eta$.

Finally, to achieve $\frac{1}{T+1}\sum_{t=0}^{T}\|\mathcal{G}_\eta(\widetilde{w}_t)\|^2 \leq \epsilon^2$, we impose

$$\tfrac{24[\Psi_0(\widetilde{w}_0)-\Psi_0^\star]+8[\Psi_0(\widetilde{w}_0)-\mathcal{L}(\widetilde{w}_0,\widetilde{u}_0)]}{\eta(T+1)} + 8\big(C_w + 15^3\kappa^6 C_u\big)\eta^2 \leq \epsilon^2.$$

If we choose $\eta := \mathcal{O}(\epsilon) \in (0, \bar{\eta}]$ sufficiently small, and $T := \mathcal{O}(\epsilon^{-3})$, then the last condition holds.

At each epoch t, Algorithm 2 requires n evaluations of both $\nabla_w\mathcal{H}_i$ and $\nabla_u\mathcal{H}_i$. Therefore, the total evaluation of $\nabla_w\mathcal{H}_i$ and $\nabla_u\mathcal{H}_i$ is $\mathcal{T}_e := nT = \mathcal{O}\big(n\epsilon^{-3}\big)$. Similarly, since each epoch t, Algorithm 2 requires one evaluation of $\mathrm{prox}_{\eta_t f}$, and one evaluation of $\mathrm{prox}_{\hat{\eta}_t h}$, the total number of both $\mathrm{prox}_{\eta_t f}$ and $\mathrm{prox}_{\eta_t f}$ evaluations is $T = \mathcal{O}(\epsilon^{-3})$. □

D Details and Additional Results of Numerical Experiments

This section provides the details of our experiments in Section 5 and also adds more experiments to illustrate our algorithms and compares them with two other methods. All the algorithms we experiment in this paper are implemented in Python and are run on a MacBook Pro. 2.8GHz Quad-Core Intel Core I7, 16Gb Memory.

D.1 Details of Numerical Experiments in Section 5

We have abbreviated Algorithm 1 by `SGM` in Figure 1. Since we have two options to construct estimator $F_i^{(t)}$ for $F(\widetilde{w}_{t-1})$, we name `SGM-Option 1` for Algorithm 1 using (21), and `SGM-Option 2` for Algorithm 1 using (22).

Implementation details and competitors. Since $\phi_0(v) = \max_{\|u\|_1\leq 1}\{\langle v, u\rangle\}$ in our model (31) is nonsmooth, we have implemented two other algorithms, `SGD` in [10] – a variant of the stochastic gradient method for compositional minimization, and `Prox-Linear` in [11] – a type of the Gauss-Newton method with variance-reduction using large mini-batches for compositional minimization. Since `SGD` only works for smooth ϕ_0, we have smoothed it as in our method, and utilized the estimator and algorithm from [10], but also updated the smoothness parameter as in our method. Here, we only compare the performance of all algorithms in terms of epochs (i.e. the number of data passes) and ignore their computational time since Prox-Linear becomes slower if p is getting large. This is due to its expensive subproblem of evaluating the prox-linear operator.

To compare with SGD and Prox-Linear, we only use Algorithm 1 since both SGD and Prox-Linear are designed to solve compositional minimization problems of the form (CO). However, Prox-Linear requires to solve a nonsmooth convex subproblem to evaluate the prox-linear operator. Therefore, we have implemented a first-order primal-dual scheme in [1] to evaluate this operator, which we believe that it is an efficient method.

Parameter selection. To boost the performance of all algorithms, we implement mini-batch variants of these methods instead of a single sample variant. Our batch size b is computed by $b := \lfloor \frac{n}{k_b} \rfloor$, where n is the number of data points and k_b is the number of blocks. In our experiments, we have also varied the number of blocks k_b to observe the performance of these algorithms. Since we want to obtain good performance, instead of using their theoretical learning rates, we have carefully tuned the learning rate η of all algorithms in a given set of candidates $\{100, 50, 10, 5, 1, 0.5, 0.1, 0.05, 0.01, 0.001, 0.0001\}$. We find $\eta = 5$ (i.e. $\eta_t = 10^{-4}$) for **w8a** and $\eta = 100$ (i.e. $\eta_t = 5 \times 10^{-5}$) for **rcv1** which work well for our method. We also update the smoothness parameter γ as $\gamma := \frac{1}{2(t+1)^{1/3}}$ w.r.t. to the epoch counter t instead of fixing it at a small value. For **w8a**, we find $\eta = 0.05$ as a good learning rate for both SGD and Prox-Linear. For **rcv1**, we get $\eta = 0.5$ for both algorithms. All experiments are run up to 200 epochs.

The convergence of gradient mapping norm. Figure 1 only reveals the objective values of (31) against the number of epochs. Figure 2 below shows the absolute norm of the gradient mapping $\|\mathcal{G}_\eta(\widetilde{w}_t)\|$ for this experiment.

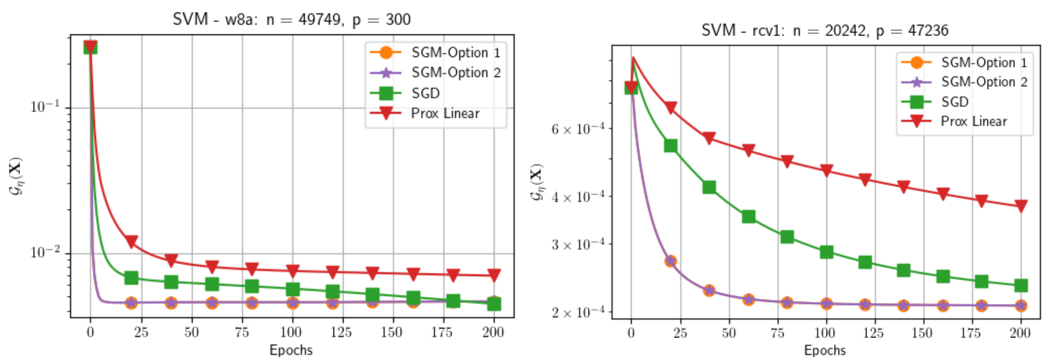

Figure 2: The performance of 4 algorithms for solving (31) in terms of gradient mapping norm.

It seems that both options, SGM-Option 1 and SGM-Option 2 are almost identical for this test. For **w8a**, our methods look like having comparable performance with both SGD and Prox-Linear, just slightly better. For **rcv1**, our methods reach a better approximate solution earlier than SGD, but after more than 200 epochs, SGD tends to approach a similar accuracy level. Prox-Linear has a significantly worse performance than ours and SGD in this particular experiment.

D.2 Additional Experiments

We provide additional experiments to test our algorithms and compare them with SGD and Prox-Linear as in Section 5.

The effect of mini-batch size. Our first test is to verify if the mini-batch size b actually affects the performance of these algorithms. We use the same datasets and the same parameters as in Section 5, but reduce b by increasing k_b from 32 to 64 blocks. Figure 3 reveals the performance of 4 algorithms on two datasets with $k_b = 64$: **w8a** corresponding to $b = 777$ and **rcv1** corresponding to $b = 316$.

With this choice of mini-batches, our algorithms still have a similar performance as SGD, while Prox-Linear does not really improve its performance, and slightly gets worse. Note that Prox-Linear requires a large mini-batch to achieve a variance reduce, and decreasing this mini-batch size indeed affects its performance.

Different learning rates. Now, let us test our algorithms using different learning rates, we only focus on **Option 2** as both options show similar results in our tests. For **w8a**, we choose 4 different learning rates $\eta = 0.5, 2.5, 5.0$, and 7.5, while maintaining $k_b = 64$. For **rcv1**, we also choose 4 different

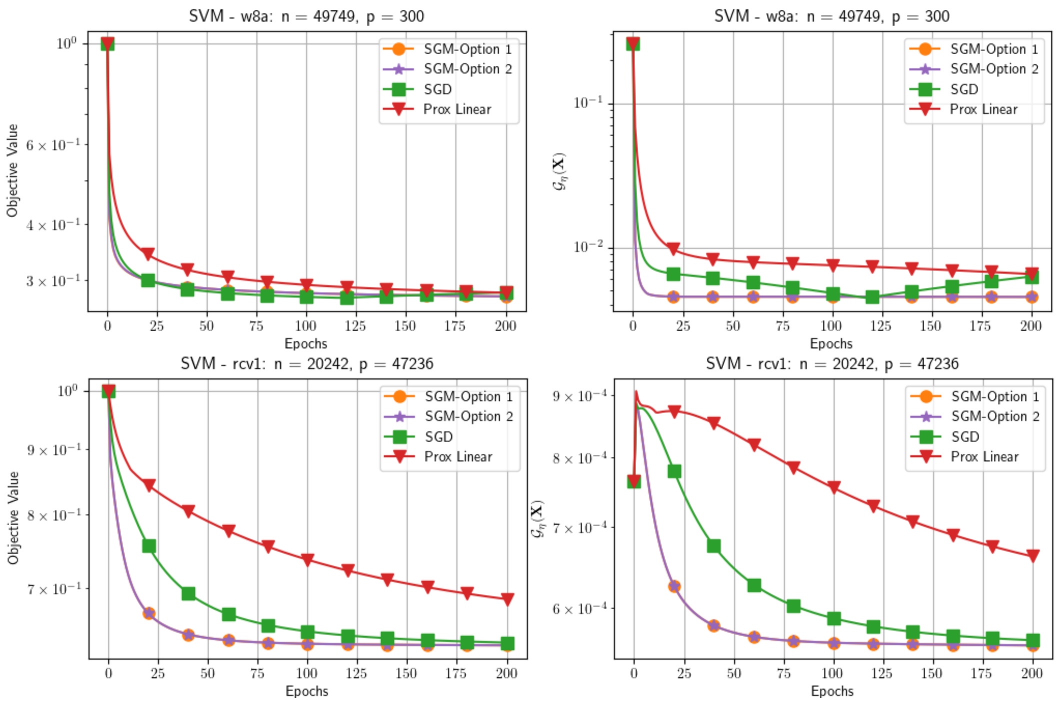

Figure 3: The performance of 4 algorithms on two different datasets with $k_b = 64$.

learning rates $\eta = 25, 50, 100$, and 125. The results of this experiment are plotted in Figure 4 for both **w8a** and **rcv1** datasets.

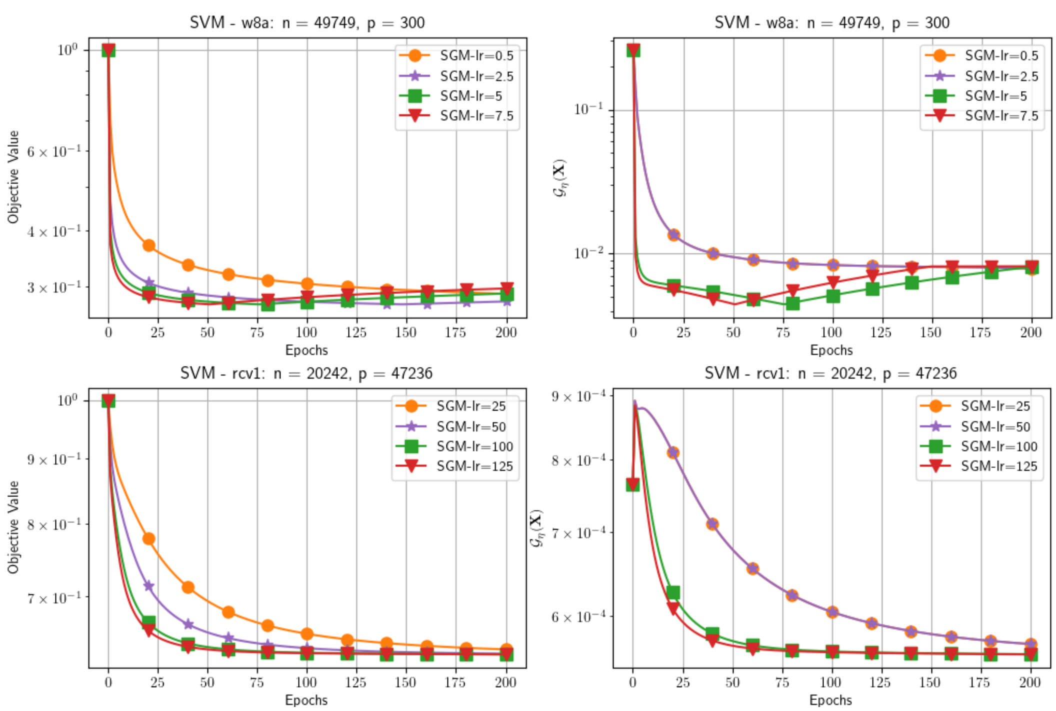

Figure 4: The performance of Algorithm 1 with 4 different learning rates η and $k_b = 64$ on 2 datasets.

As we can see from Figure 4 that

- For **w8a**, our method starts diverging when $\eta = 7.5$, while still works well for smaller learning rates. For $\eta = 0.25$, it indeed has a slow progress in early iterations as often seen in SGD.
- For **rcv1**, we also observe similar behaviors as in **w8a**, but with larger learning rates than $\eta = 125$.

Large dataset. We have also run our algorithms and their competitors on a bigger dataset from LIBSVM: **url** with $n = 2,396,130$ and $p = 3,231,951$. Here, we use a learning rate $\eta = 1$ for our methods, which corresponds to $\eta_t = 4.2 \times 10^{-7}$. For `SGD`, we use a learning rate $\eta = 0.01$ and for `Prox-Linear`, we use a learning rate $\eta = 0.01$ after tuning both methods. We also set $k_b = 64$ for all algorithms. The results of this experiment are reported in Figure 5.

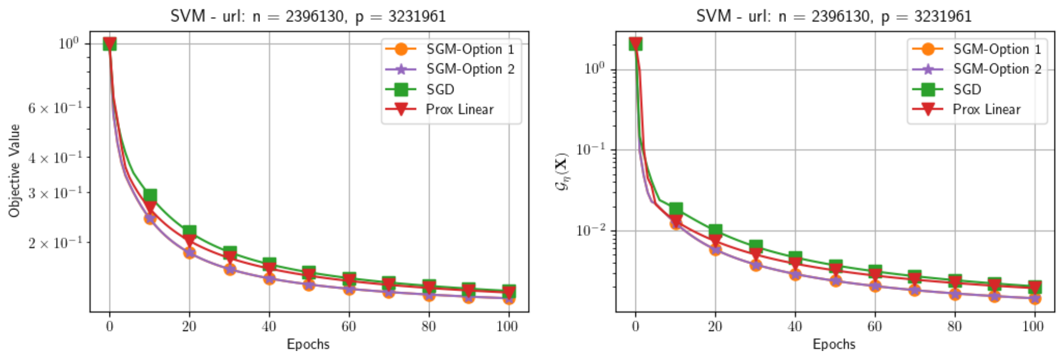

Figure 5: The performance of 4 algorithms on a large dataset: **url**.

As we can see from Figure 5, our methods have a comparable performance with their competitors. All algorithms have similar behavior in terms of convergence.

References

[1] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

[2] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Math. Program.*, 64:81–101, 1994.

[3] H. Cho and C. Yun. SGDA with shuffling: faster convergence for nonconvex-PŁ minimax optimization. *The 11th International Conference on Learning Representations*, pp. 1–10, 2022.

[4] K. Mishchenko, A. Khaled, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

[5] K. Mishchenko, A. Khaled, and P. Richtárik. Proximal and federated random reshuffling. In *International Conference on Machine Learning*, pages 15718–15749. PMLR, 2022.

[6] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

[7] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[8] L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.

[9] Q. Tran-Dinh, D. Liu, and L. M. Nguyen. Hybrid variance-reduced SGD algorithms for nonconvex-concave minimax problems. *The 34th Conference on Neural Information Processing Systems (NeurIPs 2020)*, 2020.

[10] M. Wang, E. Fang, and L. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math. Program.*, 161(1-2):419–449, 2017.

[11] J. Zhang and L. Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Mathematical Programming*, pp. 1–43, 2022.