Variance-reduced accelerated methods for decentralized stochastic double-regularized nonconvex strongly-concave minimax problems

Gabriel Mancino-Ball · Yangyang Xu

July 17, 2023

Abstract In this paper, we consider the decentralized, stochastic nonconvex strongly-concave (NCSC) minimax problem with nonsmooth regularization terms on both primal and dual variables, wherein a network of m computing agents collaborate via peer-to-peer communications. We consider when the coupling function is in expectation or finite-sum form and the double regularizers are convex functions, applied separately to the primal and dual variables. Our algorithmic framework introduces a Lagrangian multiplier to eliminate the consensus constraint on the dual variable. Coupling this with variance-reduction (VR) techniques, our proposed method, entitled `VRLM`, by a single neighbor communication per iteration, is able to achieve an $\mathcal{O}(\kappa^3\varepsilon^{-3})$ sample complexity under the general stochastic setting, with either a big-batch or small-batch VR option, where κ is the condition number of the problem and ε is the desired solution accuracy. With a big-batch VR, we can additionally achieve $\mathcal{O}(\kappa^2\varepsilon^{-2})$ communication complexity. Under the special finite-sum setting, our method with a big-batch VR can achieve an $\mathcal{O}(n+\sqrt{n}\kappa^2\varepsilon^{-2})$ sample complexity and $\mathcal{O}(\kappa^2\varepsilon^{-2})$ communication complexity, where n is the number of components in the finite sum. All complexity results match the best-known results achieved by a few existing methods for solving special cases of the problem we consider. To the best of our knowledge, this is the first work which provides convergence guarantees for NCSC minimax problems with general convex nonsmooth regularizers applied to both the primal and dual variables in the decentralized stochastic setting. Numerical experiments are conducted on two machine learning problems. Our code is downloadable from https://github.com/RPI-OPT/VRLM.

Keywords: Stochastic optimization, decentralized optimization, minimax problems, variance reduction

Mathematics Subject Classification: 90C15, 90C26, 90C47, 65K05

G. Mancino-Ball
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY
E-mail: gabriel.mancino.ball@gmail.com
Y. Xu
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY
E-mail: xuy21@rpi.edu

1 Introduction

In this paper, we consider the following minimax-structured problem

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1}} \max_{\mathbf{y}\in\mathbb{R}^{d_2}} \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x},\mathbf{y}) + g(\mathbf{x}) - h(\mathbf{y}), \text{ with } f_i(\mathbf{x},\mathbf{y}) := \mathbb{E}_{\xi\sim\mathcal{D}_i}\left[\tilde{f}_i(\mathbf{x},\mathbf{y};\xi)\right], \tag{1.1}$$

where f_i is a smooth, nonconvex strongly-concave (NCSC) function governed by $\mathcal{D}_i$ for each $i = 1,\dots,m$. In the special case where $\mathcal{D}_i$ is a discrete uniform distribution, we recover the finite-sum problem setting, and without lose of generality, we can assume each $\mathcal{D}_i$ involves the same number of scenarios, i.e.,

$$f_i(\mathbf{x},\mathbf{y}) := \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \tilde{f}_i(\mathbf{x},\mathbf{y};\xi_{ij}) \tag{1.2}$$

Thus (1.1) encompasses a broad class of problems. We assume g and h are closed convex functions, often serving as regularizers, hence we refer to (1.1) as a *double-regularized minimax* problem. Problem (1.1) has received a lot of research attention recently due to its application in many machine learning settings such as adversarial training [11, 22], distributionally robust optimization [31, 44], and reinforcement learning [54]. It has also been used to study fairness in machine learning [35] and improving generalization error [9].

Motivated by scenarios where data are distributed over many different computing devices [47], we are interested in the case where f_i is owned *privately* by the i-th one among m agents that are connected over a communication network $\mathcal{G} = (\mathcal{V},\mathcal{E})$. Here, $\mathcal{V} = \{1,\dots,m\}$ denotes the set of agents and $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$ denotes the set of feasible communication links between the agents. We assume the network $\mathcal{G}$ is connected. Furthermore, in a stochastic setting, we assume that each agent i can only access local stochastic gradients, rather than exact gradients. Such a scheme is close to a real-world setting where agents may not have access to a global communication protocol (either due to privacy restrictions [43] or high communication overhead [19]), nor access to the entire local gradient (e.g., when data arrives in a stream [49]).

In order for the m agents to collaboratively solve (1.1), each agent i will maintain a copy of the primal-dual variable $(\mathbf{x},\mathbf{y})$, denoted as $(\mathbf{x}_i,\mathbf{y}_i)$. With the introduction of local variables $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i\in\mathcal{V}}$, we can formulate (1.1) equivalently into the following decentralized consensus minimax problem

$$\min_{\mathbf{x}_1,\dots,\mathbf{x}_m} \max_{\mathbf{y}_1,\dots,\mathbf{y}_m} \frac{1}{m}\sum_{i=1}^{m} \big(f_i(\mathbf{x}_i,\mathbf{y}_i) + g(\mathbf{x}_i) - h(\mathbf{y}_i)\big), \text{ s.t. } (\mathbf{x}_i,\mathbf{y}_i) = (\mathbf{x}_j,\mathbf{y}_j),\ \forall\, i,j\in\mathcal{V}. \tag{1.3}$$

To ensure that the consensus constraint $(\mathbf{x}_i,\mathbf{y}_i) = (\mathbf{x}_j,\mathbf{y}_j)$ is satisfied for all i, $j \in \mathcal{V}$, agents exchange their information with their 1-hop neighbors via the communication links in $\mathcal{E}$. Mathematically, this communication is represented as multiplication with a *mixing matrix*, $\mathbf{W} \in \mathbb{R}^{m\times m}$, which serves as an inexact averaging operation among the agents.

The efficiency of a method to solve (1.3) will be measured by the number of samples and neighbor communications required to achieve an ε-accurate solution; see Definition 3.1. We refer to these quantities as *sample* and *communication* complexities, respectively. In this work, we propose the `VRLM` method which serves as a framework for applying state-of-the-art variance-reduction (VR) techniques to achieve low sample *and* communication complexities when solving (1.3). Our framework is based on a reformulation of (1.3) which allows for exploitation of the strong-concavity of $f_i(\mathbf{x},\cdot)$. We summarize our contributions below.

1.1 Contributions

Our contributions are three-fold. They are summarized as follows.

- *First*, we provide one decentralized method with two VR options for solving stochastic double-regularized NCSC minimax problems in the form of (1.1) or the equivalent form (1.3). Our method is based on a reformulation of (1.3) introduced in [48], by removing the consensus constraint on the dual variable with the introduction of a Lagrangian multiplier. One VR option of our method uses large-batch sampling, while the other needs only $\mathcal{O}(1)$ samples for each update per agent. To the best of our knowledge, this is the first decentralized stochastic method for solving the structured problem (1.1), while a few existing methods can only be applied to certain special cases; see more details in the section of related work. Compared to most existing decentralized stochastic methods that even can only solve special cases of (1.1), our method can take $\Theta(\kappa)$ times larger stepsize, where κ is the condition number defined in (3.8). This can potentially lead to better empirical performance; see Remark 3.4.
- *Second*, we show a last-iterate convergence in probability result for our method with the large-batch VR option on solving finite-sum structured problems. Also, we establish complexity results of our method with either VR option to produce an ε-stationary solution for any given $\varepsilon > 0$; see Definition 3.1. Utilizing just a single neighbor communication per iteration, we prove that both versions of our method can achieve a sample complexity of $\mathcal{O}\left(\kappa^3\varepsilon^{-3}\right)$ in the general stochastic setting. The small-batch version attains the same communication complexity, while the large-batch one only needs $\mathcal{O}\left(\kappa^2\varepsilon^{-2}\right)$ communication rounds. In addition, for the finite-sum structured problem, our method with the large-batch VR option can achieve a sample complexity of $\mathcal{O}(n + \sqrt{n}\kappa^2\varepsilon^{-2})$ and communication complexity of $\mathcal{O}(\kappa^2\varepsilon^{-2})$. These complexity results are optimal in terms of the dependence on ε [1,26] and match the best-known results of decentralized methods for solving special cases of (1.1). Moreover, our complexity results are graph-topology independent in certain regimes; see Remarks 3.4 and 3.5. Furthermore, when specialized to the single-agent setting, i.e., $m = 1$ in (1.1), our method also improves over a few state-of-the-art methods for NCSC finite-sum or stochastic minimax problems. With large-batch sampling, our method achieves the same-order complexity result as SREDA [27], which requires double loops and only applies to the special case of (1.1) with $g \equiv 0$ and $h = \mathbb{I}_{\mathcal{Y}}$ for a convex set $\mathcal{Y}$. If $\mathcal{O}(1)$ samples are available for each update, our method achieves a lower-order complexity than Acc-MDA [12], which needs $\tilde{\mathcal{O}}(\kappa^{\frac{9}{2}}\varepsilon^{-3})$ sample gradients to produce an ε-stationary solution.
- *Third*, we verify the performance of our proposed method on two machine learning problems in a real decentralized computing environment. Additionally, we make our code open source at https://github.com/RPI-OPT/VRLM.

1.2 Related work

A few decentralized methods have been proposed for solving minimax problems. However, most of them focus on deterministic or finite-sum structured problems. We list a few representative methods in Table 1.

The recent D-GDMax method [48] is very closely related to our proposed method, as both rely on a reformulation of (1.3) by the introduction of a Lagrangian multiplier to facilitate convergence. A key difference is that D-GDMax requires exact gradients. Thus it is not applicable to the general stochastic problem setting that we consider. Though D-GDMax can be applied to the special finite-sum case, its complexity will be significantly higher than ours if n is big and the desired accuracy ε is small. GT-DA [42]

Table 1 Representative decentralized minimax optimization methods for (1.1). The STOCH./F.S. column indicates whether the method handles a stochastic setting or a finite-sum setting in (1.2); S.C. stands for "single communication" and hence an ✗ in this column indicates a method requires multiple communications to ensure convergence. The final two columns, SAMP. COMP. and COMM. COMP., indicate the complexity results for solving the associated problem, e.g. for `VRLM`, these results are in terms of solving (2.2). Here, the $\mathcal{O}(\cdot)$ notation hides dependence on non-key values; some works do not make the dependence on the spectral gap or condition number clear - we use constants $a, b, c, d, e, f > 0$ with subscripts s and c to denote whether or not these unknowns are related to the *sample* or *communication* complexities, respectively. ¶PRECISION is guaranteed to converge in the finite-sum setting, but the dependence upon the number of local component functions n is unclear. †We assume here that $\varepsilon \leq (1-\rho)^2$. ‡We assume here that $1 \leq \kappa(1-\rho)^2$; in these two regimes, our results are graph-topology independent. For complete results, please see Corollaries 1 and 2.

METHOD	STOCH./F.S.	$(g,\ h)$	S.C.	SAMP. COMP.	COMM. COMP.
GT-DA [42]	F.S.	$(0,\ \mathbb{I}_{\mathcal{Y}})$	✓	$\tilde{\mathcal{O}}\left(\frac{n\kappa^{a_s}}{(1-\rho)^{b_s}\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{\kappa^{a_c}}{(1-\rho)^{b_c}\varepsilon^2}\right)$
GT-SRVR [54]	F.S.	$(0,\ 0)$	✓	$\mathcal{O}\left(n+\frac{\sqrt{n}\kappa^{c_s}}{(1-\rho)^{d_s}\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\kappa^{c_c}}{(1-\rho)^{d_c}\varepsilon^2}\right)$
PRECISION [23]¶	F.S.	$(\text{convex},\ \mathbb{I}_{\mathcal{X}})$	✓	$\mathcal{O}\left(\frac{\kappa^{e_s}}{(1-\rho)^{f_s}\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\kappa^{e_c}}{(1-\rho)^{f_c}\varepsilon^2}\right)$
DREAM [2]	F.S.	$(0,\ \mathbb{I}_{\mathcal{Y}})$	✗	$\mathcal{O}\left(n+\frac{\sqrt{n}\kappa^2}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\kappa^2}{\sqrt{1-\rho}\varepsilon^2}\right)$
	STOCH.			$\mathcal{O}\left(\frac{\kappa^3}{\varepsilon^3}\right)$	$\mathcal{O}\left(\frac{\kappa^2}{\sqrt{1-\rho}\varepsilon^2}\right)$
`VRLM`-STORM†	STOCH.	(convex, convex)	✓	$\mathcal{O}\left(\frac{\kappa^3}{\varepsilon^3}\right)$	$\mathcal{O}\left(\frac{\kappa^3}{\varepsilon^3}\right)$
`VRLM`-SPIDER‡	F.S.	(convex, convex)	✓	$\mathcal{O}\left(n+\frac{\sqrt{n}\kappa^2}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\kappa^2}{\varepsilon^2}\right)$
	STOCH.			$\mathcal{O}\left(\frac{\kappa^3}{\varepsilon^3}\right)$	$\mathcal{O}\left(\frac{\kappa^2}{\varepsilon^2}\right)$

is also a deterministic gradient method. It solves a variant of (1.3) by only enforcing consensus on either the $\{\mathbf{x}_i\}_{i\in\mathcal{V}}$ or $\{\mathbf{y}_i\}_{i\in\mathcal{V}}$ variables, but not both simultaneously.

The PRECISION method in [23] is designed for solving (1.3) under the finite-sum setting, and in addition, it assumes $h = \mathbb{I}_{\mathcal{Y}}$ for some convex set $\mathcal{Y}$. Similar to one option of our method, it utilizes the SPIDER-type [8] variance reduction. In order to produce an ε-accurate point, it needs $\mathcal{O}(n+\sqrt{n}\varepsilon^{-2})$ samples and $\mathcal{O}(\varepsilon^{-2})$ communications rounds without giving an explicit dependence on the problem's condition number κ; the dependence on ε and n is the best-known for the finite-sum setting. A preceding method of PRECISION, called GT-SRVR in [54], considers a more special case of (1.3) under the finite-sum setting. It assumes $g \equiv 0$ and $h \equiv 0$ and achieves the same-order complexity results by the SPIDER-type variance reduction. DSGDA [10] solves the same-structured problem as GT-SRVR and also enjoys the same-order complexity results. Different from GT-SRVR, DSGDA uses a SAGA-type acceleration technique [6]. It does not need to take large-batch samples for each update but requires a large memory to maintain n component gradients.

Like our method, DREAM [2] can be applied to both the stochastic and finite-sum settings of (1.3). It employs a loopless version of the SPIDER-type VR method. However, it assumes $g \equiv 0$ and $h = \mathbb{I}_{\mathcal{Y}}$. In addition, it performs multiple communications per update, and its convergence results rely on the multi-communication trick. This is fundamentally different from our method. With a multi-communication trick, our results can have a lower-order dependence on the spectral of the communication network; see Remark 3.4. However, we can have guaranteed convergence with a single communication per update, while the analysis of DREAM requires the multi-communication trick to prove convergence. The requirement of multiple communications can be too restrictive and even impractical in a real computing setting, as it needs more coordinations between agents [19]. All the aforementioned methods require either large-batch samplings or a large number of maintained component gradients. In contrast, DM-HSGD [44], by the STORM-type [4]

VR technique, only needs $\mathcal{O}(1)$ samples per update per agent, except for a possible large-batch sampling at the initial step. However, there is one critical error with its convergence analysis. It turns out that Eqn. (28) in [44] does not hold with the given choice of θ. In fact, θ must be $\Theta(1/L)$ instead of the given $\Theta(1/\mu)$, where L is the smoothness constant of $\{f_i\}$ and μ the strong-concavity constant of $\{f_i(\mathbf{x},\cdot)\}$. With the correct θ, it is unclear if its final claimed convergence results will hold[1].

All the methods we mentioned above assume strong concavity about the dual variable $\mathbf{y}$ and also convexity of the regularization terms or constraint sets, if there are any. A few existing decentralized methods for minimax problems make weaker or different assumptions. For example, DPOSG [22] is designed to solve smooth stochastic nonconvex nonconcave decentralized minimax problems. Instead of concavity structure, the Minty VI condition is assumed by DPOSG in order to have guaranteed convergence. However, its complexity has a high-order dependence on a given accuracy, reaching $\mathcal{O}(\varepsilon^{-12})$ to produce an ε-stationary solution. In addition, it cannot handle problems with nonsmooth regularization terms or a hard constraint. This is another fundamental difference from our method.

Though our focus is on decentralized computation, our method is also applicable in the single-agent (or non-distributed) setting, i.e., the problem (1.1) with $m = 1$. In this case, many methods have been proposed under various settings of f. The GDA [20] method can be applied to dual-constrained smooth deterministic NCSC minimax problems, i.e., $g \equiv 0$ and $h = \mathbb{I}_{\mathcal{Y}}$ in (1.1). To achieve an ε-accurate solution, it requires $\mathcal{O}(\kappa^2\varepsilon^{-2})$ gradient evaluations, which was later reduced to $\tilde{\mathcal{O}}(\sqrt{\kappa}\varepsilon^{-2})$ by the Minimax-PPA [21] method. When $g \not\equiv 0$ or $h \not\equiv 0$, proximal-AltGDAm [3] can achieve a complexity result of $\mathcal{O}(\kappa^{\frac{11}{6}}\varepsilon^{-2})$. In the stochastic setting, GDA utilizes large-batch sampling and can achieve a sample complexity of $\mathcal{O}(\kappa^3\varepsilon^{-4})$. SREDA [27] considers both stochastic and finite-sum structured minimax problems. Similar to one choice of our method in the single-agent setting, it adopts the SPIDER-type VR. However, different from our method, SREDA needs an inner loop to approximately solve the dual maximization problem, and thus it is a double-loop method. It achieves a sample complexity of $\tilde{\mathcal{O}}(n + \sqrt{n}\kappa^2\varepsilon^{-2})$ for the finite-sum case and $\mathcal{O}(\kappa^3\varepsilon^{-3})$ for the stochastic case. All aforementioned single-agent methods require a large batch-size: either all samples in a deterministic/finite-sum setting or as many as $\Theta(\varepsilon^{-2})$ in a stochastic setting where SPIDER-type VR is used. In contrast, Acc-MDA [12] can achieve a complexity of $\tilde{\mathcal{O}}(\kappa^{\frac{9}{2}}\varepsilon^{-3})$ by $\mathcal{O}(1)$ samples per update through the STORM-type variance reduction. SAPD+ [53] can also have convergence guarantees by small-batch sampling and achieves a complexity of $\mathcal{O}(\kappa\varepsilon^{-4})$. When big-batch sampling is performed, SAPD+ can have a complexity of $\mathcal{O}(\kappa^2\varepsilon^{-3})$ by variance reduction. However, different from Acc-MDA and our method in a single-agent setting, SAPD+ is a double-loop method. A comprehensive literature review for single-agent methods designed for solving (1.1) is out of the scope of this work. The interested readers can refer to the references therein of previously mentioned works for a more thorough treatment of this subject, including whether or not each method can handle $g \not\equiv 0$ or $h \not\equiv 0$. For readers interested in works that handle the nonconvex concave or nonconvex nonconcave settings, see e.g., [13, 18, 20, 36, 41, 50, 51, 53].

1.3 Notation

We denote $[m]$ as the set $\{1,\dots,m\}$. We let $\mathbf{z} := [\mathbf{x};\mathbf{y}] \in \mathbb{R}^{d_1+d_2}$ and $\mathcal{Z} = \mathrm{dom}(g) \times \mathrm{dom}(h)$ be the joint variable and domain. For each $i \in [m]$, let $\mathbf{z}_i := [\mathbf{x}_i;\mathbf{y}_i]$ be a local copy of $\mathbf{z}$ on agent i. Then we denote

$$\mathbf{X} = \big[\mathbf{x}_1,\dots,\mathbf{x}_m\big]^\top,\ \mathbf{Y} = \big[\mathbf{y}_1,\dots,\mathbf{y}_m\big]^\top,\ \mathbf{Z} = \big[\mathbf{z}_1,\dots,\mathbf{z}_m\big]^\top, \bar{\mathbf{x}} = \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbf{x}_i,\ \mathbf{X}_\perp = \mathbf{X} - \tfrac{1}{m}\mathbf{1}\bar{\mathbf{x}}^\top, \quad (1.4a)$$

[1] With $\theta = \Theta(1/L)$, the coefficient for $\mathbb{E}\left\|\bar{u}^t\right\|$ becomes positive but it is required to be negative in the analysis for DM-HSGD.

$$\nabla_{\mathbf{x}} F(\mathbf{Z}) = \left[\nabla_{\mathbf{x}} f_1(\mathbf{z}_1), \ldots, \nabla_{\mathbf{x}} f_m(\mathbf{z}_m)\right]^\top, \quad \nabla_{\mathbf{y}} F(\mathbf{Z}) = \left[\nabla_{\mathbf{y}} f_1(\mathbf{z}_1), \ldots, \nabla_{\mathbf{y}} f_m(\mathbf{z}_m)\right]^\top. \tag{1.4b}$$

We use the superscript $^{(t)}$ for the t-th iteration. For a set $\mathcal{X} \subseteq \mathbb{R}^d$, we denote its indicator function by $\mathbb{I}_{\mathcal{X}}(\mathbf{x})$, i.e., $\mathbb{I}_{\mathcal{X}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{X}$ and ∞ otherwise. For a closed convex function r, we define its proximal mapping as $\mathbf{prox}_r(\mathbf{x}) := \arg\min_{\mathbf{y}} \{r(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2\}$. Finally, given a set of random samples $\mathcal{B}$, we denote

$$G_i(\mathcal{B}) := \tfrac{1}{|\mathcal{B}|} \textstyle\sum_{\xi \in \mathcal{B}} \tilde{\nabla} f_i(\mathbf{x}_i, \mathbf{y}_i; \xi), \quad G_i^{(t)}(\mathcal{B}) := \tfrac{1}{|\mathcal{B}|} \textstyle\sum_{\xi \in \mathcal{B}} \tilde{\nabla} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}; \xi). \tag{1.5}$$

1.4 Outline

The rest of this paper proceeds as follows. In Sect. 2 we introduce our proposed method. We provide convergence results in Sect. 3 and numerical experiments in Sect. 4. In Sect. 5 we make concluding remarks.

2 Proposed method

The primary challenges with providing convergence guarantees for decentralized methods that solve (1.3) are caused by the non-linearity of the proximal mapping associated with the non-smooth terms g and h, the nonconvexity of $\{f_i(\cdot, \mathbf{y})\}$, and the stochasticity of gradient information.

To address these challenges, we adopt a reformulation of (1.3), introduced in the recent work [48], by using a Lagrangian multiplier $\{\boldsymbol{\lambda}_i\}_{i \in \mathcal{V}}$ to remove the consensus constraint on $\{\mathbf{y}_i\}_{i \in \mathcal{V}}$. Formally, we define

$$\Phi(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{Y}) := \tfrac{1}{m} \textstyle\sum_{i=1}^m \left(f_i(\mathbf{x}_i, \mathbf{y}_i) - h(\mathbf{y}_i)\right) - \tfrac{L}{2\sqrt{m}} \langle \boldsymbol{\Lambda}, (\mathbf{W} - \mathbf{I})\mathbf{Y} \rangle \tag{2.1}$$

such that when $\mathrm{dom}(h)$ has nonempty relative interior, the problem (1.3) can be reformulated equivalently into

$$\min_{\mathbf{X}, \boldsymbol{\Lambda}} \max_{\mathbf{Y}} \ \Phi(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{Y}) + \tfrac{1}{m} \textstyle\sum_{i=1}^m g_i(\mathbf{x}_i), \ \text{s.t. } \mathbf{x}_i = \mathbf{x}_j, \forall\, i, j \in [m]. \tag{2.2}$$

In addition, we incorporate VR techniques to facilitate a better estimate of the true local gradients. Specifically, we provide convergence guarantees when agents utilize either the SPIDER [8,34] or STORM-type [4,49] VR technique. By additionally using gradient tracking [15,24,25,32,52] in the $\mathbf{x}$-variable, we make the least restrictive assumptions on the data distribution among the agents [40], namely, heterogeneous data is allowed. Combining all of the above ideas leads to our `Variance Reduced Lagrangian Multiplier` based method for decentralized double-regularized minimax problems, `VRLM`. Its pseudocode is summarized in Algorithm 1.

The updates in (2.3) utilize the SPIDER-type VR technique, which requires large-batch sampling, meaning that the number of samples at each iteration to compute a local gradient estimator depends on a desired solution accuracy ε. As shown in the next section and in Table 1, the large batches can lead to an improved communication complexity result, but may lead to poor generalization on some machine learning tasks [14]. Furthermore, in scenarios where the data arrives in a stream it may be impractical to wait for enough samples to compute a large-batch gradient estimator. Hence, we also allow for agents to utilize the STORM VR technique in (2.4). By the STORM technique, agents only need $\mathcal{O}(1)$ samples to compute a stochastic gradient estimator, except for a possible large-batch sampling at the initial step. We find in practice (see Section 4) that the STORM estimator outperforms the SPIDER estimator on more complex tasks. Nevertheless, we will provide convergence analysis for both methods.

Algorithm 1: Variance Reduced Lagrangian Multiplier based method: VRLM (agent view)

Input: $\mathbf{x}_i^{(0)} = \mathbf{x}^{(0)} \in \mathrm{dom}(g)$, $\mathbf{y}_i^{(0)} = \mathbf{y}^{(0)} \in \mathrm{dom}(h)$, $\boldsymbol{\lambda}_i^{(0)} = \mathbf{0}$ for all $i \in [m]$; step-sizes $\eta_{\mathbf{x}}, \eta_{\mathbf{y}}, \eta_{\boldsymbol{\Lambda}} > 0$; VR-tag

Initial step: set $\mathbf{v}_i^{(0)} = \mathbf{d}_i^{(0)} = G_i^{(0)}(\mathcal{B}_i^{(0)})$ where $|\mathcal{B}_i^{(0)}| = \mathcal{S}_0$ for all $i \in [m]$

for $t = 1, \ldots$ **do**

  for agents $i \in [m]$ **in parallel do**

    if VR-tag == SPIDER **then**

$$\mathbf{d}_i^{(t)} = [\mathbf{d}_{\mathbf{x},i}^{(t)}; \mathbf{d}_{\mathbf{y},i}^{(t)}] = \begin{cases} G_i^{(t)}(\tilde{\mathcal{B}}_i^{(t)}) & \text{if } \mathrm{mod}(t,q) = 0, \text{ where } |\tilde{\mathcal{B}}_i^{(t)}| = \mathcal{S}_1 \\ G_i^{(t)}(\mathcal{B}_i^{(t)}) - G_i^{(t-1)}(\mathcal{B}_i^{(t)}) + \mathbf{d}_i^{(t-1)} & \text{otherwise, where } |\mathcal{B}_i^{(t)}| = \mathcal{S}_2. \end{cases} \tag{2.3}$$

    else

$$\mathbf{d}_i^{(t)} = [\mathbf{d}_{\mathbf{x},i}^{(t)}; \mathbf{d}_{\mathbf{y},i}^{(t)}] = G_i^{(t)}(\mathcal{B}_i^{(t)}) + (1-\beta)\left(\mathbf{d}_i^{(t-1)} - G_i^{(t-1)}(\mathcal{B}_i^{(t)})\right) \text{ where } |\mathcal{B}_i^{(t)}| = \mathcal{S}_t. \tag{2.4}$$

    Let $\mathbf{v}_{\mathbf{x},i}^{(t)} = \sum_{j=1}^m w_{ij}\left(\mathbf{v}_{\mathbf{x},j}^{(t-1)} + \mathbf{d}_{\mathbf{x},j}^{(t)} - \mathbf{d}_{\mathbf{x},j}^{(t-1)}\right)$ and $\mathbf{v}_{\mathbf{y},i}^{(t)} = \mathbf{d}_{\mathbf{y},i}^{(t)} - \frac{L\sqrt{m}}{2}\left(\sum_{j=1}^m w_{ji}\boldsymbol{\lambda}_j^{(t)} - \boldsymbol{\lambda}_i^{(t)}\right)$.

    Update the Lagrangian multiplier by $\boldsymbol{\lambda}_i^{(t+1)} = \boldsymbol{\lambda}_i^{(t)} + \frac{L\eta_{\boldsymbol{\Lambda}}}{2\sqrt{m}}\left(\sum_{j=1}^m w_{ij}\mathbf{y}_j^{(t)} - \mathbf{y}_i^{(t)}\right)$.

    Let $\mathbf{x}_i^{(t+1)} = \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\sum_{j=1}^m w_{ij}\mathbf{x}_j^{(t)} - \eta_{\mathbf{x}}\mathbf{v}_{\mathbf{x},i}^{(t)}\right)$ and $\mathbf{y}_i^{(t+1)} = \mathbf{prox}_{\eta_{\mathbf{y}} h}\left(\mathbf{y}_i^{(t)} + \eta_{\mathbf{y}}\mathbf{v}_{\mathbf{y},i}^{(t)}\right)$.

Algorithm 1 provides the updates from the viewpoint of the agents. To facilitate ease of expression and readability, we form $\mathbf{D}_{\mathbf{x}}, \mathbf{D}_{\mathbf{y}}$ and re-write the last three lines of Algorithm 1 in the following matrix form

$$\mathbf{D}_{\mathbf{x}}^{(t)} = [\mathbf{d}_{\mathbf{x},1}, \ldots, \mathbf{d}_{\mathbf{x},m}]^\top, \quad \mathbf{D}_{\mathbf{y}}^{(t)} = [\mathbf{d}_{\mathbf{y},1}, \ldots, \mathbf{d}_{\mathbf{y},m}]^\top, \tag{2.5}$$

$$\mathbf{V}_{\mathbf{x}}^{(t)} = \mathbf{W}\left(\mathbf{V}_{\mathbf{x}}^{(t-1)} + \mathbf{D}_{\mathbf{x}}^{(t)} - \mathbf{D}_{\mathbf{x}}^{(t-1)}\right), \quad \mathbf{V}_{\mathbf{y}}^{(t)} = \mathbf{D}_{\mathbf{y}}^{(t)} - \frac{L\sqrt{m}}{2}(\mathbf{W}-\mathbf{I})^\top \boldsymbol{\Lambda}^{(t)}, \tag{2.6}$$

$$\boldsymbol{\Lambda}^{(t+1)} = \boldsymbol{\Lambda}^{(t)} + \frac{L\eta_{\boldsymbol{\Lambda}}}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\mathbf{Y}^{(t)}, \tag{2.7}$$

$$\mathbf{X}^{(t+1)} = \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\widetilde{\mathbf{X}}^{(t)} - \eta_{\mathbf{x}}\mathbf{V}_{\mathbf{x}}^{(t)}\right), \quad \widetilde{\mathbf{X}}^{(t)} = \mathbf{W}\mathbf{X}^{(t)}, \quad \mathbf{Y}^{(t+1)} = \mathbf{prox}_{\eta_{\mathbf{y}} h}\left(\mathbf{Y}^{(t)} + \eta_{\mathbf{y}}\mathbf{V}_{\mathbf{y}}^{(t)}\right) \tag{2.8}$$

where the **prox** operator acts row-wisely on the input.

3 Convergence results

In this section, we give the convergence results of Algorithm 1 for two VR options (by setting the VR-tag in Algorithm 1). The following definitions are used throughout our analysis with Φ given in (2.1).

$$P(\mathbf{x}, \boldsymbol{\Lambda}) := \max_{\mathbf{Y}} \Phi(\mathbf{1}\mathbf{x}^\top, \mathbf{Y}, \boldsymbol{\Lambda}), \quad Q(\mathbf{X}, \boldsymbol{\Lambda}) := \max_{\mathbf{Y}} \Phi(\mathbf{X}, \mathbf{Y}, \boldsymbol{\Lambda}), \quad S_\Phi(\mathbf{X}, \boldsymbol{\Lambda}) := \arg\max_{\mathbf{Y}} \Phi(\mathbf{X}, \boldsymbol{\Lambda}, \mathbf{Y}), \tag{3.1}$$

$$f(\mathbf{x}, \mathbf{y}) := \tfrac{1}{m}\textstyle\sum_{i=1}^m f_i(\mathbf{x}, \mathbf{y}), \quad p(\mathbf{x}) := \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - h(\mathbf{y}), \quad \phi(\mathbf{x}, \boldsymbol{\Lambda}) := P(\mathbf{x}, \boldsymbol{\Lambda}) + g(\mathbf{x}), \tag{3.2}$$

$$\widehat{\mathbf{Y}}^{(t)} := \arg\max_{\mathbf{Y}} \Phi(\mathbf{1}\bar{\mathbf{x}}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}), \quad \widetilde{\mathbf{Y}}^{(t)} := \arg\max_{\mathbf{Y}} \Phi(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}), \tag{3.3}$$

$$\mathbf{R}_{\mathbf{x}}^{(t)} := \mathbf{D}_{\mathbf{x}}^{(t)} - \nabla_{\mathbf{x}} F(\mathbf{Z}^{(t)}), \quad \mathbf{R}_{\mathbf{y}}^{(t)} := \mathbf{D}_{\mathbf{y}}^{(t)} - \nabla_{\mathbf{y}} F(\mathbf{Z}^{(t)}), \quad \mathbf{R}^{(t)} := \mathbf{D}^{(t)} - \nabla F(\mathbf{Z}^{(t)}), \tag{3.4}$$

$$\Gamma_t(\mathbf{Y}) := \tfrac{1}{m}\textstyle\sum_{i=1}^m f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i) - \frac{L}{2\sqrt{m}}\big\langle (\mathbf{W}-\mathbf{I})\,\mathbf{Y}, \boldsymbol{\Lambda}^{(t)}\big\rangle. \tag{3.5}$$

By Danskin's theorem [5], with $\mathbf{Y} = S_\Phi(\mathbf{1}\mathbf{x}^\top, \boldsymbol{\Lambda})$, we have

$$\nabla P(\mathbf{x}, \boldsymbol{\Lambda}) = \left(\tfrac{1}{m} \textstyle\sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}, \mathbf{y}_i),\ -\tfrac{L}{2\sqrt{m}} (\mathbf{W} - \mathbf{I})\mathbf{Y} \right). \tag{3.6}$$

The analysis will be conducted based on the following assumptions, which are standard in the minimax and decentralized optimization literature [38, 54].

Assumption 3.1 *The function $f_i(\mathbf{x}, \cdot)$ is μ-strongly convex with $\mu > 0$, for each $i \in [m]$; there exists $\phi^* \in \mathbb{R}$ such that $\phi(\mathbf{x}, \boldsymbol{\Lambda}) \geq \phi^*$ for all $\mathbf{x}, \boldsymbol{\Lambda}$ where ϕ is defined in (3.2); $\mathrm{dom}(h)$ has a nonempty relative interior.*

Assumption 3.2 *The mixing matrix $\mathbf{W} \in \mathbb{R}^{m \times m}$ satisfies the conditions:* (i) $\mathrm{Null}(\mathbf{W} - \mathbf{I}) = \mathrm{Span}\{\mathbf{1}\}$ *and* $\mathbf{W}^\top \mathbf{1} = \mathbf{1}$; (ii) $\rho := \|\mathbf{W} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\|_2 < 1$; (iii) $\|\mathbf{W} - \mathbf{I}\|_2 \leq 2$.

Under Assumption 3.2 and by the notation in (2.8), since $\mathbf{W} - \mathbf{I} = (\mathbf{W} - \mathbf{I})(\mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top)$, it holds

$$\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F^2 \leq 2\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 + 2\left\|(\mathbf{W} - \mathbf{I})\,\mathbf{X}_\perp^{(t)}\right\|_F^2 \leq 2\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 + 8\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2. \tag{3.7}$$

Assumption 3.3 *For each $i \in [m]$, f_i is L-smooth. In addition, in the stochastic case, $\tilde{\nabla} f_i(\mathbf{x}, \mathbf{y}; \xi)$ satisfies* (i) $\mathbb{E}_\xi[\tilde{\nabla} f_i(\mathbf{x}, \mathbf{y}; \xi)] = \nabla f_i(\mathbf{x}, \mathbf{y})$, (ii) *there exists $\sigma^2 \geq 0$ such that $\mathbb{E}_\xi \left\|\tilde{\nabla} f_i(\mathbf{x}, \mathbf{y}; \xi) - \nabla f_i(\mathbf{x}, \mathbf{y})\right\|_2^2 \leq \sigma^2$, and* (iii) *for any $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^{d_1 + d_2}$, it holds $\mathbb{E}_\xi \left\|\tilde{\nabla} f_i(\mathbf{x}, \mathbf{y}; \xi) - \tilde{\nabla} f_i(\mathbf{x}', \mathbf{y}'; \xi)\right\|_2^2 \leq L^2 \left(\|\mathbf{x} - \mathbf{x}'\|_2^2 + \|\mathbf{y} - \mathbf{y}'\|_2^2\right)$.*

Throughout this paper, we denote the condition number by

$$\kappa := \tfrac{L}{\mu}. \tag{3.8}$$

To quantify the sample and communication complexities, we are interested in finding a near-stationary point of (2.2), defined below.

Definition 3.1 For any $\varepsilon > 0$, a random point $(\mathbf{X}, \boldsymbol{\Lambda})$ is called an ε-stationary point in expectation of (2.2) if there is some $\eta_{\mathbf{x}} > 0$ such that

$$\mathbb{E}\left\|\tfrac{1}{\eta_{\mathbf{x}}}\left(\bar{\mathbf{x}} - \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\bar{\mathbf{x}} - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} P(\bar{\mathbf{x}}, \boldsymbol{\Lambda})\right)\right)\right\|_2^2 + \tfrac{L^2}{m}\mathbb{E}\left\|\mathbf{X}_\perp\right\|_F^2 \leq \varepsilon^2 \text{ and } \mathbb{E}\left\|\nabla_{\boldsymbol{\Lambda}} P(\bar{\mathbf{x}}, \boldsymbol{\Lambda})\right\|_F^2 \leq \varepsilon^2. \tag{3.9}$$

Remark 3.1 Notice that P is the objective function of the primal problem of (2.2). The stationarity measure in Definition 3.1 is sometimes called optimization stationarity [55]. Another related notion is the so-called game stationarity, which also involves the dual variable $\mathbf{y}$. In addition, [48] shows that if $(\mathbf{X}, \boldsymbol{\Lambda})$ is an ε-stationary point of (2.2), then $\mathbf{X}$ is an $\mathcal{O}(\varepsilon)$-stationary point of the original decentralized formulation (1.3) when h is smooth. This claim can actually be extended to the case where the function $P(\mathbf{x}, \cdot)$ defined in (3.1) satisfies a quadratic-growth condition [7] for all $\mathbf{x} \in \mathrm{dom}(g)$. Hence, we adopt the notion in Definition 3.1.

3.1 Preparatory results

We begin with the following preparatory propositions. The first one is directly from [48].

Proposition 3.1 *Let P and S_Φ be defined in (3.1). Then with $L_P = L\sqrt{4\kappa^2+1}$, it holds*

$$\|\nabla P(\mathbf{x},\boldsymbol{\Lambda}) - \nabla P(\tilde{\mathbf{x}},\tilde{\boldsymbol{\Lambda}})\|_F^2 \le L_P^2\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2 + \|\boldsymbol{\Lambda}-\tilde{\boldsymbol{\Lambda}}\|_F^2\right), \forall\, \mathbf{x},\tilde{\mathbf{x}} \in \mathrm{dom}(g);\ \forall\, \boldsymbol{\Lambda},\tilde{\boldsymbol{\Lambda}}, \tag{3.10a}$$

$$\|S_\Phi(\mathbf{X},\boldsymbol{\Lambda}) - S_\Phi(\widetilde{\mathbf{X}},\boldsymbol{\Lambda})\|_F^2 \le \kappa^2\|\mathbf{X}-\widetilde{\mathbf{X}}\|_F^2, \forall\, \mathbf{x}_i,\tilde{\mathbf{x}}_i \in \mathrm{dom}(g), \forall\, i;\ \forall\, \boldsymbol{\Lambda}, \tag{3.10b}$$

$$\|S_\Phi(\mathbf{X},\boldsymbol{\Lambda}) - S_\Phi(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}})\|_F^2 \le 2\kappa^2\|\mathbf{X}-\widetilde{\mathbf{X}}\|_F^2 + 2m\kappa^2\|\boldsymbol{\Lambda}-\tilde{\boldsymbol{\Lambda}}\|_F^2, \forall\, \mathbf{x}_i,\tilde{\mathbf{x}}_i \in \mathrm{dom}(g), \forall\, i;\ \forall\, \boldsymbol{\Lambda},\tilde{\boldsymbol{\Lambda}}. \tag{3.10c}$$

Based on Proposition 3.1 and (3.3), we have

$$\left\|\widehat{\mathbf{Y}}^{(t)} - \widetilde{\mathbf{Y}}^{(t)}\right\|_F^2 \le \kappa^2\left\|\bar{\mathbf{X}}^{(t)} - \mathbf{X}^{(t)}\right\|_F^2 = \kappa^2\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2. \tag{3.11}$$

Proposition 3.2 *Let Q be defined in (3.1). Then with $L_Q = L\sqrt{4\kappa^2+1}$, it holds*

$$\begin{aligned} &m\left\|\nabla_{\mathbf{X}} Q(\mathbf{X},\boldsymbol{\Lambda}) - \nabla_{\mathbf{X}} Q(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}})\right\|_F^2 + \left\|\nabla_{\boldsymbol{\Lambda}} Q(\mathbf{X},\boldsymbol{\Lambda}) - \nabla_{\boldsymbol{\Lambda}} Q(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}})\right\|_F^2 \\ \le& L_Q^2\left(\tfrac{1}{m}\|\mathbf{X}-\widetilde{\mathbf{X}}\|_F^2 + \|\tilde{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_F^2\right), \forall\, \mathbf{x}_i,\tilde{\mathbf{x}}_i \in \mathrm{dom}(g), \forall\, i;\ \forall\, \boldsymbol{\Lambda},\tilde{\boldsymbol{\Lambda}}. \end{aligned} \tag{3.12}$$

Proof. By the notation in (1.4), let $\mathbf{Y} = S_\Phi(\mathbf{X},\boldsymbol{\Lambda})$ and $\widetilde{\mathbf{Y}} = S_\Phi(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}})$. We have

$$\nabla Q(\mathbf{X},\boldsymbol{\Lambda}) = \left(\tfrac{1}{m}\nabla F(\mathbf{Z})^\top,\ -\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\mathbf{Y}\right), \quad \nabla Q(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}}) = \left(\tfrac{1}{m}\nabla F(\tilde{\mathbf{Z}})^\top,\ -\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\widetilde{\mathbf{Y}}\right).$$

Hence, by the L-smoothness of each f_i, it follows

$$\begin{aligned} &m\left\|\nabla_{\mathbf{X}} Q(\mathbf{X},\boldsymbol{\Lambda}) - \nabla_{\mathbf{X}} Q(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}})\right\|_F^2 + \left\|\nabla_{\boldsymbol{\Lambda}} Q(\mathbf{X},\boldsymbol{\Lambda}) - \nabla_{\boldsymbol{\Lambda}} Q(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}})\right\|_F^2 \\ =& \tfrac{1}{m}\left\|\nabla F(\mathbf{Z}) - \nabla F(\tilde{\mathbf{Z}})\right\|_F^2 + \tfrac{L^2}{4m}\left\|(\mathbf{W}-\mathbf{I})(\mathbf{Y}-\widetilde{\mathbf{Y}})\right\|_F^2 \le \tfrac{L^2}{m}\left(\|\mathbf{X}-\widetilde{\mathbf{X}}\|_F^2 + \|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2\right) + \tfrac{L^2}{m}\|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2. \end{aligned}$$

Now use (3.10c) in the inequality above to obtain the desired result. □

The inequality in (3.12) indicates the smoothness of Q under a weighted norm. By [33, Eqn. 2.12], we have that for any $\mathbf{X},\widetilde{\mathbf{X}}$ with $\mathbf{x}_i,\tilde{\mathbf{x}}_i \in \mathrm{dom}(g), \forall\, i \in [m]$ and any $\boldsymbol{\Lambda},\tilde{\boldsymbol{\Lambda}}$,

$$\left|Q(\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}}) - Q(\mathbf{X},\boldsymbol{\Lambda}) - \left\langle \nabla Q(\mathbf{X},\boldsymbol{\Lambda}), (\widetilde{\mathbf{X}},\tilde{\boldsymbol{\Lambda}}) - (\mathbf{X},\boldsymbol{\Lambda})\right\rangle\right| \le \tfrac{L_Q}{2}\left(\tfrac{1}{m}\|\mathbf{X}-\widetilde{\mathbf{X}}\|_F^2 + \|\tilde{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_F^2\right) \tag{3.13}$$

The lemma below relates the stationarity violation of (2.2) to terms that appear in our analysis. This lemma will be utilized to provide our final convergence rate results. Its proof is given in Appendix A.

Lemma 3.1 *Let $\{(\mathbf{X}^{(t)},\boldsymbol{\Lambda}^{(t)},\mathbf{Y}^{(t)})\}$ be generated from Algorithm 1. For any $t \ge 0$, it holds that*

$$\begin{aligned} &\mathbb{E}\left\|\tfrac{1}{\eta_{\mathbf{x}}}\left(\bar{\mathbf{x}}^{(t)} - \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\bar{\mathbf{x}}^{(t)} - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right)\right)\right\|_2^2 + \tfrac{L^2}{m}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 \\ \le& \tfrac{5}{m\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + \left(\tfrac{2L^2(3+5\kappa^2)}{m} + \tfrac{5}{m\eta_{\mathbf{x}}^2}\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 + \tfrac{10L^2}{m}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2 \\ &+ \tfrac{5}{m}\left\|\mathbf{R}^{(t)}\right\|_F^2 + \tfrac{5}{m}\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2, \end{aligned} \tag{3.14}$$

$$\mathbb{E}\left\|\nabla_{\boldsymbol{\Lambda}} P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right\|_F^2 \le \tfrac{2}{\eta_{\boldsymbol{\Lambda}}^2}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2 + \tfrac{4L^2\kappa^2}{m}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 + \tfrac{4L^2}{m}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2, \tag{3.15}$$

where P and $\mathbf{R}^{(t)}$ are defined in (3.2) and (3.4).

3.2 One-iteration progress inequality

Our analysis relies on establishing a one-iteration progress inequality about ϕ based on the updates in lines 8-10 in Algorithm 1. Its proof is deferred to Appendix B.

Lemma 3.2 *Let $\{(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)})\}$ be generated from Algorithm 1. For all $t \geq 0$, it holds that*

$$\begin{aligned}
&\phi(\bar{\mathbf{x}}^{(t+1)}, \boldsymbol{\Lambda}^{(t+1)}) - \phi(\bar{\mathbf{x}}^{(t)}, \boldsymbol{\Lambda}^{(t)}) \\
\leq& -\tfrac{1}{2m}\left(\tfrac{1}{\eta_{\mathbf{x}}} - L(\kappa+1+c_1) - L_P(1+c_3)\right)\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 \\
&-\left(\tfrac{1}{\eta_{\boldsymbol{\Lambda}}} - \tfrac{L_P}{2} - Lc_2\right)\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2 - \tfrac{1}{2m\eta_{\mathbf{x}}}\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 \\
&+\tfrac{1}{2m}\left(L(\kappa+1) + \tfrac{\kappa^2}{c_2} + \tfrac{\rho^2}{\eta_{\mathbf{x}}}\right)\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 + \tfrac{1}{2m}\left(\tfrac{L}{c_1} + \tfrac{1}{c_2}\right)\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2 \\
&+\tfrac{1}{2mL_Pc_3}\textstyle\sum_{i=1}^m\left\|\tfrac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \mathbf{v}_{\mathbf{x},i}^{(t)}\right\|_2^2,
\end{aligned} \tag{3.16}$$

where $c_1, c_2, c_3 > 0$ are arbitrary constants, and $\phi, \widetilde{\mathbf{X}}^{(t)}$ and $\widetilde{\mathbf{Y}}^{(t)}$ are defined in (3.2), (2.8), and (3.3).

From Lemma 3.2, we see that the change in $\phi(\bar{\mathbf{x}}, \boldsymbol{\Lambda})$ is increasing in

$$\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2, \quad \left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2, \quad \textstyle\sum_{i=1}^m\left\|\tfrac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \mathbf{v}_{\mathbf{x},i}^{(t)}\right\|_2^2. \tag{3.17}$$

Hence we need to ensure these terms can be well controlled to establish convergence. The next subsection is devoted to providing upper bounds on each term in (3.17).

3.3 Consensus and dual error bounds.

The following proof can be found in Lemma C.7 of [28].

Lemma 3.3 *Let $\{(\mathbf{X}^{(t)}, \mathbf{V}^{(t)})\}$ be generated from Algorithm 1. For all $t \geq 0$, it holds that*

$$\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2 \leq \rho\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 + \tfrac{\eta_{\mathbf{x}}^2}{1-\rho}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2, \tag{3.18}$$

where ρ is defined in Assumption 3.2.

Next, we provide an upper bound on the last term in (3.17). Its proof is given in Appendix C.

Lemma 3.4 *Let $\{(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{V}^{(t)})\}$ be generated from Algorithm 1 and $\mathbf{R}^{(t)}$ defined in (3.4). Then*

$$\textstyle\sum_{i=1}^m \mathbb{E}\left\|\tfrac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \mathbf{v}_{\mathbf{x},i}^{(t)}\right\|_2^2 \leq 2\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2 + 2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2, \forall\, t \geq 0. \tag{3.19}$$

Finally, we provide upper bounds to the dual errors. The proofs are given in Appendix C.

Lemma 3.5 *Let $\{(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)})\}$ be generated from Algorithm 1. Then provided $\eta_{\mathbf{y}} \le \frac{1}{4L}$, it holds that for any $c_4, c_5 > 0$,*

$$\begin{aligned}\left\|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\right\|_F^2 \le\ & 4m\eta_{\mathbf{y}}\left(\hat{\delta}_t - \hat{\delta}_{t+1}\right) + 4\eta_{\mathbf{y}}^2\left\|\mathbf{R}^{(t)}\right\|_F^2 + 4\eta_{\mathbf{y}}\left(\tfrac{L^2}{2c_4} + \tfrac{L\sqrt{m}}{c_5}\right)\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \mathbf{Y}^{(t+1)}\right\|_F^2 \\ & + 4\eta_{\mathbf{y}}\left(\tfrac{L_Q+L}{2} + \tfrac{c_4}{2}\right)\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F^2 + 4\eta_{\mathbf{y}}\left(\tfrac{mL_Q}{2} + \tfrac{c_5 L\sqrt{m}}{4}\right)\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2,\end{aligned} \tag{3.20}$$

where $\widetilde{\mathbf{Y}}^{(t)}$ is defined in (3.3), $\mathbf{R}^{(t)}$ is defined in (3.4), $L_Q = L\sqrt{4\kappa^2+1}$, and

$$\hat{\delta}_t := Q(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}) - \left(\Gamma_t(\mathbf{Y}^{(t)}) - \tfrac{1}{m}\textstyle\sum_{i=1}^m h(\mathbf{y}_i^{(t)})\right), \tag{3.21}$$

with $Q(\mathbf{X}, \boldsymbol{\Lambda})$ and $\Gamma_t(\cdot)$ defined in (3.1) and (3.5).

Lemma 3.6 *Let $\{(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)})\}$ be generated from Algorithm 1 and $\mathbf{R}^{(t)}$ defined in (3.4). Suppose $\eta_{\mathbf{y}} \le \frac{1}{4L}$. Then it holds that*

$$\begin{aligned}&\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \mathbf{Y}^{(t+1)}\right\|_F^2 \\ &\le (1-\eta_{\mathbf{y}}\mu)\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2 + \tfrac{4\eta_{\mathbf{y}}}{\mu}\left\|\mathbf{R}^{(t)}\right\|_F^2 + \tfrac{4\kappa^2}{\eta_{\mathbf{y}}\mu}\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F^2 + \tfrac{4\kappa^2 m}{\eta_{\mathbf{y}}\mu}\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned} \tag{3.22}$$

3.4 Convergence results by SPIDER-type variance reduction

In this subsection, we set VR-tag = SPIDER in Algorithm 1, and we consider both the general stochastic case and the special finite-sum setting. The proofs of all the lemmas are given in Appendix D. We first bound the consensus error of the tracked gradient and the error of the gradient estimator.

Lemma 3.7 *Let $\{(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{V}^{(t)})\}$ be generated from Algorithm 1 and $\mathbf{R}^{(t)}$ defined in (3.4). When (1.2) holds, we set $\mathcal{S}_0 = \mathcal{S}_1 = n$ and take all data samples. Define $n_t \in \mathbb{Z}_+$ as the unique integer such that $n_t q \le t < (n_t+1)q$ for all $t \ge 0$. Then it holds that*

$$\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2 \le \rho\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2 + \frac{6m\Upsilon}{1-\rho} + \frac{3L^2}{1-\rho}\mathbb{E}\left(\left\|\mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)}\right\|_F^2 + \frac{2}{\mathcal{S}_2}\sum_{r=n_t q}^{t}\left\|\mathbf{Z}^{(r+1)} - \mathbf{Z}^{(r)}\right\|_F^2\right), \tag{3.23}$$

$$\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2 \le \frac{L^2}{\mathcal{S}_2}\sum_{r=n_t q}^{t-1}\mathbb{E}\left\|\mathbf{Z}^{(r+1)} - \mathbf{Z}^{(r)}\right\|_F^2 + m\Upsilon, \tag{3.24}$$

where $\mathbf{Z}^{(t)} = (\mathbf{X}^{(t)}, \mathbf{Y}^{(t)})$ by the notation in (1.4a), and

$$\Upsilon := \tfrac{\sigma^2}{\mathcal{S}_1}, \text{ for general distributions } \{\mathcal{D}_i\};\ \Upsilon = 0, \text{ for the special finite-sum setting in (1.2).} \tag{3.25}$$

Remark 3.2 Notice that for any $t \ge 0$, we have $n_t q \le t \le (n_t+1)q - 1$. Therefore

$$\sum_{t=0}^{T-1}\sum_{r=n_t q}^{t}\mathbb{E}\left\|\mathbf{Z}^{(r+1)} - \mathbf{Z}^{(r)}\right\|_F^2 \le q\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)}\right\|_F^2. \tag{3.26}$$

The relation in (3.26) is standard in the analysis of SPIDER-type methods; e.g., see [46, Eqn. (85)].

In the rest of this subsection, we set

$$q = \mathcal{S}_2,\ c_1 = c_2 = 32\kappa^2,\ c_3 = 60,\ c_4 = 16\kappa^2 L,\ c_5 = 32\sqrt{m}\kappa^2,\ \eta_{\mathbf{y}} = \tfrac{1}{4L}. \tag{3.27}$$

Lemma 3.8 *Let $\{(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)})\}$ be generated from Algorithm 1, $\widetilde{\mathbf{Y}}^{(t)}$ defined in (3.3), and $\hat{\delta}_t$ be defined in (3.21). Then for any integer $T \geq 1$, it holds that*

$$\begin{aligned}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \mathbf{Y}^{(t+1)}\right\|_F^2 \leq\ & 16\kappa^2(20\kappa^2+\kappa+2)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left(\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 + 4\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2\right) \\ &+ (6\kappa - \tfrac{3}{2})\left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2 + \tfrac{8m\kappa^2}{L}\hat{\delta}_0 + \tfrac{8mT\Upsilon}{\mu^2} + 8m\kappa^2(20\kappa^2+\kappa+1)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2\end{aligned} \tag{3.28}$$

and

$$\begin{aligned}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\right\|_F^2 \leq\ & 2(24\kappa^2+2\kappa+3)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left(\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 + 4\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2\right) \\ &+ \tfrac{1}{2\kappa}\left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2 + \tfrac{2m}{L}\hat{\delta}_0 + \tfrac{mT\Upsilon}{L^2} + 2m(12\kappa^2+\kappa+1)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned} \tag{3.29}$$

Below we show the square summability of the generated sequence, which is crucial to obtain our convergence and complexity results.

Theorem 3.1 *Under Assumptions 3.1-3.3, let $\{(\mathbf{X}^{(t)}, \widetilde{\mathbf{X}}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{V}^{(t)})\}_{t\geq 0}$ be generated from Algorithm 1 with `VR-tag` == SPIDER, $q = \mathcal{S}_2$, $\eta_{\mathbf{y}} = \frac{1}{4L}$, and $\eta_{\mathbf{x}}$ and $\eta_{\boldsymbol{\Lambda}}$ set to*

$$\eta_{\mathbf{x}} = \min\left\{\frac{(1-\rho)^2}{180L_P},\ \frac{1}{20(L+1)(12\kappa^2+2\kappa+5)}\right\}, \tag{3.30a}$$

$$\eta_{\boldsymbol{\Lambda}} = \min\left\{\frac{5L_P(1-\rho)^2}{24L^2(12\kappa^2+\kappa+1)},\ \frac{1}{2L_P + 128L\kappa^2 + \frac{(L+1)(20\kappa^2+\kappa+1)}{2} + \frac{4L^2(12\kappa^2+\kappa+1)}{30L_P}}\right\}, \tag{3.30b}$$

where $L_P = L\sqrt{4\kappa^2+1}$ is given in Proposition 3.1. Then it holds for any $T \geq 1$,

$$\begin{aligned}&\tfrac{1}{4m\eta_{\mathbf{x}}}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + \tfrac{1}{2\eta_{\boldsymbol{\Lambda}}}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2 + \tfrac{1}{4m\eta_{\mathbf{x}}}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 \\ &+ \tfrac{3}{4m\eta_{\mathbf{x}}}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 + \tfrac{1}{60mL_P}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2 \\ \leq\ & C_0 + T\left(\tfrac{1}{30L_P} + \tfrac{1}{L_P(1-\rho)^2} + \tfrac{L+1}{8L^2}\right)\Upsilon,\end{aligned} \tag{3.31}$$

where Υ is defined in (3.25), and

$$\begin{aligned}C_0 :=\ & \phi(\bar{\mathbf{x}}^{(0)}, \boldsymbol{\Lambda}^{(0)}) - \phi^* + \left(\tfrac{1}{20mL_P} + \tfrac{\rho}{15mL_P(1-\rho)}\right)\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(0)}\right\|^2 \\ &+ \left(\tfrac{6(L+1)}{64m\kappa} + \tfrac{L^2}{120mL_P\kappa} + \tfrac{3L^2}{10mL_P\kappa(1-\rho)^2}\right)\left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2 + \left(\tfrac{1+1/L}{8} + \tfrac{L}{30L_P} + \tfrac{6L}{5L_P(1-\rho)^2}\right)\hat{\delta}_0.\end{aligned}$$

Proof. We first take the expectation of (3.16), apply (3.19), and plug in the values of c_1, c_2 and c_3 to have

$$\begin{aligned}\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\phi(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right] &\le -\tfrac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2-\tfrac{1}{2m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2\\ &-\left(\tfrac{1}{\eta_{\boldsymbol{\Lambda}}}-\tfrac{L_P}{2}-32L\kappa^2\right)\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2+\tfrac{1}{2m}\left(L(\kappa+1)+\tfrac{1}{32}+\tfrac{\rho^2}{\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2\\ &+\tfrac{L+1}{64m\kappa^2}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2+\tfrac{1}{60mL_P}\mathbb{E}\left(\left\|\mathbf{R}^{(t)}\right\|_F^2+\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\right),\end{aligned}\tag{3.32}$$

where the coefficient of $\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2$ is obtained by the arguments

$$L(\kappa+1+c_1)+L_P(1+c_3)=L(\kappa+1+32\kappa^2)+61L\sqrt{4\kappa^2+1}\le\tfrac{L(\kappa+1)+1+20(L+1)(5+\kappa+12\kappa^2)}{2}\le\tfrac{1}{2\eta_{\mathbf{x}}}.$$

Next, for any $\gamma_1>0$ and $\gamma_2>0$, we add $\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2,\gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2$ to both sides of (3.32) and use the results of Lemmas 3.3 and 3.7 to obtain

$$\begin{aligned}&\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\phi(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right]+\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2+\gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2\\ &\le-\tfrac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2-\tfrac{1}{2m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2-\left(\tfrac{1}{\eta_{\boldsymbol{\Lambda}}}-\tfrac{L_P}{2}-32L\kappa^2\right)\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\\ &+\tfrac{1}{2m}\left(L(\kappa+1)+\tfrac{1}{32}+\tfrac{\rho^2}{\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\tfrac{L+1}{64m\kappa^2}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\\ &+\tfrac{1}{60mL_P}\mathbb{E}\left(\tfrac{L^2}{\mathcal{S}_2}\textstyle\sum_{r=n_tq}^{t-1}\mathbb{E}\left\|\mathbf{Z}^{(r+1)}-\mathbf{Z}^{(r)}\right\|_F^2+m\Upsilon+\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\right)+\gamma_1\rho\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\tfrac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\\ &+\gamma_2\left(\rho\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2+\tfrac{6m\Upsilon}{1-\rho}+\tfrac{3L^2}{1-\rho}\mathbb{E}\left[\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2+\tfrac{2}{\mathcal{S}_2}\textstyle\sum_{r=n_tq}^{t}\left\|\mathbf{Z}^{(r+1)}-\mathbf{Z}^{(r)}\right\|_F^2\right]\right).\end{aligned}\tag{3.33}$$

Sum up (3.33) over $t=0$ to $T-1$, utilize (3.26), and recall $\mathcal{S}_2=q$ to have

$$\begin{aligned}&\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(T)},\boldsymbol{\Lambda}^{(T)})-\phi(\bar{\mathbf{x}}^{(0)},\boldsymbol{\Lambda}^{(0)})\right]+\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left(\gamma_1\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2+\gamma_2\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2\right)\\ &\le\textstyle\sum_{t=0}^{T-1}\mathbb{E}\Big[-\tfrac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2-\tfrac{1}{2m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2\\ &\qquad-\left(\tfrac{1}{\eta_{\boldsymbol{\Lambda}}}-\tfrac{L_P}{2}-32L\kappa^2\right)\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2+\tfrac{L+1}{64m\kappa^2}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\Big]\\ &+\tfrac{1}{2m}\left(L(\kappa+1)+\tfrac{1}{32}+\tfrac{\rho^2}{\eta_{\mathbf{x}}}+2m\gamma_1\rho\right)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\left(\tfrac{L^2}{60mL_P}+\tfrac{9\gamma_2L^2}{1-\rho}\right)\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2\\ &+\left(\tfrac{1}{60mL_P}+\tfrac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}+\gamma_2\rho\right)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2+T\left(\tfrac{1}{60L_P}+\tfrac{6m\gamma_2}{1-\rho}\right)\Upsilon.\end{aligned}\tag{3.34}$$

Now plug (3.28) and (3.29) into (3.34) and also bound $\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2$ by (3.7) to have

$$\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(T)},\boldsymbol{\Lambda}^{(T)})-\phi(\bar{\mathbf{x}}^{(0)},\boldsymbol{\Lambda}^{(0)})\right]+\sum_{t=0}^{T-1}\mathbb{E}\left(\gamma_1\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2+\gamma_2\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2+\tfrac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2\right)$$

$$\leq -\sum_{t=0}^{T-1}\mathbb{E}\Bigg[\tfrac{1}{2m}\left(\tfrac{1}{\eta_{\mathbf{x}}}-\tfrac{(L+1)(20\kappa^2+\kappa+2)}{2}-2(24\kappa^2+2\kappa+4)\left(\tfrac{L^2}{30L_P}+\tfrac{18m\gamma_2L^2}{1-\rho}\right)\right)\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2 \tag{3.35}$$

$$+\left(\tfrac{1}{\eta_{\boldsymbol{\Lambda}}}-\tfrac{L_P}{2}-32L\kappa^2-\tfrac{(L+1)(20\kappa^2+\kappa+1)}{8}-(12\kappa^2+\kappa+1)\left(\tfrac{L^2}{30L_P}+\tfrac{18m\gamma_2L^2}{1-\rho}\right)\right)\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\Bigg]$$

$$+\tfrac{1}{2m}\Bigg(L(\kappa+1)+\tfrac{1}{32}+\tfrac{\rho^2}{\eta_{\mathbf{x}}}+2m\gamma_1\rho+2(L+1)(20\kappa^2+\kappa+2)$$

$$+8(24\kappa^2+2\kappa+4)\left(\tfrac{L^2}{30L_P}+\tfrac{18m\gamma_2L^2}{1-\rho}\right)\Bigg)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2$$

$$+\left(\tfrac{1}{60mL_P}+\tfrac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}+\gamma_2\rho\right)\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2+T\left(\tfrac{1}{30L_P}+\tfrac{15m\gamma_2}{1-\rho}+\tfrac{L+1}{8L^2}\right)\Upsilon.$$

$$+\left(\tfrac{6(L+1)}{64m\kappa}+\tfrac{L^2}{120mL_P\kappa}+\tfrac{9\gamma_2L^2}{2\kappa(1-\rho)}\right)\left\|\widetilde{\mathbf{Y}}^{(0)}-\mathbf{Y}^{(0)}\right\|_F^2+\left(\tfrac{L+1}{8m}+\tfrac{L^2}{30mL_P}+\tfrac{18\gamma_2L^2}{1-\rho}\right)\tfrac{m}{L}\hat{\delta}_0.$$

Set γ_1 and γ_2 to

$$\gamma_1=\tfrac{3}{2m(1-\rho)}\left(L(\kappa+1)+\tfrac{1}{32}+\tfrac{1}{\eta_{\mathbf{x}}}+2(L+1)(20\kappa^2+\kappa+2)+8(24\kappa^2+2\kappa+4)\left(\tfrac{L^2}{30L_P}+\tfrac{3L^2}{5L_P(1-\rho)^2}\right)\right), \tag{3.36}$$

$$\gamma_2=\tfrac{2}{1-\rho}\left(\tfrac{1}{60mL_P}+\tfrac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}\right). \tag{3.37}$$

By $\eta_{\mathbf{x}}\leq\frac{(1-\rho)^2}{180L_P}$ and $L_P=L\sqrt{4\kappa^2+1}$, it is straightforward to have $\frac{144\eta_{\mathbf{x}}^2L^2(24\kappa^2+2\kappa+4)}{(1-\rho)^3}\leq\frac{1-\rho}{6}$. Then we have from (3.36) and (3.37) that

$$\begin{aligned}&\tfrac{1}{2m}\left(L(\kappa+1)+\tfrac{1}{32}+\tfrac{\rho^2}{\eta_{\mathbf{x}}}+2(L+1)(20\kappa^2+\kappa+2)+8(24\kappa^2+2\kappa+4)\left(\tfrac{L^2}{30L_P}+\tfrac{18m\gamma_2L^2}{1-\rho}\right)\right)\\ \leq&\tfrac{1}{2m}\Bigg(L(\kappa+1)+\tfrac{1}{32}+\tfrac{1}{\eta_{\mathbf{x}}}+2(L+1)(20\kappa^2+\kappa+2)+8(24\kappa^2+2\kappa+4)\left(\tfrac{L^2}{30L_P}+\tfrac{3L^2}{10L_P(1-\rho)^2}\right)\\ &+8(24\kappa^2+2\kappa+4)\tfrac{36m\gamma_1\eta_x^2L^2}{(1-\rho)^3}\Bigg)\leq\tfrac{(1-\rho)\gamma_1}{3}+\tfrac{(1-\rho)\gamma_1}{6}=\tfrac{(1-\rho)\gamma_1}{2}.\end{aligned} \tag{3.38}$$

In addition, by the choice of γ_2 in (3.37), it follows

$$\gamma_2-\left(\tfrac{1}{60mL_P}+\tfrac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}+\gamma_2\rho\right)\geq\tfrac{(1-\rho)\gamma_2}{2}. \tag{3.39}$$

By the choice of γ_1 and $\eta_{\mathbf{x}}\leq\frac{(1-\rho)^2}{180L_P}$, it follows that

$$\tfrac{m\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}\leq\tfrac{\eta_{\mathbf{x}}}{120L_P}\left(L(\kappa+1)+\tfrac{1}{32}+\tfrac{1}{\eta_{\mathbf{x}}}+2(L+1)(20\kappa^2+\kappa+2)+8(24\kappa^2+2\kappa+4)\left(\tfrac{L^2}{30L_P}+\tfrac{3L^2}{5L_P(1-\rho)^2}\right)\right). \tag{3.40}$$

Also, $\frac{\eta_{\mathbf{x}}\cdot8(24\kappa^2+2\kappa+4)\cdot3L^2}{5L_P(1-\rho)^2}\leq\frac{8(24\kappa^2+2\kappa+4)\cdot3L^2}{5\cdot180L_P^2}=\frac{8(24\kappa^2+2\kappa+4)\cdot3L^2}{5\cdot180L^2(4\kappa^2+1)}\leq\frac{1}{5}$, and by $L_P\geq2L\kappa$, it holds

$$\begin{aligned}&L(\kappa+1)+\tfrac{1}{32}+2(L+1)(20\kappa^2+\kappa+2)+\tfrac{8(24\kappa^2+2\kappa+4)L^2}{30L_P}\\ \leq\ &L(\kappa+1)+1+2(L+1)(20\kappa^2+\kappa+2)+\tfrac{8(24\kappa^2+2\kappa+4)L^2}{60L\kappa}\leq16(L+1)(12\kappa^2+2\kappa+5).\end{aligned}$$

Hence from (3.40), we have $\frac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho} \le \frac{1}{60mL_P}$. Thus $m\gamma_2 \le \frac{1}{15L_P(1-\rho)}$ follows from (3.37), and we have

$$\begin{aligned}&\tfrac{1}{\eta_{\mathbf{x}}} - \tfrac{(L+1)(20\kappa^2+\kappa+2)}{2} - 2(24\kappa^2+2\kappa+4)\left(\tfrac{L^2}{30L_P} + \tfrac{18m\gamma_2L^2}{1-\rho}\right)\\ \ge&\tfrac{1}{\eta_{\mathbf{x}}} - \tfrac{(L+1)(20\kappa^2+\kappa+2)}{2} - \tfrac{2L^2(24\kappa^2+2\kappa+4)}{30L_P}\left(1+\tfrac{36}{(1-\rho)^2}\right) \ge \tfrac{1}{2\eta_{\mathbf{x}}},\end{aligned} \tag{3.41}$$

where the last inequality can be verified by plugging the value of $\eta_{\mathbf{x}}$ given in (3.30a). Similarly, by the choice of $\eta_{\boldsymbol{\Lambda}}$ in (3.30b), it is straightforward to have

$$\tfrac{1}{\eta_{\boldsymbol{\Lambda}}} - \tfrac{L_P}{2} - 32L\kappa^2 - \tfrac{(L+1)(20\kappa^2+\kappa+1)}{8} - (12\kappa^2+\kappa+1)\left(\tfrac{L^2}{30L_P} + \tfrac{18m\gamma_2L^2}{1-\rho}\right) \ge \tfrac{1}{2\eta_{\boldsymbol{\Lambda}}}. \tag{3.42}$$

Moreover, by $m\gamma_2 \le \frac{1}{15L_P(1-\rho)}$, it holds

$$\left(\tfrac{L+1}{8m} + \tfrac{L^2}{30mL_P} + \tfrac{18\gamma_2L^2}{1-\rho}\right)\tfrac{m}{L}\hat{\delta}_0 \le \left(\tfrac{1+1/L}{8} + \tfrac{L}{30L_P} + \tfrac{6L}{5L_P(1-\rho)^2}\right)\hat{\delta}_0. \tag{3.43}$$

Therefore, we obtain (3.31) from (3.35) by using (3.38)-(3.43), the lower bounds $\frac{(1-\rho)\gamma_1}{2} \ge \frac{3}{4m\eta_{\mathbf{x}}}$ and $\frac{(1-\rho)\gamma_2}{2} \ge \frac{1}{60mL_P}$, the upper bounds $\frac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho} \le \frac{1}{60mL_P}$ and $\gamma_2 \le \frac{1}{15mL_P(1-\rho)}$, and $\phi(\bar{\mathbf{x}}^{(T)}, \boldsymbol{\Lambda}^{(T)}) \ge \phi^*$. □

By Theorem 3.1, we first show a last-iterate convergence in probability for the finite-sum case.

Theorem 3.2 (Convergence in probability for finite-sum case) *Under Assumptions 3.1-3.3, let* $\{(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)})\}_{t\ge0}$ *be generated from Algorithm 1 with* `VR-tag` *= SPIDER and* $\eta_{\mathbf{x}}, \eta_{\boldsymbol{\Lambda}}, \eta_{\mathbf{y}}$ *chosen as in (3.30) and (3.27). If (1.2) holds and* $\mathcal{S}_0 = \mathcal{S}_1 = n$ *in Algorithm 1, then for any* $\varepsilon > 0$,

$$\lim_{t\to\infty} \text{Prob}\left\{\left\|\tfrac{1}{\eta_{\mathbf{x}}}\left(\bar{\mathbf{x}}^{(t)} - \mathbf{prox}_{\eta_{\mathbf{x}}g}\left(\bar{\mathbf{x}}^{(t)} - \eta_{\mathbf{x}}\nabla_{\mathbf{x}}P(\bar{\mathbf{x}}^{(t)}, \boldsymbol{\Lambda}^{(t)})\right)\right)\right\|_2^2 + \tfrac{L^2}{m}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 \ge \varepsilon\right\} = 0, \tag{3.44a}$$

$$\lim_{t\to\infty} \text{Prob}\left\{\left\|\nabla_{\boldsymbol{\Lambda}}P(\bar{\mathbf{x}}^{(t)}, \boldsymbol{\Lambda}^{(t)})\right\|_F^2 \ge \varepsilon\right\} = 0. \tag{3.44b}$$

Proof. Recall that $\Upsilon = 0$ when (1.2) holds and $\mathcal{S}_0 = \mathcal{S}_1 = n$. Hence, Theorem 3.1 indicates

$$\textstyle\sum_{t=0}^{\infty}\mathbb{E}\left[\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + \left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2 + \left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 + \left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 + \left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\right] < \infty,$$

which together with (3.28) further implies $\sum_{t=0}^{\infty}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \mathbf{Y}^{(t+1)}\right\|_F^2 < \infty$. Therefore, each of the terms $\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2$, $\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2$, $\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2$, $\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2$, $\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2$, and $\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \mathbf{Y}^{(t+1)}\right\|_F^2$ approaches 0 in expectation as $t \to \infty$. Now the desired results follow immediately from Lemma 3.1 and the Markov inequality. □

Remark 3.3 In order to show the last-iterate convergence in probability for the general stochastic case, we need to increase $\mathcal{S}_1$ periodically such that the cumulated variance is finite. This requires us to make $\mathcal{S}_1$ dependent on the number of periods, which will cause confusion on the notation. We do not extend

the analysis here. In addition, Theorem 3.2 together with Remark 3.1 implies that $\{\mathbf{X}^{(t)}\}_{t\geq 0}$ satisfies the convergence in probability for (1.3) when (1.2) holds and $\mathcal{S}_0 = \mathcal{S}_1 = n$ in Algorithm 1, i.e.,

$$\lim_{t\to\infty} \text{Prob}\left\{ \left\| \tfrac{1}{\eta_{\mathbf{x}}} \left(\bar{\mathbf{x}}^{(t)} - \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\bar{\mathbf{x}}^{(t)} - \eta_{\mathbf{x}} \nabla p(\bar{\mathbf{x}}^{(t)}) \right) \right) \right\|_2^2 + \tfrac{L^2}{m} \left\| \mathbf{X}_{\perp}^{(t)} \right\|_F^2 \geq \varepsilon \right\} = 0,$$

where $p(\cdot)$ is defined in (3.2).

Moreover, by Theorem 3.1, we have the expected convergence rate result below.

Theorem 3.3 *Under Assumptions 3.1-3.3, let $\{(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)})\}_{t\geq 0}$ be generated from Algorithm 1 with* VR*-tag = SPIDER and $\eta_{\mathbf{x}}, \eta_{\boldsymbol{\Lambda}}, \eta_{\mathbf{y}}$ chosen as in (3.30) and (3.27). For any integer $T \geq 1$, select τ uniformly at random from $\{0, 1, \ldots, T-1\}$. Then*

$$\begin{aligned}
&\mathbb{E}\left\| \tfrac{1}{\eta_{\mathbf{x}}} \left(\bar{\mathbf{x}}^{(\tau)} - \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\bar{\mathbf{x}}^{(\tau)} - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} P(\bar{\mathbf{x}}^{(\tau)}, \boldsymbol{\Lambda}^{(\tau)}) \right) \right) \right\|_2^2 + \tfrac{L^2}{m} \mathbb{E}\left\| \mathbf{X}_{\perp}^{(\tau)} \right\|_F^2 \\
\leq & \tfrac{L^2(60\kappa+3)}{mT} \left\| \widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)} \right\|_F^2 + (80\kappa^2+10)\Upsilon + \tfrac{(80\kappa^2+10)L\hat{\delta}_0}{T} \\
& + \left(\tfrac{C_0}{T} + \left(\tfrac{1}{30L_P} + \tfrac{1}{L_P(1-\rho)^2} + \tfrac{L+1}{8L^2} \right) \Upsilon \right) \cdot \Big[10L^2 \Big(150\eta_{\mathbf{x}} \kappa^2 (20\kappa^2 + \kappa + 4) \\
& \qquad + 16\eta_{\boldsymbol{\Lambda}} \kappa^2 (20\kappa^2 + \kappa + 3) \Big) + \tfrac{20}{\eta_{\mathbf{x}}} + \left(3L^2 \eta_{\mathbf{x}} (3 + 5\kappa^2) + \tfrac{7}{\eta_{\mathbf{x}}} \right) + 300 L_P \Big]
\end{aligned} \tag{3.45}$$

and

$$\begin{aligned}
\mathbb{E}\left\| \nabla_{\boldsymbol{\Lambda}} P(\bar{\mathbf{x}}^{(\tau)}, \boldsymbol{\Lambda}^{(\tau)}) \right\|_F^2 \leq & \tfrac{24\kappa L^2}{mT} \left\| \widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)} \right\|_F^2 + 32\kappa^2 \Upsilon + \tfrac{32L\kappa^2\hat{\delta}_0}{T} \\
& + \left(\tfrac{C_0}{T} + \left(\tfrac{1}{30L_P} + \tfrac{1}{L_P(1-\rho)^2} + \tfrac{L+1}{8L^2} \right) \Upsilon \right) \cdot \Big[4L^2 \Big(150\eta_{\mathbf{x}} \kappa^2 (20\kappa^2 + \kappa + 2) \\
& \qquad + 16\eta_{\boldsymbol{\Lambda}} \kappa^2 (20\kappa^2 + \kappa + 1) \Big) + \tfrac{4}{\eta_{\boldsymbol{\Lambda}}} + 6L^2\kappa^2\eta_{\mathbf{x}} \Big],
\end{aligned} \tag{3.46}$$

where Υ is given in (3.25), $L_P = L\sqrt{4\kappa^2+1}$, and C_0 is defined in Theorem 3.1.

Proof. By the selection of τ, we have $\mathbb{E}\left\| \widetilde{\mathbf{Y}}^{(\tau)} - \mathbf{Y}^{(\tau)} \right\|_F^2 = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\| \widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)} \right\|_F^2$. Hence, it follows from (3.28) and (3.31) that

$$\begin{aligned}
&\mathbb{E}\left\| \widetilde{\mathbf{Y}}^{(\tau)} - \mathbf{Y}^{(\tau)} \right\|_F^2 \leq \tfrac{6\kappa}{T} \left\| \widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)} \right\|_F^2 + \tfrac{8m\Upsilon}{\mu^2} + \tfrac{8m\kappa^2\hat{\delta}_0}{TL} \\
& + \left(\tfrac{C_0}{T} + \left(\tfrac{1}{30L_P} + \tfrac{1}{L_P(1-\rho)^2} + \tfrac{L+1}{8L^2} \right) \Upsilon \right) \cdot \left(150m\eta_{\mathbf{x}} \kappa^2 (20\kappa^2 + \kappa + 2) + 16m\eta_{\boldsymbol{\Lambda}} \kappa^2 (20\kappa^2 + \kappa + 1) \right).
\end{aligned} \tag{3.47}$$

Similarly, we have from (3.29) and (3.31) that

$$\begin{aligned}
&\mathbb{E}\left\| \mathbf{Y}^{(\tau+1)} - \mathbf{Y}^{(\tau)} \right\|_F^2 = \tfrac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\| \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)} \right\|_F^2 \leq \tfrac{2m}{TL}\hat{\delta}_0 + \tfrac{1}{2\kappa T} \left\| \widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)} \right\|_F^2 + \tfrac{m\Upsilon}{L^2} \\
& + \left(\tfrac{C_0}{T} + \left(\tfrac{1}{30L_P} + \tfrac{1}{L_P(1-\rho)^2} + \tfrac{L+1}{8L^2} \right) \Upsilon \right) \cdot \left(20m\eta_{\mathbf{x}} (24\kappa^2 + 2\kappa + 3) + 4m\eta_{\boldsymbol{\Lambda}} (12\kappa^2 + \kappa + 1) \right).
\end{aligned} \tag{3.48}$$

In addition, summing up (3.24) over $t = 0, \ldots, T-1$ and using (3.26) gives

$$\sum_{t=0}^{T-1} \mathbb{E} \left\|\mathbf{R}^{(t)}\right\|_F^2 \leq L^2 \sum_{t=0}^{T-1} \left(\mathbb{E} \left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F^2 + \mathbb{E} \left\|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\right\|_F^2 \right) + Tm\Upsilon. \tag{3.49}$$

Hence, by the choice of τ and (3.7), we obtain

$$\begin{aligned} \mathbb{E} \left\|\mathbf{R}^{(\tau)}\right\|_F^2 &\leq \tfrac{L^2}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(2 \left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 + 8 \left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 + \left\|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\right\|_F^2 \right) + m\Upsilon \\ &\overset{(3.31),(3.48)}{\leq} \tfrac{2mL}{T} \hat{\delta}_0 + \tfrac{L^2}{2\kappa T} \left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2 + 2m\Upsilon \\ &\quad + L^2 \left(\tfrac{C_0}{T} + \left(\tfrac{1}{30L_P} + \tfrac{1}{L_P(1-\rho)^2} + \tfrac{L+1}{8L^2} \right) \Upsilon \right) \cdot \left(20m\eta_{\mathbf{x}}(24\kappa^2 + 2\kappa + 4) + 4m\eta_{\boldsymbol{\Lambda}}(12\kappa^2 + \kappa + 1) \right). \end{aligned} \tag{3.50}$$

Therefore, plugging (3.47) and (3.50) into (3.14) and (3.15) with $t = \tau$, we obtain the desired results from (3.31) and by combining like terms. □

Below we give the total sample and communication complexity result. The proof is given in Appendix D.

Corollary 1 *Let $\varepsilon > 0$ be given and assume $L \geq 1$. Under the same assumptions as in Theorem 3.3, suppose*

$$\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(0)}\right\|_F^2 = \mathcal{O}\big(mL_P(1-\rho)\big), \quad \left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2 = \mathcal{O}\left(\min\left\{ \tfrac{m\kappa}{L}, \ \tfrac{mL_P\kappa(1-\rho)^2}{L^2} \right\} \right).$$

Then Algorithm 1 with VR-tag = SPIDER, $\mathcal{S}_1 = \Theta\left(\frac{\sigma^2 \cdot \max\{\kappa^2, (1-\rho)^{-4}\}}{\varepsilon^2} \right)$ for the general stochastic case and $\mathcal{S}_0 = \mathcal{S}_1 = n$ for the special finite-sum case, and $\mathcal{S}_2 = q = \lceil \sqrt{\mathcal{S}_1} \rceil$ can find an ε-stationary point in expectation of (2.2) by T_s stochastic gradients and T_c local neighbor communications, where

$$T_c = \Theta\left(\tfrac{L\kappa^2}{\varepsilon^2 \cdot \min\{1, \kappa(1-\rho)^2\}} \left(\phi(\bar{\mathbf{x}}^{(0)}, \boldsymbol{\Lambda}^{(0)}) - \phi^* + \tfrac{\hat{\delta}_0}{\min\{1, \kappa(1-\rho)^2\}} + 1 \right) \right)$$

and $T_s = n + \sqrt{n} T_c$ for the finite-sum case and $T_s = \frac{\sigma}{\varepsilon} \cdot \max\{\kappa, (1-\rho)^{-2}\} T_c$ for the general stochastic case.

Remark 3.4 The assumption on $\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(0)}\right\|_F^2$ and $\left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2$ is mild and can easily hold if $\kappa(1-\rho)$ is not small. Otherwise, multiple communications can be performed at the *initial* step to satisfy the condition. In addition, when $\kappa(1-\rho)^2 \geq 1$, the complexity results will be independent of the graph topology. In this case, our stepsize $\eta_{\mathbf{x}}$ and $\eta_{\boldsymbol{\Lambda}}$ in (3.30) are both in the order of $\frac{1}{L\kappa^2}$, matching to that used by a centralized method for NCSC minimax problems, e.g., the GDA in [20]. This can be significantly larger than the stepsize in the order of $\frac{1}{L\kappa^3}$ taken by state-of-the-art decentralized methods for minimax problems, e.g., PRECISION [23] and DSGDA [10]. When $\kappa(1-\rho)^2 \ll 1$, then T_c linearly depends on $(1-\rho)^{-4}$, which is worse than existing results with a dependence of $(1-\rho)^{-2}$ for decentralized methods with a *single* communication per iteration on solving composite nonconvex problems; see [37] for example. To improve the dependence, we can again perform multiple communications in the *initial* step to have $\hat{\delta}_0 = \mathcal{O}(\kappa(1-\rho)^2)$. Moreover, if the multi-communication trick, which needs more coordinations between agents, is applied for every update, we can change the mixing matrix $\mathbf{W}$ in our analysis to a polynomial in $\mathbf{W}$, denoted by $q(\mathbf{W})$. If $\mathbf{W}$ is symmetric, we can apply the Chebyshev polynomial (e.g., see [29]) to have an accelerated averaging. This way, the sample complexity will be independent of ρ and the communication complexity will linearly depend on $(1-\rho)^{-\frac{1}{2}}$.

3.5 Convergence results by STORM-type variance reduction

In this subsection, we set VR-tag = STORM in Algorithm 1. The general proof structure mimics that of Section 3.4. The proofs of all lemmas are given in Appendix E.

Lemma 3.9 *Let $\{(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{V}^{(t)})\}$ be generated from Algorithm 1 and $\mathbf{R}^{(t)}$ defined in (3.4). Then*

$$\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2 \le \rho\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2 + \tfrac{1}{1-\rho}\left(3L^2\mathbb{E}\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2 + 3\beta^2\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2 + 3m\beta^2\Upsilon_{t+1}\right), \tag{3.51}$$

$$\mathbb{E}\left\|\mathbf{R}^{(t+1)}\right\|_F^2 \le 2(1-\beta)^2L^2\mathbb{E}\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2 + (1-\beta)^2\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2 + 2m\beta^2\Upsilon_{t+1}, \tag{3.52}$$

where $\mathbf{Z}^{(t)} = (\mathbf{X}^{(t)}, \mathbf{Y}^{(t)})$ by the notation in (1.4a) and $\Upsilon_t := \frac{\sigma^2}{\mathcal{S}_t}$ for any $t \ge 0$.

In the rest of this subsection, we set

$$c_1 = c_2 = \tfrac{32\kappa^2}{\sqrt{\beta}},\ c_3 = \tfrac{60}{\sqrt{\beta}},\ c_4 = \tfrac{30\sqrt{2}\kappa^2 L}{\sqrt{\beta}},\ c_5 = \tfrac{60\sqrt{2m}\kappa^2}{\sqrt{\beta}},\ \eta_{\mathbf{y}} = \tfrac{\sqrt{\beta}}{4\sqrt{2}L}. \tag{3.53}$$

With Lemma 3.9, we can use (3.16) to show a result similar to Theorem 3.1.

Lemma 3.10 *Under Assumptions 3.1-3.3, let $\{(\mathbf{X}^{(t)}, \widetilde{\mathbf{X}}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{V}^{(t)})\}_{t\ge 0}$ be generated from Algorithm 1 with VR-tag = STORM, $\beta \in (0,1)$, $\eta_{\mathbf{y}} = \frac{\sqrt{\beta}}{4\sqrt{2}L}$, and $\eta_{\mathbf{x}}$ and $\eta_{\boldsymbol{\Lambda}}$ set to*

$$\eta_{\mathbf{x}} = \min\left\{\frac{\kappa(1-\rho)^2}{40L(24\kappa^2+8\kappa+5)},\ \frac{\sqrt{\beta}}{48(L+1)(24\kappa^2+7\kappa+4)}\right\}, \tag{3.54a}$$

$$\eta_{\boldsymbol{\Lambda}} = \min\left\{\frac{(1-\rho)^2}{4L(20\kappa+3)},\ \frac{\sqrt{\beta}}{4(L+1)(52\kappa^2+\kappa+1)}\right\}. \tag{3.54b}$$

Then it holds for any integer $T \ge 1$,

$$\begin{aligned}
&\tfrac{1}{4m\eta_{\mathbf{x}}}\textstyle\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2 + \tfrac{1}{6m\eta_{\mathbf{x}}}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}_{\perp}^{(t)}\right\|_F^2 + \tfrac{1}{2\eta_{\boldsymbol{\Lambda}}}\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2 \\
&+ \tfrac{\sqrt{\beta}(L+1)}{160m\kappa^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2 + \tfrac{\eta_{\mathbf{x}}}{m(1-\rho)^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2 + \tfrac{\sqrt{\beta}}{16mL}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2 \\
\le& C_0 + \left(\tfrac{1}{(1-\rho)^2\kappa L} + \tfrac{(1+1/L)}{\sqrt{\beta}L}\right)\beta^2\textstyle\sum_{t=0}^{T-1}\Upsilon_{t+1},
\end{aligned} \tag{3.55}$$

where $\mathbf{R}^{(t)}$ is defined in (3.4), Υ_t is defined in Lemma 3.9, and

$$\begin{aligned}
C_0 :=& \phi(\bar{\mathbf{x}}^{(0)}, \boldsymbol{\Lambda}^{(0)}) - \phi^* + \tfrac{1}{40m(1-\rho)\kappa L}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(0)}\right\|_F^2 + \left(\tfrac{(1+1/L)}{2\sqrt{\beta}mL} + \tfrac{1}{4m\kappa L(1-\rho)^2}\right)\left\|\mathbf{R}^{(0)}\right\|_F^2 \\
&+ \left(\tfrac{L+1}{5m\kappa} + \tfrac{\sqrt{\beta}L}{10m\kappa^2(1-\rho)^2}\right)\left\|\widetilde{\mathbf{Y}}^{(0)}-\mathbf{Y}^{(0)}\right\|_F^2 + \tfrac{\sqrt{\beta}}{\sqrt{2}}\left(\tfrac{1}{(1-\rho)^2\kappa} + \tfrac{(1+1/L)}{\sqrt{\beta}}\right)\hat{\delta}_0.
\end{aligned} \tag{3.56}$$

We have the following convergence rate result by Lemma 3.10.

Theorem 3.4 *Under Assumptions 3.1-3.3, let* $\{(\mathbf{X}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{Y}^{(t)})\}_{t\geq 0}$ *be generated from Algorithm 1 with* VR*-tag = STORM and* $\eta_{\mathbf{x}}, \eta_{\boldsymbol{\Lambda}}, \eta_{\mathbf{y}}$ *chosen as in Lemma 3.10. For any integer* $T \geq 1$*, select* τ *uniformly at random from* $\{0, \ldots, T-1\}$*. Then it holds that*

$$\mathbb{E}\left\|\tfrac{1}{\eta_{\mathbf{x}}}\left(\bar{\mathbf{x}}^{(\tau)} - \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\bar{\mathbf{x}}^{(\tau)} - \eta_{\mathbf{x}}\nabla P(\bar{\mathbf{x}}^{(\tau)}, \boldsymbol{\Lambda}^{(\tau)})\right)\right)\right\|_2^2 + \tfrac{L^2}{m}\mathbb{E}\left\|\mathbf{X}_{\perp}^{(\tau)}\right\|_F^2 \tag{3.57}$$
$$\leq \left(\tfrac{C_0}{T} + \left(\tfrac{1}{(1-\rho)^2\kappa L} + \tfrac{(1+1/L)}{\sqrt{\beta}L}\right)\beta^2\textstyle\sum_{t=0}^{T-1}\tfrac{\Upsilon_{t+1}}{T}\right)\cdot\left[\tfrac{20}{\eta_{\mathbf{x}}} + 6\eta_{\mathbf{x}}\left(2L^2(3+5\kappa^2) + \tfrac{5}{\eta_{\mathbf{x}}^2}\right) + \tfrac{1600\kappa^2 L}{\sqrt{\beta}} + \tfrac{80L}{\sqrt{\beta}} + \tfrac{5(1-\rho)^2}{\eta_{\mathbf{x}}}\right]$$

and

$$\mathbb{E}\left\|\nabla_{\boldsymbol{\Lambda}} P(\bar{\mathbf{x}}^{(\tau)}, \boldsymbol{\Lambda}^{(\tau)})\right\|_F^2 \leq \left(\tfrac{C_0}{T} + \left(\tfrac{1}{(1-\rho)^2\kappa L} + \tfrac{(1+1/L)}{\sqrt{\beta}L}\right)\beta^2\textstyle\sum_{t=0}^{T-1}\tfrac{\Upsilon_{t+1}}{T}\right)\cdot\left[\tfrac{4}{\eta_{\boldsymbol{\Lambda}}} + 24\kappa^2L^2\eta_{\mathbf{x}} + \tfrac{640\kappa^2 L}{\sqrt{\beta}}\right], \tag{3.58}$$

where Υ_t *is defined in Lemma 3.9, and* C_0 *is defined in* (3.56)*.*

Proof. The results follow directly from taking $\frac{1}{T}$ times of (3.55) and utilizing (3.14) and (3.15) with $t = \tau$. □

Based on Theorem 3.4, we give, in the following corollary, the complexity results of VRLM-STORM. For simplicity, we focus on the high-accuracy regime, i.e., ε is sufficiently small.

Corollary 2 *Let* $\varepsilon \in \left(0, \sigma(1-\rho)^2\right]$ *be given and assume* $L \geq 1$*. Under the same conditions as in Theorem 3.4, assume* $\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(0)}\right\|_F^2 \leq 40m(1-\rho)\kappa L$*,* $\left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2 = \mathcal{O}\left(\frac{m\kappa}{L}\right)$*, and* $\mathbb{E}\left\|\mathbf{R}^{(0)}\right\|_F^2 \leq \frac{m\sigma^2}{\mathcal{S}_0}$ *with* $\mathcal{S}_0 = \left\lceil\left(\frac{1}{\sqrt{\beta}L} + \frac{1}{4(1-\rho)^2\kappa L}\right)\sigma^2\right\rceil$ *and* $\mathcal{S}_t = \mathcal{O}(1)$ *for all* $t \geq 1$*. Then, Algorithm 1 with* VR*-tag = STORM and*

$$\beta = \frac{\varepsilon^2}{1440\sigma^2(24\kappa^2 + 7\kappa + 4)}, \tag{3.59}$$

can find an ε*-stationary point in expectation of* (2.2) *by* $T_s = \Theta\left(\frac{\sigma\kappa^3 L}{\varepsilon^3} + \frac{\sigma^3\kappa}{\varepsilon L}\right)$ *stochastic gradients and* $T_c = \Theta\left(\frac{\sigma\kappa^3 L}{\varepsilon^3}\right)$ *local neighbor communications.*

A few remarks are in order regarding the above corollary.

Remark 3.5 First, similar to other recent works [28, 47], our complexity results are stated for the high-accuracy regime (i.e. small enough ε). Technically, this accuracy requirement removes the final dependence of VRLM-STORM on the spectrum of the graph. Second, the initialization requirements in Corollary 2 are akin to those in [28]. If necessary, e.g., when $1-\rho$ is much smaller than $\frac{1}{\kappa L}$, we can perform multiple communications at the initial step to have $\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(0)}\right\|_F^2 \leq 40m(1-\rho)\kappa L$. Third, the recent Acc-MDA [12] method can attain a convergence rate of $\tilde{\mathcal{O}}(\kappa^{4.5}\varepsilon^{-3})$, however, *this is only in the single-agent setting* where $m = 1$ and further, this result only holds when $g \equiv h \equiv 0$. When $m = 1$, we have $\mathbf{W} = 1$, which eliminates the variable $\boldsymbol{\Lambda}$, as well as the consensus constraint on $\{\mathbf{x}_i\}_{i\in\mathcal{V}}$. Hence, (2.2) is actually identical to (1.1) which indicates our analysis provides a better dependence on κ, as well proves convergence on a broader class of problems (i.e., $g \not\equiv 0$ and/or $h \not\equiv 0$).

4 Numerical experiments

In this section, we empirically validate our proposed methods on two benchmark problems which fit (1.3). We compare our methods to three methods: DPSOG [22], DM-HSGD [44], and GT-SRVR [54]; since these methods are only presented for the case of (1.3) with $g \equiv 0$, we simply wrap their $\mathbf{x}_i$ updates with $\mathbf{prox}_{\eta_{\mathbf{x}} g}(\cdot)$. For our experiments, we use $m = 8$ agents, where each agent is an NVIDIA Tesla V100 GPU. The agents are connected via a ring structured graph, where self-weighting and neighbor weighting are $\frac{1}{3}$. Our code is made available at https://github.com/RPI-OPT/VRLM.

4.1 Sparse distributionally robust optimization

We test our proposed method on the decentralized distributionally robust optimization problem [48] using both the MNIST [17] and Fashion-MNIST [45] datasets. Each agent maintains a local dataset, $\{\mathbf{a}_{ij}, b_{ij}\}_{j=1}^{n}$, where $\mathbf{a}_{ij} \in \mathbb{R}^{28\times 28}$ is the j-th image on the i-th agent and $b_{ij} \in \{1, \dots, C\}$ is the corresponding label among C classes. The total number of data points is given by $N = mn$ and we let $\mathcal{J}_i \subseteq \{1, \dots, N\}$ be an index set which contains the indices of data points on agent i. The agents' local objective functions are given by

$$f_i(\mathbf{x}_i, \mathbf{y}_i) = m \sum_{j\in\mathcal{J}_i} (\mathbf{y}_{i,j}) \ell\left(Z_{\mathbf{x}_i}(\mathbf{a}_{ij}), b_{ij}\right) - \tfrac{\mu}{2} \left\|\mathbf{y}_i - \tfrac{\mathbf{1}}{N}\right\|_2^2, \quad h(\mathbf{y}_i) \equiv \mathbb{I}_{\Delta_N}(\mathbf{y}_i), \quad g(\mathbf{x}_i) \equiv \lambda \left\|\mathbf{x}_i\right\|_1. \tag{4.1}$$

Here, $\mathbf{y}_{i,j}$ denotes the j-th component of $\mathbf{y}_i$, ℓ is the cross-entropy loss function taken over each class, $Z_{\mathbf{x}_i}$ is a neural network governed by the parameter $\mathbf{x}_i$, and μ is a parameter controling the deviation from the uniform distribution. The set $\Delta_N := \{\mathbf{y} : \mathbf{y}^\top \mathbf{1} = 1, \mathbf{y} \geq \mathbf{0}\}$ is the standard probability simplex. The regularizer g induces sparsity on the $\mathbf{x}$ variable. We choose $Z_{\mathbf{x}_i}$ to be a two-layer neural network with 200 hidden units and Tanh activation function. Further, we let $\mu = 10.0$ and $\lambda = 5 \times 10^{-4}$.

The dataset is split uniformly at random among the agents, hence each agent receives $n = 7{,}500$ local data points. We let the mini-batch size for all methods be 100 and tune $\eta_{\mathbf{y}} \in \{0.1, 0.01, 0.001, 0.0001\}$ and $\frac{\eta_{\mathbf{x}}}{\eta_{\mathbf{y}}} \in \{1, 0.1, 0.01, 0.001\}$. For DM-HSGD, we tune $\beta_{\mathbf{x}} = \beta_{\mathbf{y}} = 0.01$ following the paper guidelines and set the initial batch-size to 3000. For GT-SRVR, we tune $q \in \{100, 300\}$ and the large mini-batch size from $\{3000, 7500\}$. For `VRLM`-STORM we let $\beta = 0.01$, $L = \frac{2}{\sqrt{m}}$, and $\eta_{\boldsymbol{\Lambda}} = 0.001$. For `VRLM`-SPIDER, we tune $q \in \{100, 300\}$ and the large mini-batch size from $\{3000, 7500\}$ and let $L = \frac{2}{\sqrt{m}}$ and $\eta_{\boldsymbol{\Lambda}} = 0.001$. We run all methods to 50,000 iterations and compare the training loss value (as if performing centralized training; note this is not the objective value), testing accuracy, average number of non-zeros, and stationarity violation, where the stationarity violation is computed as

$$\left\|\bar{\mathbf{x}} - \mathbf{prox}_g\left(\bar{\mathbf{x}} - \tfrac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\bar{\mathbf{x}}, \mathbf{y}^{(*)})\right)\right\|_2^2 + \left\|\mathbf{X}_\perp\right\|_F^2 + \left\|\mathbf{Y}_\perp\right\|_F^2 \tag{4.2}$$

where $\mathbf{y}^{(*)} := \arg\max_{\mathbf{y}} \frac{1}{m}\sum_{i=1}^m f_i(\bar{\mathbf{x}}, \mathbf{y})$ for $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i$. We run each experiment with 5 different initial seeds and report the average results across both iterations and epochs.

From Figure 1, we can see that `VRLM` (both variants) outperform all competing methods in terms of training loss and testing accuracy, while `VRLM`-STORM can additionally find sparser solutions. For this problem, DM-HSGD is not competitive when $\lambda > 0$. All methods appear to struggle with reducing (4.2) for both datasets, but we remark that this does not appear to affect each method's ability to minimize the training loss and obtain reasonable testing accuracy.

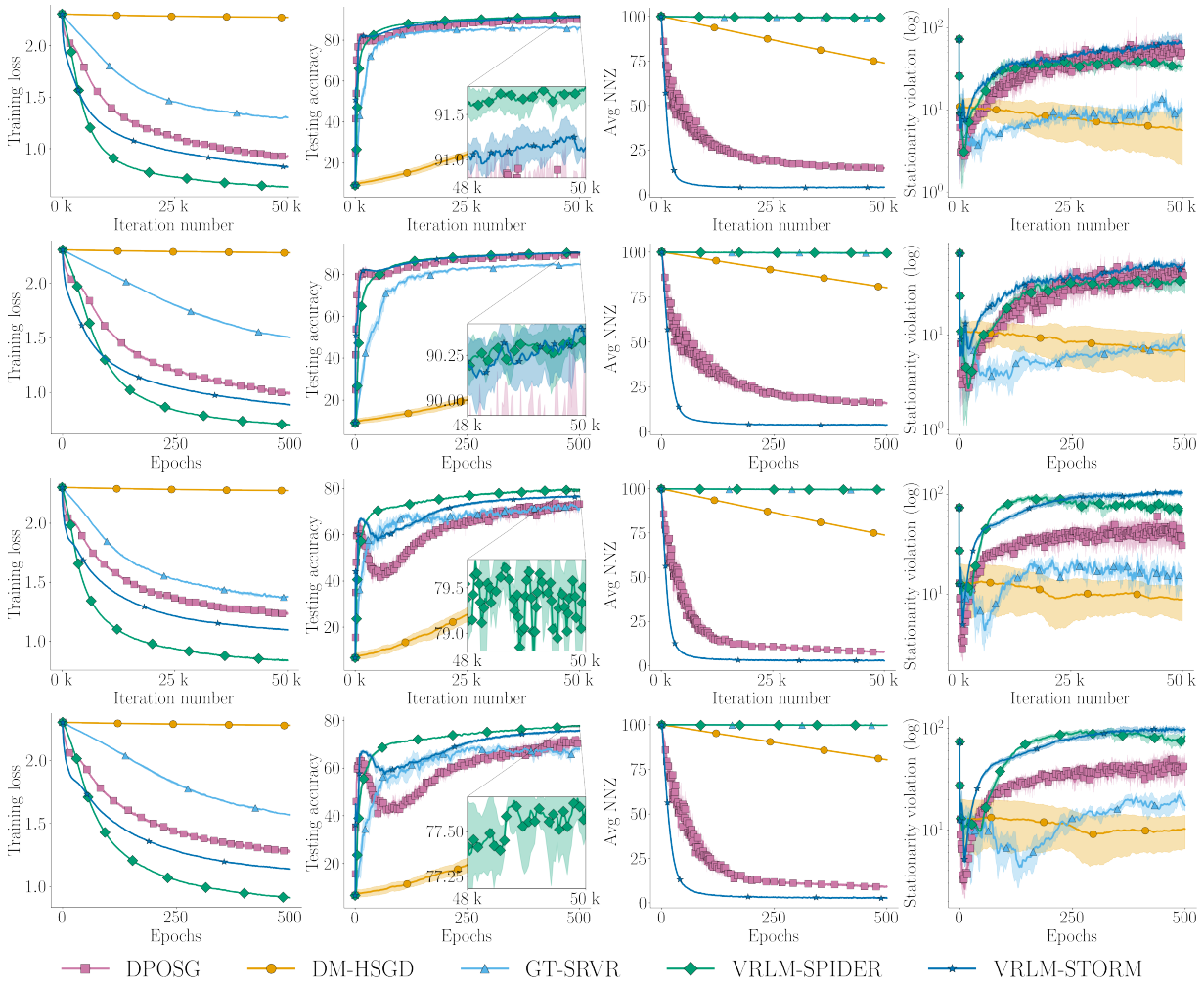

Fig. 1 Experimental results for the sparse distributionally robust optimization problem (4.1) using the MNIST and Fashion-MNIST datasets. The top two row depict results for the MNIST dataset in terms of iteration number and epochs, respectively. The bottom two rows depict the same results for the Fashion-MNIST dataset. Shaded regions represent 95% confidence intervals computed over 5 trials.

4.2 Fair Classification

We also test our method on the Fair Classification problem [30] using the CIFAR-10 [16] dataset. Each agent maintains a local dataset, $\{\mathbf{a}_{ij}, b_{ij}\}_{j=1}^{n}$, where $\mathbf{a}_{ij} \in \mathbb{R}^{32\times 32}$ is the j-th image on the i-th agent and $b_{ij} \in \{1, \ldots, C\}$ is the corresponding label. The agents' local objective functions are given by

$$f_i(\mathbf{x}_i, \mathbf{y}_i) = \sum_{c=1}^{C} (\mathbf{y}_{i,c}) \ell \left(Z_{\mathbf{x}_i}(\{\mathbf{a}_{ij}\}_{b_{ij}=c}), \{b_{ij}\}_{b_{ij}=c} \right) - \tfrac{\mu}{2} \|\mathbf{y}_i\|_2^2, \quad h(\mathbf{y}_i) \equiv \mathbb{I}_{\Delta_C}(\mathbf{y}_i), \quad g(\mathbf{x}_i) \equiv 0, \tag{4.3}$$

where $\mathbf{y}_{i,c}$ denotes the c-th component of $\mathbf{y}_i$, ℓ is the cross-entropy loss function which computes the average loss over each class, $Z_{\mathbf{x}_i}$ is a neural network governed by the parameter $\mathbf{x}_i$, μ is a parameter to tune, and Δ_C is the standard probability simplex. We choose $Z_{\mathbf{x}_i}$ to be the All-CNN network [39] and let $\mu = 0.1$.

The dataset is split uniformly at random among the agents, hence each agent receives $n = 6{,}250$ local data points. We let the mini-batch size for all methods be 100 and tune $\eta_{\mathbf{y}} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\frac{\eta_{\mathbf{x}}}{\eta_{\mathbf{y}}} \in \{1, 0.1, 0.01, 0.001\}$. For DM-HSGD, we tune $\beta_{\mathbf{x}} = \beta_{\mathbf{y}} \in \{0.01, 0.1, 0.5, 0.8, 0.9, 0.99\}$ and set the initial batch-size to 3000. For GT-SRVR, we tune $q \in \{100, 300\}$ and the large mini-batch size from $\{2000, 3000\}$. For `VRLM`-STORM, we tune $\beta \in \{0.01, 0.1, 0.5, 0.8, 0.9, 0.99\}$ and let $L = \frac{2}{\sqrt{m}}$ and $\eta_{\boldsymbol{\Lambda}} = 0.1$. We found `VRLM`-SPIDER noncompetitive on this instance, and the results are omitted. We run all methods to 20,000 iterations and compare the training loss value (as if performing centralized training; note this is not the objective value), testing accuracy, and stationarity violation, where the stationarity violation is computed as

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\bar{\mathbf{x}}, \mathbf{y}_i^{(*)}) \right\|_2^2 + \|\mathbf{X}_\perp\|_F^2 + \|\mathbf{Y}_\perp\|_F^2 \tag{4.4}$$

where $\mathbf{y}_i^{(*)} := \arg\max_{\mathbf{y}_i} f_i(\bar{\mathbf{x}}, \mathbf{y}_i)$ for $\bar{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i$. We use (4.4) instead of (4.2) due to the memory constraint of our computing environment which prohibits us from putting all the data on one machine. Again, we run each experiment with 5 different initial seeds and report the average results across both iterations and epochs.

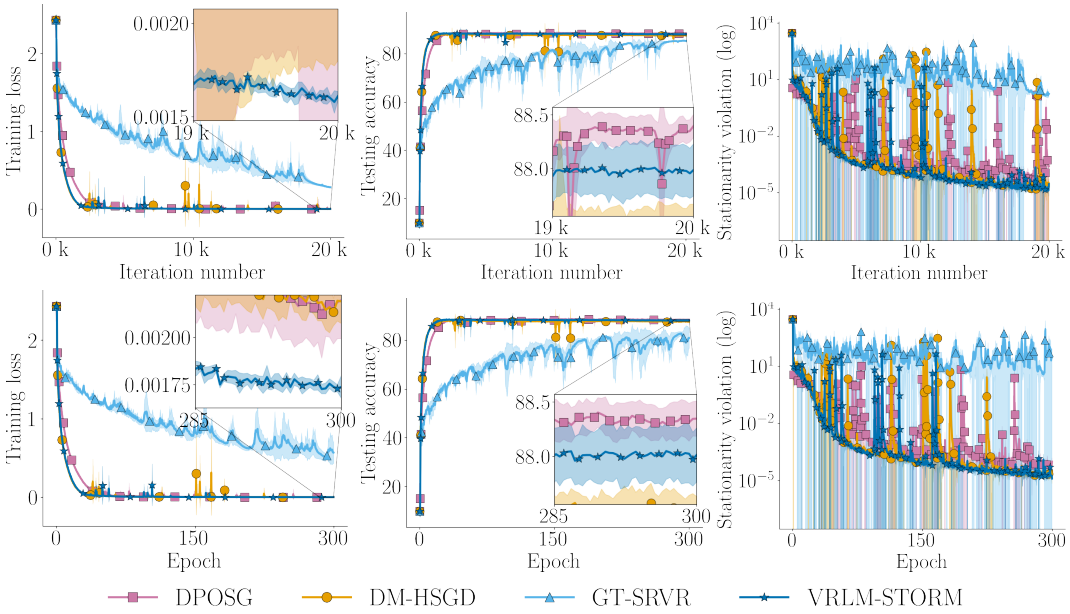

Fig. 2 Experimental results for the Fair Classification problem (4.3) using the All-CNN network with the CIFAR-10 dataset. The top row depicts results in terms of iteration number, while the bottom is in terms of epochs. Shaded regions represent 95% confidence intervals computed over 5 trials.

From Figure 2, we can see that DM-HSGD and our proposed method perform similarly in terms of stationarity violation. However, our method can achieve higher testing accuracy and lower training loss. DPOSG yields high testing accuracy, but its convergence for the double-regularized problem (2.2) has not been studied. The GT-SRVR method is not competitive on this problem, which could be due to the lack of theoretical guarantees for the double-regularized problem (2.2), or due to the periodic large-batch stochastic gradient computation. Overall, we find our proposed method to be competitive against other state-of-the-art methods, while providing improved theoretical guarantees.

5 Conclusions

In this work, we have presented the Variance Reduced Lagrangian Multiplier (VRLM) based method for solving the decentralized, double-regularized, stochastic nonconvex strongly-concave minimax problem. We analyzed VRLM with both big-batch and small-batch variance-reduction techniques. Under mild assumptions, both versions are able to achieve the best-known complexity results that are achieved by existing methods for solving special cases of the problem we consider. Finally, we demonstrated the effectiveness of our proposed methods in a real decentralized computing environment on two benchmark machine learning problems.

A Proof of Lemma 3.1

By the definition of $\widehat{\mathbf{Y}}^{(t)}$ in (3.3) and (3.6), we have

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{prox}_{\eta_{\mathbf{x}}g}\left(\bar{\mathbf{x}}^{(t)}-\eta_{\mathbf{x}}\nabla_{\mathbf{x}}P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right)\right\|_2=\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{prox}_{\eta_{\mathbf{x}}g}\left(\bar{\mathbf{x}}^{(t)}-\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}_j^{(t)})\right)\right\|_2\\
\leq&\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{prox}_{\eta_{\mathbf{x}}g}\left(\tilde{\mathbf{x}}_i^{(t)}-\eta_{\mathbf{x}}\mathbf{v}_{\mathbf{x},i}^{(t)}\right)\right\|_2+\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}_j^{(t)})-\left(\tilde{\mathbf{x}}_i^{(t)}-\eta_{\mathbf{x}}\mathbf{v}_{\mathbf{x},i}^{(t)}\right)\right\|_2\\
\leq&\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t+1)}\right\|_2+\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\tilde{\mathbf{x}}_i^{(t)}\right\|_2+\mathbb{E}\left\|\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\mathbf{x}_j^{(t)},\mathbf{y}_j^{(t)})-\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}_j^{(t)})\right\|_2\\
&+\left\|\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\mathbf{x}_j^{(t)},\mathbf{y}_j^{(t)})-\eta_{\mathbf{x}}\mathbf{v}_{\mathbf{x},i}^{(t)}\right\|_2\\
\leq&\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t+1)}\right\|_2+\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\tilde{\mathbf{x}}_i^{(t)}\right\|_2+\mathbb{E}\left\|\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\mathbf{x}_j^{(t)},\mathbf{y}_j^{(t)})-\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}_j^{(t)})\right\|_2\\
&+\left\|\tfrac{\eta_{\mathbf{x}}}{m}\textstyle\sum_{j=1}^m\nabla_{\mathbf{x}}f_j(\mathbf{x}_j^{(t)},\mathbf{y}_j^{(t)})-\eta_{\mathbf{x}}\bar{\mathbf{d}}_{\mathbf{x}}^{(t)}\right\|_2+\left\|\eta_{\mathbf{x}}\bar{\mathbf{v}}_{\mathbf{x}}^{(t)}-\eta_{\mathbf{x}}\mathbf{v}_{\mathbf{x},i}^{(t)}\right\|_2,
\end{aligned}\tag{A.1}$$

where the first inequality follows from the triangle inequality and the nonexpansiveness of $\mathbf{prox}_{\eta_{\mathbf{x}}g}$, the second inequality uses the update of $\mathbf{x}_i^{(t+1)}$ and the triangle inequality, and the last inequality holds by $\bar{\mathbf{d}}_{\mathbf{x}}^{(t)}=\bar{\mathbf{v}}_{\mathbf{x}}^{(t)}$ for all $t\geq 0$. Squaring both sides of (A.1), using Young's inequality, and summing over $i=1,\ldots,m$ with the definition of $\mathbf{R}_{\mathbf{x}}$ in (3.4) yields

$$\begin{aligned}
&m\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{prox}_{\eta_{\mathbf{x}}g}\left(\bar{\mathbf{x}}^{(t)}-\eta_{\mathbf{x}}\nabla_{\mathbf{x}}P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right)\right\|_2^2\\
\leq&5\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+5\mathbb{E}\left\|\bar{\mathbf{X}}^{(t)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2+5\eta_{\mathbf{x}}^2L^2\left(\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\mathbb{E}\left\|\widehat{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\right)\\
&+5\eta_{\mathbf{x}}^2\left\|\mathbf{R}_{\mathbf{x}}^{(t)}\right\|_F^2+5\eta_{\mathbf{x}}^2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\\
\overset{(3.11)}{\leq}&5\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\left(5\eta_{\mathbf{x}}^2L^2+5+10\eta_{\mathbf{x}}^2L^2\kappa^2\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+10\eta_{\mathbf{x}}^2L^2\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\\
&+5\eta_{\mathbf{x}}^2\left\|\mathbf{R}_{\mathbf{x}}^{(t)}\right\|_F^2+5\eta_{\mathbf{x}}^2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2
\end{aligned}\tag{A.2}$$

where we have used $\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top=(\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)(\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)$ to have $\left\|\bar{\mathbf{X}}^{(t)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2\le\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2$. By multiplying $\frac{1}{m\eta_{\mathbf{x}}^2}$ to (A.2), adding $\frac{L^2}{m}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2$, and using $\left\|\mathbf{R}_{\mathbf{x}}^{(t)}\right\|_F^2\le\left\|\mathbf{R}^{(t)}\right\|_F^2$, we obtain

$$\begin{aligned}&\mathbb{E}\left\|\tfrac{1}{\eta_{\mathbf{x}}}\left(\bar{\mathbf{x}}^{(t)}-\mathbf{prox}_{\eta_{\mathbf{x}}g}\left(\bar{\mathbf{x}}^{(t)}-\eta_{\mathbf{x}}\nabla_{\mathbf{x}}P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right)\right)\right\|_2^2+\tfrac{L^2}{m}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2\\ \le&\tfrac{5}{m\eta_{\mathbf{x}}^2}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\left(\tfrac{2L^2(3+5\kappa^2)}{m}+\tfrac{5}{m\eta_{\mathbf{x}}^2}\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\tfrac{10L^2}{m}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\\ &+\tfrac{5}{m}\left\|\mathbf{R}^{(t)}\right\|_F^2+\tfrac{5}{m}\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2.\end{aligned}\tag{A.3}$$

This completes the proof of (3.14). Again, by the definition of $\widehat{\mathbf{Y}}^{(t)}$ in (3.3), in conjunction with (3.6) and Young's inequality, we have

$$\begin{aligned}\mathbb{E}\left\|\nabla_{\boldsymbol{\Lambda}}P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right\|_F^2\le&2\mathbb{E}\left\|\tfrac{L}{2\sqrt{m}}(\mathbf{I}-\mathbf{W})\mathbf{Y}^{(t)}\right\|_F^2+\tfrac{L^2}{2m}\mathbb{E}\left\|(\mathbf{I}-\mathbf{W})(\widehat{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)})\right\|_F^2\\ \le&\tfrac{2}{\eta_{\boldsymbol{\Lambda}}^2}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2+\tfrac{2L^2}{m}\mathbb{E}\left\|\widehat{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\\ \le&\tfrac{2}{\eta_{\boldsymbol{\Lambda}}^2}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2+\tfrac{4L^2\kappa^2}{m}\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\tfrac{4L^2}{m}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2,\end{aligned}\tag{A.4}$$

where we have used (2.7), $\|\mathbf{I}-\mathbf{W}\|_2^2\le 4$ by Assumption 3.2, and (3.11). This completes the proof.

B Proof of Lemma 3.2

To prove Lemma 3.2, we first state a few supporting lemmas. The following lemma can be proved in the same way as the proof of [28, Lemma C.3].

Lemma B.1 *For all $t\ge 0$ and for all $i=1,\dots,m$,*

$$g(\mathbf{x}_i^{(t+1)})-g(\bar{\mathbf{x}}^{(t)})\le-\frac{1}{2\eta_{\mathbf{x}}}\left(\left\|\mathbf{x}_i^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2+\left\|\mathbf{x}_i^{(t+1)}-\tilde{\mathbf{x}}^{(t)}\right\|_2^2-\left\|\bar{\mathbf{x}}^{(t)}-\tilde{\mathbf{x}}^{(t)}\right\|_2^2\right)-\left\langle\mathbf{x}_i^{(t+1)}-\bar{\mathbf{x}}^{(t)},\mathbf{v}_{\mathbf{x},i}^{(t)}\right\rangle.\tag{B.1}$$

Lemma B.2 *For all $t\ge 0$ and arbitrary constants $c_1,c_2>0$, the following inequality holds*

$$\begin{aligned}&\phi(\bar{\mathbf{x}}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\phi(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\\ \le&\frac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle-\frac{1}{m}\sum_{i=1}^m\left\langle\mathbf{x}_i^{(t+1)}-\bar{\mathbf{x}}^{(t)},\mathbf{v}_{\mathbf{x},i}^{(t)}\right\rangle\\ &-\frac{1}{2m\eta_{\mathbf{x}}}\left(\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2-\left\|\widetilde{\mathbf{X}}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2\right)\\ &+\left(\frac{L(\kappa+1)}{2m}+\frac{\kappa^2}{2mc_2}\right)\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\frac{L(\kappa+1+c_1)+L_P}{2}\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2+\left(\frac{L}{2mc_1}+\frac{1}{2mc_2}\right)\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\\ &+\frac{Lc_2}{4}\left\|(\mathbf{W}-\mathbf{I})^\top\left(\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right)\right\|_F^2-\left(\frac{1}{\eta_{\boldsymbol{\Lambda}}}-\frac{L_P}{2}\right)\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned}\tag{B.2}$$

Proof. By the L_P-smoothness of P defined in (3.1), it holds that

$$\begin{aligned}&\phi(\bar{\mathbf{x}}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\phi(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\\ \le&\frac{L_P}{2}\left(\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2+\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\right)+\left\langle\nabla P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)}),\left(\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right)\right\rangle+g(\bar{\mathbf{x}}^{(t+1)})-g(\bar{\mathbf{x}}^{(t)})\\ \le&\frac{L_P}{2}\left(\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2+\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\right)+\left\langle\nabla P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)}),\left(\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right)\right\rangle\end{aligned}\tag{B.3}$$

$$-\frac{1}{2m\eta_{\mathbf{x}}}\left(\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2-\left\|\widetilde{\mathbf{X}}^{(t)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2\right)-\frac{1}{m}\sum_{i=1}^m\left\langle\mathbf{x}_i^{(t+1)}-\bar{\mathbf{x}}^{(t)},\mathbf{v}_{\mathbf{x},i}^{(t)}\right\rangle$$

where the second inequality uses the convexity of g to have $g(\bar{\mathbf{x}}^{(t+1)})\le\frac{1}{m}\sum_{i=1}^m g(\mathbf{x}_i^{(t+1)})$ and (B.1). By the definition of $\nabla P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})$ in (3.6) and the definition of $\widehat{\mathbf{Y}}^{(t)}$ in (3.3), we have

$$\begin{aligned}&\left\langle\nabla P(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)}),(\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)})\right\rangle\\=&\frac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle-\frac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})\widehat{\mathbf{Y}}^{(t)},\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\rangle.\end{aligned}\tag{B.4}$$

By the L-smoothness of $\{f_i\}$ and the Peter-Paul inequality, we further have

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle\\=&\frac{1}{m}\sum_{i=1}^m\left(\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle+\left\langle\nabla_{\mathbf{x}}f_i(\bar{\mathbf{x}}^{(t)},\hat{\mathbf{y}}_i^{(t)})-\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\hat{\mathbf{y}}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle\right)\\+&\frac{1}{m}\sum_{i=1}^m\left(\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\hat{\mathbf{y}}_i^{(t)})-\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\tilde{\mathbf{y}}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle+\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\tilde{\mathbf{y}}_i^{(t)})-\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle\right)\\\le&\frac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle+\frac{L}{2m}\sum_{i=1}^m\left(\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t)}\right\|_2^2+\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2\right)\\&+\frac{L}{2m}\sum_{i=1}^m\left(\frac{1}{\kappa}\left\|\hat{\mathbf{y}}^{(t)}-\tilde{\mathbf{y}}^{(t)}\right\|_2^2+\kappa\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2\right)+\frac{L}{2m}\sum_{i=1}^m\left(\frac{1}{c_1}\left\|\tilde{\mathbf{y}}^{(t)}-\mathbf{y}_i^{(t)}\right\|_2^2+c_1\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2\right)\\\le&\frac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle+\frac{L(\kappa+1)}{2m}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2\\&+\frac{L(\kappa+1+c_1)}{2m}\left\|\bar{\mathbf{X}}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2+\frac{L}{2mc_1}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\end{aligned}\tag{B.5}$$

where the last inequality uses (3.10b). Additionally, by the Peter-Paul inequality,

$$\begin{aligned}&-\frac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})\widehat{\mathbf{Y}}^{(t)},\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\rangle\\=&-\frac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})(\widehat{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)})+(\mathbf{W}-\mathbf{I})\mathbf{Y}^{(t)},\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\rangle\\\overset{(2.7)}{=}&-\frac{L}{2\sqrt{m}}\left\langle\widehat{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)},(\mathbf{W}-\mathbf{I})^\top(\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)})\right\rangle-\frac{1}{\eta_{\boldsymbol{\Lambda}}}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\\\le&\frac{L}{4mc_2}\left\|\widehat{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2+\frac{Lc_2}{4}\left\|(\mathbf{W}-\mathbf{I})^\top\left(\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right)\right\|_F^2-\frac{1}{\eta_{\boldsymbol{\Lambda}}}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\\\le&\frac{1}{2mc_2}\left(\left\|\widehat{\mathbf{Y}}^{(t)}-\widetilde{\mathbf{Y}}^{(t)}\right\|_F^2+\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\right)+\frac{Lc_2}{4}\left\|(\mathbf{W}-\mathbf{I})^\top\left(\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right)\right\|_F^2-\frac{1}{\eta_{\boldsymbol{\Lambda}}}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\\\le&\frac{1}{2mc_2}\left(\kappa^2\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2\right)+\frac{Lc_2}{4}\left\|(\mathbf{W}-\mathbf{I})^\top\left(\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right)\right\|_F^2-\frac{1}{\eta_{\boldsymbol{\Lambda}}}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\end{aligned}\tag{B.6}$$

where the last inequality uses (3.10b). Plugging (B.5) and (B.6) into (B.4) results in

$$\begin{aligned}
&\left\langle \nabla P(\bar{\mathbf{x}}^{(t)}, \boldsymbol{\Lambda}^{(t)}), \left(\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}, \boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right)\right\rangle \\
\leq & \frac{1}{m}\sum_{i=1}^{m}\left\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}), \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\rangle + \left(\frac{L(\kappa+1)}{2m} + \frac{\kappa^2}{2mc_2}\right)\left\|\mathbf{X}_{\perp}^{(t)}\right\|_F^2 \\
&+ \frac{L(\kappa+1+c_1)}{2m}\left\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + \left(\frac{L}{2mc_1} + \frac{1}{2mc_2}\right)\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2 \\
&+ \frac{Lc_2}{4}\left\|(\mathbf{W}-\mathbf{I})^{\top}\left(\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right)\right\|_F^2 - \frac{1}{\eta_{\boldsymbol{\Lambda}}}\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2.
\end{aligned} \tag{B.7}$$

Utilizing (B.7) in (B.3) and noting $\frac{1}{m}\left\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 = \left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2$ completes the proof. □

Lemma B.3 *For all $t \geq 0$, it holds that*

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\left\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}), \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\rangle - \frac{1}{m}\sum_{i=1}^{m}\left\langle \mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t)}, \mathbf{v}_{\mathbf{x},i}^{(t)}\right\rangle \\
\leq & \frac{1}{2m}\sum_{i=1}^{m}\left(L_P c_3\left\|\mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{1}{L_P c_3}\left\|\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \mathbf{v}_{\mathbf{x},i}^{(t)}\right\|_2^2\right),
\end{aligned} \tag{B.8}$$

where $c_3 > 0$ is an arbitrary constant.

Proof. Notice

$$\begin{aligned}
&\left\langle \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}), \bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\rangle - \frac{1}{m}\sum_{i=1}^{m}\left\langle \mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t)}, \mathbf{v}_{\mathbf{x},i}^{(t)}\right\rangle \\
= & \frac{1}{m}\sum_{i=1}^{m}\left\langle \frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}), \mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\rangle - \frac{1}{m}\sum_{i=1}^{m}\left\langle \mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t)}, \mathbf{v}_{\mathbf{x},i}^{(t)}\right\rangle \\
= & \frac{1}{m}\sum_{i=1}^{m}\left\langle \frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \mathbf{v}_{\mathbf{x},i}^{(t)}, \mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\rangle.
\end{aligned}$$

Then the desired result follows from the Peter-Paul inequality. □

Proof. (Of Lemma 3.2) The proof follows from applying (B.8) to (B.2) to have

$$\begin{aligned}
\phi(\bar{\mathbf{x}}^{(t+1)}, \boldsymbol{\Lambda}^{(t+1)}) - \phi(\bar{\mathbf{x}}^{(t)}, \boldsymbol{\Lambda}^{(t)}) \leq & \frac{1}{2m}\sum_{i=1}^{m}\left(L_P c_3\left\|\mathbf{x}_i^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{1}{L_P c_3}\left\|\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \mathbf{v}_{\mathbf{x},i}^{(t)}\right\|_2^2\right) \\
&- \frac{1}{2m\eta_{\mathbf{x}}}\left(\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + \left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 - \left\|\widetilde{\mathbf{X}}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2\right) + \left(\frac{L(\kappa+1)}{2m} + \frac{\kappa^2}{2mc_2}\right)\left\|\mathbf{X}_{\perp}^{(t)}\right\|_F^2 \\
&+ \frac{L(\kappa+1+c_1)+L_P}{2}\left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \left(\frac{L}{2mc_1} + \frac{1}{2mc_2}\right)\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2 \\
&+ \frac{Lc_2}{4}\left\|(\mathbf{W}-\mathbf{I})^{\top}\left(\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right)\right\|_F^2 - \left(\frac{1}{\eta_{\boldsymbol{\Lambda}}} - \frac{L_P}{2}\right)\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2
\end{aligned} \tag{B.9}$$

and using the inequalities $\left\|\widetilde{\mathbf{X}}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 = \left\|\left(\mathbf{W} - \frac{1}{m}\mathbf{1}\mathbf{1}^{\top}\right)\mathbf{X}_{\perp}^{(t)}\right\|_F^2 \leq \rho^2\left\|\mathbf{X}_{\perp}^{(t)}\right\|_F^2$, $\left\|\bar{\mathbf{x}}^{(t+1)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 \leq \frac{1}{m}\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2$, and $\|\mathbf{W}-\mathbf{I}\|_2^2 \leq 4$ to further upper bound the right-hand side of (B.9). □

C Proofs of Lemmas in Section 3.3

Proof. [Of Lemma 3.4] By Young's inequality, it holds that

$$\begin{aligned}
&\textstyle\sum_{i=1}^m \left\| \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \mathbf{v}_{\mathbf{x},i}^{(t)} \right\|_2^2 \\
\le& 2\textstyle\sum_{i=1}^m \left(\left\| \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{x}} f_j(\mathbf{x}_j^{(t)}, \mathbf{y}_j^{(t)}) - \bar{\mathbf{d}}_{\mathbf{x}}^{(t)} \right\|_2^2 + \left\| \bar{\mathbf{d}}_{\mathbf{x}}^{(t)} - \mathbf{v}_{\mathbf{x},i}^{(t)} \right\|_2^2 \right) \\
\le& 2\left\| \nabla_{\mathbf{x}} F(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}) - \mathbf{D}_{\mathbf{x}}^{(t)} \right\|_F^2 + 2\left\| \mathbf{V}_{\perp,\mathbf{x}}^{(t)} \right\|_F^2 \le 2\left\| \nabla F(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}) - \mathbf{D}^{(t)} \right\|_F^2 + 2\left\| \mathbf{V}_{\perp,\mathbf{x}}^{(t)} \right\|_F^2,
\end{aligned}$$

where the second inequality follows from Jensen's inequality and $\bar{\mathbf{d}}_{\mathbf{x}}^{(t)} = \bar{\mathbf{v}}_{\mathbf{x}}^{(t)}, \forall\, t \ge 0$. Taking the expectation yields (3.19). □

Proof. [Of Lemma 3.5] By the $\mathbf{Y}$ update defined in (2.8), it holds for all $i = 1, \dots, m$ that

$$\mathbf{0} \in \partial h(\mathbf{y}_i^{(t+1)}) + \frac{1}{\eta_{\mathbf{y}}} \left(\mathbf{y}_i^{(t+1)} - \left(\mathbf{y}_i^{(t)} + \eta_{\mathbf{y}} \mathbf{v}_{\mathbf{y},i}^{(t)} \right) \right) \tag{C.1}$$

and hence for some $\tilde{\nabla} h(\mathbf{y}_i^{(t+1)}) \in \partial h(\mathbf{y}_i^{(t+1)})$ and any $\mathbf{y}_i \in \mathrm{dom}(h)$, it holds that

$$0 = \left\langle \mathbf{y}_i^{(t+1)} - \mathbf{y}_i, \tilde{\nabla} h(\mathbf{y}_i^{(t+1)}) + \frac{1}{\eta_{\mathbf{y}}} \left(\mathbf{y}_i^{(t+1)} - \left(\mathbf{y}_i^{(t)} + \eta_{\mathbf{y}} \mathbf{v}_{\mathbf{y},i}^{(t)} \right) \right) \right\rangle. \tag{C.2}$$

By the convexity of h, we further have

$$h(\mathbf{y}_i) \ge h(\mathbf{y}_i^{(t+1)}) + \left\langle \mathbf{y}_i - \mathbf{y}_i^{(t+1)}, \tilde{\nabla} h(\mathbf{y}_i^{(t+1)}) \right\rangle \overset{(C.2)}{=} h(\mathbf{y}_i^{(t+1)}) + \left\langle \mathbf{y}_i^{(t+1)} - \mathbf{y}_i, \tfrac{1}{\eta_{\mathbf{y}}} \left(\mathbf{y}_i^{(t+1)} - \left(\mathbf{y}_i^{(t)} + \eta_{\mathbf{y}} \mathbf{v}_{\mathbf{y},i}^{(t)} \right) \right) \right\rangle. \tag{C.3}$$

Summing (C.3) over $i = 1, \dots, m$ and taking $\mathbf{y}_i = \mathbf{y}_i^{(t)}$ for all $i = 1, \dots, m$ gives

$$\tfrac{1}{\eta_{\mathbf{y}}} \left\| \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)} \right\|_F^2 - \left\langle \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}, \mathbf{V}_{\mathbf{y}}^{(t)} \right\rangle \le \textstyle\sum_{i=1}^m \left(h(\mathbf{y}_i^{(t)}) - h(\mathbf{y}_i^{(t+1)}) \right). \tag{C.4}$$

Now by the definition of $\Gamma_t(\mathbf{Y})$ from (3.5) and the update for $\mathbf{V}_{\mathbf{y}}$ from (2.6), it holds that

$$m \nabla \Gamma_t(\mathbf{Y}^{(t)}) - \mathbf{V}_{\mathbf{y}}^{(t)} = \nabla_{\mathbf{y}} F(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}) - \mathbf{D}_{\mathbf{y}}^{(t)}. \tag{C.5}$$

By Assumption 3.3, $-\Gamma_t(\cdot)$ is $\frac{L}{m}$-smooth for all $t \ge 0$ and hence

$$\begin{aligned}
-\left\langle \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}, \mathbf{V}_{\mathbf{y}}^{(t)} \right\rangle =& -m \left\langle \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}, \nabla \Gamma_t(\mathbf{Y}^{(t)}) \right\rangle - \left\langle \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}, \mathbf{D}_{\mathbf{y}}^{(t)} - \nabla_{\mathbf{y}} F(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}) \right\rangle \\
\ge& m \left(\Gamma_t(\mathbf{Y}^{(t)}) - \Gamma_t(\mathbf{Y}^{(t+1)}) \right) - \frac{L}{2} \left\| \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)} \right\|_F^2 - \left\langle \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}, \mathbf{R}_{\mathbf{y}}^{(t)} \right\rangle,
\end{aligned} \tag{C.6}$$

where we have used the definition of $\mathbf{R}_{\mathbf{y}}$ from (3.4). Next, we compute

$$\begin{aligned}
&- \Gamma_t(\mathbf{Y}^{(t+1)}) + \Gamma_{t+1}(\mathbf{Y}^{(t+1)}) \\
=& -\frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t+1)}) + \frac{L}{2\sqrt{m}} \left\langle (\mathbf{W} - \mathbf{I}) \mathbf{Y}^{(t+1)}, \boldsymbol{\Lambda}^{(t)} \right\rangle + \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}_i^{(t+1)}, \mathbf{y}_i^{(t+1)}) - \frac{L}{2\sqrt{m}} \left\langle (\mathbf{W} - \mathbf{I}) \mathbf{Y}^{(t+1)}, \boldsymbol{\Lambda}^{(t+1)} \right\rangle \\
\ge& \frac{1}{m} \sum_{i=1}^m \left(\left\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{(t+1)}, \mathbf{y}_i^{(t+1)}), \mathbf{x}_i^{(t+1)} - \mathbf{x}_i^{(t)} \right\rangle - \frac{L}{2} \left\| \mathbf{x}_i^{(t+1)} - \mathbf{x}_i^{(t)} \right\|_2^2 \right) + \frac{L}{2\sqrt{m}} \left\langle (\mathbf{W} - \mathbf{I}) \mathbf{Y}^{(t+1)}, \boldsymbol{\Lambda}^{(t)} - \boldsymbol{\Lambda}^{(t+1)} \right\rangle \\
=& \frac{1}{m} \sum_{i=1}^m \left(\left\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{(t+1)}, \tilde{\mathbf{y}}_i^{(t+1)}), \mathbf{x}_i^{(t+1)} - \mathbf{x}_i^{(t)} \right\rangle - \frac{L}{2} \left\| \mathbf{x}_i^{(t+1)} - \mathbf{x}_i^{(t)} \right\|_2^2 \right) \\
&+ \frac{1}{m} \left\langle \nabla_{\mathbf{x}} F(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}) - \nabla_{\mathbf{x}} F(\mathbf{X}^{(t+1)}, \widetilde{\mathbf{Y}}^{(t+1)}), \mathbf{X}^{(t+1)} - \mathbf{X}^{(t)} \right\rangle
\end{aligned} \tag{C.7}$$

$$+\frac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})\widetilde{\mathbf{Y}}^{(t+1)},\boldsymbol{\Lambda}^{(t)}-\boldsymbol{\Lambda}^{(t+1)}\right\rangle+\frac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})(\mathbf{Y}^{(t+1)}-\widetilde{\mathbf{Y}}^{(t+1)}),\boldsymbol{\Lambda}^{(t)}-\boldsymbol{\Lambda}^{(t+1)}\right\rangle$$

where the inequality uses Assumption 3.3 and $\widetilde{\mathbf{Y}}^{(t)}$ is defined in (3.3) for all $t\geq 0$. By

$$\nabla Q(\mathbf{X}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})=\left(\tfrac{1}{m}\nabla F(\mathbf{X}^{(t+1)},\widetilde{\mathbf{Y}}^{(t+1)})^{\top},\ -\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\widetilde{\mathbf{Y}}^{(t+1)}\right),$$

we obtain from (3.13) that

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{(t+1)},\tilde{\mathbf{y}}_i^{(t+1)}),\mathbf{x}_i^{(t+1)}-\mathbf{x}_i^{(t)}\right\rangle+\frac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})\widetilde{\mathbf{Y}}^{(t+1)},\boldsymbol{\Lambda}^{(t)}-\boldsymbol{\Lambda}^{(t+1)}\right\rangle\\ \geq&-Q(\mathbf{X}^{(t)},\boldsymbol{\Lambda}^{(t)})+Q(\mathbf{X}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\frac{L_Q}{2m}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2-\frac{L_Q}{2}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned}\tag{C.8}$$

Applying (C.8) to (C.7) and rearranging results in

$$\begin{aligned}&-\Gamma_t(\mathbf{Y}^{(t+1)})+Q(\mathbf{X}^{(t)},\boldsymbol{\Lambda}^{(t)})+\Gamma_{t+1}(\mathbf{Y}^{(t+1)})-Q(\mathbf{X}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})\\ \geq&\frac{1}{m}\left\langle\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)})-\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\widetilde{\mathbf{Y}}^{(t+1)}),\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\rangle\\ &+\frac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})(\mathbf{Y}^{(t+1)}-\widetilde{\mathbf{Y}}^{(t+1)}),\boldsymbol{\Lambda}^{(t)}-\boldsymbol{\Lambda}^{(t+1)}\right\rangle-\frac{L_Q+L}{2m}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2-\frac{L_Q}{2}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned}\tag{C.9}$$

Adding m times of (C.9) to (C.6) results in

$$\begin{aligned}&-\left\langle\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)},\mathbf{V}_{\mathbf{y}}^{(t)}\right\rangle+m\left(\Gamma_{t+1}(\mathbf{Y}^{(t+1)})-Q(\mathbf{X}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\Gamma_t(\mathbf{Y}^{(t)})+Q(\mathbf{X}^{(t)},\boldsymbol{\Lambda}^{(t)})\right)\\ \geq&-\frac{L}{2}\left\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\right\|_F^2-\left\langle\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)},\mathbf{R}_{\mathbf{y}}^{(t)}\right\rangle+\left\langle\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)})-\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\widetilde{\mathbf{Y}}^{(t+1)}),\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\rangle\\ &+\frac{L\sqrt{m}}{2}\left\langle(\mathbf{W}-\mathbf{I})(\mathbf{Y}^{(t+1)}-\widetilde{\mathbf{Y}}^{(t+1)}),\boldsymbol{\Lambda}^{(t)}-\boldsymbol{\Lambda}^{(t+1)}\right\rangle-\frac{L_Q+L}{2}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2-\frac{mL_Q}{2}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned}\tag{C.10}$$

Applying (C.10) to (C.4), using the definition of $\hat{\delta}_t$ from (3.21), and rearranging terms results in

$$\begin{aligned}&m(\hat{\delta}_t-\hat{\delta}_{t+1})\\ \geq&\left(\tfrac{1}{\eta_{\mathbf{y}}}-\tfrac{L}{2}\right)\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2-\left\langle\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)},\mathbf{R}_{\mathbf{y}}^{(t)}\right\rangle+\left\langle\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)})-\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\widetilde{\mathbf{Y}}^{(t+1)}),\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\rangle\\ &+\frac{L\sqrt{m}}{2}\left\langle(\mathbf{W}-\mathbf{I})(\mathbf{Y}^{(t+1)}-\widetilde{\mathbf{Y}}^{(t+1)}),\boldsymbol{\Lambda}^{(t)}-\boldsymbol{\Lambda}^{(t+1)}\right\rangle-\frac{L_Q+L}{2}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2-\frac{mL_Q}{2}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned}\tag{C.11}$$

By the Peter-Paul inequality, we have that for any $c_4,c_5>0$,

$$\left\langle\mathbf{Y}^{(t)}-\mathbf{Y}^{(t+1)},\mathbf{R}_{\mathbf{y}}^{(t)}\right\rangle\leq\frac{1}{4\eta_{\mathbf{y}}}\left\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\right\|_F^2+\eta_{\mathbf{y}}\left\|\mathbf{R}_{\mathbf{y}}^{(t)}\right\|_F^2,\tag{C.12}$$

$$\left\langle\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)})-\nabla_{\mathbf{x}}F(\mathbf{X}^{(t+1)},\widetilde{\mathbf{Y}}^{(t+1)}),\mathbf{X}^{(t)}-\mathbf{X}^{(t+1)}\right\rangle\leq\frac{L^2}{2c_4}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2+\frac{c_4}{2}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2,\tag{C.13}$$

$$\frac{L\sqrt{m}}{2}\left\langle(\mathbf{W}-\mathbf{I})(\mathbf{Y}^{(t+1)}-\widetilde{\mathbf{Y}}^{(t+1)}),\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\rangle\leq\frac{L\sqrt{m}}{c_5}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2+\frac{c_5L\sqrt{m}}{4}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2,\tag{C.14}$$

where we have used Assumption 3.3 in (C.13) and Assumption 3.2 in (C.14). Applying (C.12)-(C.14) to (C.11) results in

$$\begin{aligned}\left(\tfrac{1}{\eta_{\mathbf{y}}}-\tfrac{L}{2}-\tfrac{1}{4\eta_{\mathbf{y}}}\right)\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2\leq&\,m(\hat{\delta}_t-\hat{\delta}_{t+1})+\eta_{\mathbf{y}}\left\|\mathbf{R}_{\mathbf{y}}^{(t)}\right\|_F^2+\left(\tfrac{L^2}{2c_4}+\tfrac{L\sqrt{m}}{c_5}\right)\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2\\ &+\left(\tfrac{L_Q+L}{2}+\tfrac{c_4}{2}\right)\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2+\left(\tfrac{mL_Q}{2}+\tfrac{c_5L\sqrt{m}}{4}\right)\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned}\tag{C.15}$$

By $\eta_{\mathbf{y}} \le \frac{1}{4L}$, it holds that $\frac{1}{4\eta_{\mathbf{y}}} \le \frac{1}{\eta_{\mathbf{y}}} - \frac{L}{2} - \frac{1}{4\eta_{\mathbf{y}}}$ and hence

$$\begin{aligned}\left\|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\right\|_F^2 \le 4m\eta_{\mathbf{y}}\left(\hat{\delta}_t - \hat{\delta}_{t+1}\right) + 4\eta_{\mathbf{y}}^2\left\|\mathbf{R}_{\mathbf{y}}^{(t)}\right\|_F^2 + 4\eta_{\mathbf{y}}\left(\tfrac{L^2}{2c_4} + \tfrac{L\sqrt{m}}{c_5}\right)\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \mathbf{Y}^{(t+1)}\right\|_F^2 \\ +4\eta_{\mathbf{y}}\left(\tfrac{L_Q+L}{2} + \tfrac{c_4}{2}\right)\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F^2 + 4\eta_{\mathbf{y}}\left(\tfrac{mL_Q}{2} + \tfrac{c_5 L\sqrt{m}}{4}\right)\left\|\boldsymbol{\Lambda}^{(t+1)} - \boldsymbol{\Lambda}^{(t)}\right\|_F^2.\end{aligned} \tag{C.16}$$

Applying $\left\|\mathbf{R}_{\mathbf{y}}^{(t)}\right\|_F^2 \le \left\|\mathbf{R}^{(t)}\right\|_F^2$ completes the proof. □

Proof. [Of Lemma 3.6] By the definition of $\widetilde{\mathbf{Y}}^{(t)}$ in (3.3), it holds that

$$\mathbf{0} \in \frac{1}{m}\left[\nabla_{\mathbf{y}} f_1(\mathbf{x}_1^{(t)}, \tilde{\mathbf{y}}_1^{(t)}) - \partial h(\tilde{\mathbf{y}}_1^{(t)}), \ldots, \nabla_{\mathbf{y}} f_m(\mathbf{x}_m^{(t)}, \tilde{\mathbf{y}}_m^{(t)}) - \partial h(\tilde{\mathbf{y}}_m^{(t)})\right]^\top - \frac{L}{2\sqrt{m}}\left(\mathbf{W} - \mathbf{I}\right)^\top \boldsymbol{\Lambda}^{(t)}. \tag{C.17}$$

Defining $\tilde{\boldsymbol{\Lambda}}^{(t)} := (\mathbf{W} - \mathbf{I})^\top \boldsymbol{\Lambda}^{(t)}$, we have for all $i = 1, \ldots, m$,

$$\tilde{\mathbf{y}}_i^{(t)} = \arg\max_{\mathbf{y}_i}\left\{\left\langle \mathbf{y}_i, \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \frac{L\sqrt{m}}{2}\tilde{\boldsymbol{\lambda}}_i^{(t)}\right\rangle - \frac{1}{2\eta_{\mathbf{y}}}\left\|\mathbf{y}_i - \tilde{\mathbf{y}}_i^{(t)}\right\|_2^2 - h(\mathbf{y}_i)\right\} \tag{C.18}$$

where $(\tilde{\boldsymbol{\lambda}}_i^{(t)})^\top$ denotes the i-th row of $\tilde{\boldsymbol{\Lambda}}^{(t)}$. Hence,

$$\tilde{\mathbf{y}}_i^{(t)} = \mathbf{prox}_{\eta_{\mathbf{y}} h}\left(\tilde{\mathbf{y}}_i^{(t)} + \eta_{\mathbf{y}}\left(\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \frac{L\sqrt{m}}{2}\tilde{\boldsymbol{\lambda}}_i^{(t)}\right)\right). \tag{C.19}$$

By the non-expansiveness of the proximal operator,

$$\begin{aligned}\left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t+1)}\right\|_2^2 &= \left\|\mathbf{prox}_{\eta_{\mathbf{y}} h}\left(\tilde{\mathbf{y}}_i^{(t)} + \eta_{\mathbf{y}}\left(\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \frac{L\sqrt{m}}{2}\tilde{\boldsymbol{\lambda}}_i^{(t)}\right)\right) - \mathbf{prox}_{\eta_{\mathbf{y}} h}\left(\mathbf{y}_i^{(t)} + \eta_{\mathbf{y}}\mathbf{v}_{\mathbf{y},i}^{(t)}\right)\right\|_2^2 \\ &\le \left\|\tilde{\mathbf{y}}_i^{(t)} + \eta_{\mathbf{y}}\left(\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \frac{L\sqrt{m}}{2}\tilde{\boldsymbol{\lambda}}_i^{(t)}\right) - \mathbf{y}_i^{(t)} - \eta_{\mathbf{y}}\mathbf{v}_{\mathbf{y},i}^{(t)}\right\|_2^2.\end{aligned} \tag{C.20}$$

Utilizing the $\mathbf{V}_{\mathbf{y}}$-update (2.6) in (C.20) and the Peter-Paul inequality we obtain

$$\begin{aligned}\left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t+1)}\right\|_F^2 \le &(1+b)\left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)} + \eta_{\mathbf{y}}\left(\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)})\right)\right\|_2^2 \\ &+ (1 + \tfrac{1}{b})\eta_{\mathbf{y}}^2\left\|\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}) - \mathbf{d}_{\mathbf{y},i}^{(t)}\right\|_2^2\end{aligned} \tag{C.21}$$

where $b > 0$ is an arbitrary constant. By the L-smoothness and μ-strong convexity of $-f_i(\mathbf{x}, \cdot)$, it holds for all $i \in [m]$ that

$$\begin{aligned}&\left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)} + \eta_{\mathbf{y}}\left(\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)})\right)\right\|_2^2 \\ &= \left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)}\right\|_2^2 + 2\eta_{\mathbf{y}}\left\langle \tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)}, \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)})\right\rangle + \eta_{\mathbf{y}}^2\left\|\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)})\right\|_2^2 \\ &\le \left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)}\right\|_2^2 + \left(2\eta_{\mathbf{y}} - \eta_{\mathbf{y}}^2 L\right)\left\langle \tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)}, \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)})\right\rangle \\ &\le \left(1 - 2\eta_{\mathbf{y}}\mu + \eta_{\mathbf{y}}^2\mu L\right)\left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)}\right\|_2^2 \le \left(1 - \frac{7}{4}\eta_{\mathbf{y}}\mu\right)\left\|\tilde{\mathbf{y}}_i^{(t)} - \mathbf{y}_i^{(t)}\right\|_2^2\end{aligned} \tag{C.22}$$

where the second to last inequality uses the strong concavity of $f_i(\mathbf{x}, \cdot)$ and the last uses $\eta_{\mathbf{y}} \le \frac{1}{4L}$ to have $1 - 2\eta_{\mathbf{y}}\mu + \eta_{\mathbf{y}}^2\mu L \le 1 - \frac{7}{4}\eta_{\mathbf{y}}\mu$. Hence, by the Peter-Paul inequality and utilizing (C.22) within (C.21), we have

$$\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \mathbf{Y}^{(t+1)}\right\|_F^2 \le (1+a)\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t+1)}\right\|_F^2 + (1 + \tfrac{1}{a})\left\|\widetilde{\mathbf{Y}}^{(t+1)} - \widetilde{\mathbf{Y}}^{(t)}\right\|_F^2$$

$$
\begin{aligned}
&\leq (1+a)(1+b)\left(1-\tfrac{7}{4}\eta_{\mathbf{y}}\mu\right)\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2+(1+a)(1+\tfrac{1}{b})\eta_{\mathbf{y}}^2\left\|\mathbf{R}_{\mathbf{y}}^{(t)}\right\|_F^2+(1+\tfrac{1}{a})\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\widetilde{\mathbf{Y}}^{(t)}\right\|_F^2\\
&\overset{(3.10c)}{\leq} (1+a)(1+b)\left(1-\tfrac{7}{4}\eta_{\mathbf{y}}\mu\right)\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2+(1+a)(1+\tfrac{1}{b})\eta_{\mathbf{y}}^2\left\|\mathbf{R}_{\mathbf{y}}^{(t)}\right\|_F^2 \qquad \text{(C.23)}\\
&\quad+2(1+\tfrac{1}{a})\kappa^2\left(\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2+m\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\right)
\end{aligned}
$$

where $a>0$ is an arbitrary constant. Setting $a=\frac{\eta_{\mathbf{y}}\mu}{2-3\eta_{\mathbf{y}}\mu}=\frac{1-\eta_{\mathbf{y}}\mu}{1-\frac{3}{2}\eta_{\mathbf{y}}\mu}-1$ and $b=\frac{\frac{1}{4}\eta_{\mathbf{y}}\mu}{1-\frac{7}{4}\eta_{\mathbf{y}}\mu}=\frac{1-\frac{3}{2}\eta_{\mathbf{y}}\mu}{1-\frac{7}{4}\eta_{\mathbf{y}}\mu}-1$ in (C.23) and utilizing $\eta_{\mathbf{y}}\leq\frac{1}{4L}$, it holds that

$$(1+a)(1+b)\left(1-\frac{7}{4}\eta_{\mathbf{y}}\mu\right)=1-\eta_{\mathbf{y}}\mu, \qquad \text{(C.24)}$$

$$(1+a)(1+\frac{1}{b})\eta_{\mathbf{y}}^2=\left(\frac{1-\eta_{\mathbf{y}}\mu}{1-\frac{3}{2}\eta_{\mathbf{y}}\mu}\right)\left(\frac{1-\frac{3}{2}\eta_{\mathbf{y}}\mu}{\frac{1}{4}\eta_{\mathbf{y}}\mu}\right)\eta_{\mathbf{y}}^2=\frac{4\eta_{\mathbf{y}}-4\eta_{\mathbf{y}}^2\mu}{\mu}\leq\frac{4\eta_{\mathbf{y}}}{\mu}, \qquad \text{(C.25)}$$

$$1+\frac{1}{a}=\frac{2(1-\eta_{\mathbf{y}}\mu)}{\eta_{\mathbf{y}}\mu}\leq\frac{2}{\eta_{\mathbf{y}}\mu}. \qquad \text{(C.26)}$$

Applying the bounds in (C.24)-(C.26) to (C.23) and utilizing $\left\|\mathbf{R}_{\mathbf{y}}^{(t)}\right\|_F^2\leq\left\|\mathbf{R}^{(t)}\right\|_F^2$ completes the proof. □

D Proofs of Lemmas and Corollary in Section 3.4

Proof. [Of Lemma 3.7] We first prove (3.24), for which we break the proof into two cases. For the first case, we assume $t=n_tq$, i.e., t is divisible by q. Then for all $i\in[m]$, we have by the definition of Υ that

$$\mathbb{E}\left\|\mathbf{r}_i^{(t)}\right\|_2^2=\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)})-\mathbf{d}_i^{(t)}\right\|_2^2\overset{(2.3)}{=}\mathbb{E}\left[\mathbb{E}_{\tilde{\mathcal{B}}_i^{(t)}}\left\|\nabla f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)})-G_i^{(t)}(\tilde{\mathcal{B}}_i^{(t)})\right\|_2^2\right]\leq\Upsilon \qquad \text{(D.1)}$$

where the second equality uses the law of total expectation and the inequality uses Assumption 3.3(ii) and $|\tilde{\mathcal{B}}_i^{(t)}|=\mathcal{S}_1$. Next, we assume $n_tq<t<(n_t+1)q$ and compute

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{r}_i^{(t)}\right\|_2^2=&\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)})-\mathbf{d}_i^{(t)}\right\|_2^2\overset{(2.3)}{=}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(t)},\mathbf{y}_i^{(t)})-G_i^{(t)}(\mathcal{B}_i^{(t)})+G_i^{(t-1)}(\mathcal{B}_i^{(t)})-\mathbf{d}_i^{(t-1)}\right\|_2^2\\
\leq&\frac{L^2}{\mathcal{S}_2}\left(\mathbb{E}\left\|\mathbf{x}_i^{(t)}-\mathbf{x}_i^{(t-1)}\right\|_2^2+\mathbb{E}\left\|\mathbf{y}_i^{(t)}-\mathbf{y}_i^{(t-1)}\right\|_2^2\right)+\mathbb{E}\left\|\mathbf{r}_i^{(t-1)}\right\|_2^2,
\end{aligned} \qquad \text{(D.2)}
$$

where the inequality comes from Eqn. (A.4) in [8]. Recursively utilizing (D.2) for $n_tq<t<(n_t+1)q$ and using (D.1) with $t=n_tq$ yields

$$\mathbb{E}\left\|\mathbf{r}_i^{(t)}\right\|_2^2\leq\frac{L^2}{\mathcal{S}_2}\sum_{r=n_tq}^{t-1}\left(\mathbb{E}\left\|\mathbf{x}_i^{(r+1)}-\mathbf{x}_i^{(r)}\right\|_2^2+\mathbb{E}\left\|\mathbf{y}_i^{(r+1)}-\mathbf{y}_i^{(r)}\right\|_2^2\right)+\Upsilon. \qquad \text{(D.3)}$$

Summing up (D.3) over $i=1,\ldots,m$ proves (3.24).

For (3.23), we utilize the technique from Lemma C.7 in [28] to have

$$\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2\leq\rho\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2+\frac{1}{1-\rho}\left\|\mathbf{D}_{\mathbf{x}}^{(t+1)}-\mathbf{D}_{\mathbf{x}}^{(t)}\right\|_F^2. \qquad \text{(D.4)}$$

By $\left\|\mathbf{d}_{\mathbf{x},i}^{(t+1)}-\mathbf{d}_{\mathbf{x},i}^{(t)}\right\|_2^2\leq\left\|\mathbf{d}_i^{(t+1)}-\mathbf{d}_i^{(t)}\right\|_2^2,\forall\, i\in[m]$, we use Young's inequality to have

$$\left\|\mathbf{D}_{\mathbf{x}}^{(t+1)}-\mathbf{D}_{\mathbf{x}}^{(t)}\right\|_F^2\leq\left\|\mathbf{D}^{(t+1)}-\mathbf{D}^{(t)}\right\|_F^2$$

$$\begin{aligned}
&\leq 3\left\|\mathbf{D}^{(t+1)}-\nabla F(\mathbf{Z}^{(t+1)})\right\|_F^2+3\left\|\nabla F(\mathbf{Z}^{(t+1)})-\nabla F(\mathbf{Z}^{(t)})\right\|_F^2+3\left\|\nabla F(\mathbf{Z}^{(t)})-\mathbf{D}^{(t)}\right\|_F^2\\
&\leq 3\left\|\mathbf{R}^{(t+1)}\right\|_F^2+3L^2\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2+3\left\|\mathbf{R}^{(t)}\right\|_F^2,
\end{aligned}\tag{D.5}$$

where the last inequality further uses Assumption 3.3. Take the expectation of (D.5) and utilize (3.24) to have

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{D}_{\mathbf{x}}^{(t+1)}-\mathbf{D}_{\mathbf{x}}^{(t)}\right\|_F^2\\
\leq& 3L^2\mathbb{E}\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2+\frac{3L^2}{\mathcal{S}_2}\sum_{r=(n_{t+1})q}^{t}\mathbb{E}\left\|\mathbf{Z}^{(r+1)}-\mathbf{Z}^{(r)}\right\|_F^2+3m\Upsilon+\frac{3L^2}{\mathcal{S}_2}\sum_{r=n_tq}^{t-1}\mathbb{E}\left\|\mathbf{Z}^{(r+1)}-\mathbf{Z}^{(r)}\right\|_F^2+3m\Upsilon.
\end{aligned}\tag{D.6}$$

Applying (D.6) to the expectation of (D.4) and further using the non-negativity of the 2-norm completes the proof. □

Proof. [Of Lemma 3.8] With constants in (3.27) and $L_Q=L\sqrt{4\kappa^2+1}\leq L(2\kappa+1)$, (3.20) implies

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\right\|_F^2&\leq\mathbb{E}\left[\frac{m}{L}(\hat{\delta}_0-\hat{\delta}_T)\right]+\frac{1}{4L^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2+\frac{1}{16\kappa^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2\\
&\quad+(8\kappa^2+\kappa+1)\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2+m(8\kappa^2+\kappa+1)\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\\
&\overset{(3.24),(3.26)}{\leq}\frac{m}{L}\hat{\delta}_0+\frac{1}{16\kappa^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2+(8\kappa^2+\kappa+1)\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2\\
&\quad+m(8\kappa^2+\kappa+1)\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2+\frac{1}{4}\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2+\frac{m\Upsilon T}{4L^2},
\end{aligned}$$

where we have used the fact $\hat{\delta}_T\geq 0$. Since $\mathbf{Z}^{(t)}=(\mathbf{X}^{(t)},\mathbf{Y}^{(t)})$, the inequality above clearly indicates

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\right\|_F^2&\leq\left(\frac{4}{3}(8\kappa^2+\kappa+1)+\frac{1}{3}\right)\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2\\
&\quad+\frac{1}{12\kappa^2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2+\frac{4}{3}m(8\kappa^2+\kappa+1)\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2+\frac{4m}{3L}\hat{\delta}_0+\frac{m\Upsilon T}{3L^2}.
\end{aligned}\tag{D.7}$$

On the other hand, summing up (3.22) and using (3.24) and (3.26), we have

$$\begin{aligned}
\frac{1}{4\kappa}\sum_{t=0}^{T-1}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2&\leq\left(1-\frac{1}{4\kappa}\right)\left\|\widetilde{\mathbf{Y}}^{(0)}-\mathbf{Y}^{(0)}\right\|_F^2+\frac{Tm\Upsilon}{L\mu}\\
&\quad+(16\kappa^3+\kappa)\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2+16m\kappa^3\sum_{t=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2+\kappa\sum_{t=0}^{T-1}\mathbb{E}\left\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\right\|_F^2.
\end{aligned}\tag{D.8}$$

Utilize (D.7) within (D.8), solve it for $\sum_{t=0}^{T-1}\mathbb{E}\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\|_F^2$, and bound $\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2$ by (3.7). We obtain (3.28). Now substitute (3.28) into (D.7), use (3.7) again, and combine like terms to have (3.29) and complete the proof. □

Proof. [Of Corollary 1] By the choice of initialization and the definition of C_0 from Theorem 3.1, we have

$$\tfrac{L^2(6\kappa+3)}{mT}\left\|\widetilde{\mathbf{Y}}^{(0)}-\mathbf{Y}^{(0)}\right\|_F^2=\mathcal{O}\left(\tfrac{L\kappa^2}{T}\right),\quad C_0=\mathcal{O}\left(\phi(\bar{\mathbf{x}}^{(0)},\boldsymbol{\Lambda}^{(0)})-\phi^*+\tfrac{L\hat{\delta}_0}{L_P(1-\rho)^2}+1\right).$$

Additionally by (3.30) and $L_P = L\sqrt{4\kappa^2+1}$, the stepsizes in (3.30) are $\eta_{\mathbf{x}} = \Theta\left(\frac{\min\{1,\kappa(1-\rho)^2\}}{L\kappa^2}\right)$ and $\eta_{\mathbf{\Lambda}} = \Theta\left(\frac{\min\{1,\kappa(1-\rho)^2\}}{\kappa^2 L}\right)$. Thus we have from (3.45) that

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{\eta_{\mathbf{x}}}\left(\bar{\mathbf{x}}^{(\tau)} - \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\bar{\mathbf{x}}^{(\tau)} - \eta_{\mathbf{x}}\nabla_{\mathbf{x}} P(\bar{\mathbf{x}}^{(\tau)}, \mathbf{\Lambda}^{(\tau)})\right)\right)\right\|_2^2 + \frac{L^2}{m}\mathbb{E}\left\|\mathbf{X}_\perp^{(\tau)}\right\|_F^2 \\
=&\mathcal{O}\left(\kappa^2\Upsilon + \frac{L\kappa^2\hat{\delta}_0}{T} + \left(\frac{1}{T}\left(\phi(\bar{\mathbf{x}}^{(0)}, \mathbf{\Lambda}^{(0)}) - \phi^* + \tfrac{L\hat{\delta}_0}{L_P(1-\rho)^2} + 1\right) + \frac{\Upsilon}{L\cdot\min\{1,\kappa(1-\rho)^2\}}\right)\cdot\frac{L\kappa^2}{\min\{1,\kappa(1-\rho)^2\}}\right) \\
=&\mathcal{O}\left(\left(\frac{1}{T}\left(\phi(\bar{\mathbf{x}}^{(0)}, \mathbf{\Lambda}^{(0)}) - \phi^* + \tfrac{\hat{\delta}_0}{\min\{1,\kappa(1-\rho)^2\}} + 1\right) + \frac{\Upsilon}{L\cdot\min\{1,\kappa(1-\rho)^2\}}\right)\cdot\frac{L\kappa^2}{\min\{1,\kappa(1-\rho)^2\}}\right)
\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathbb{E}\left\|\nabla_{\mathbf{\Lambda}} P(\bar{\mathbf{x}}^{(\tau)}, \mathbf{\Lambda}^{(\tau)})\right\|_F^2 \\
=&\mathcal{O}\left(\left(\frac{1}{T}\left(\phi(\bar{\mathbf{x}}^{(0)}, \mathbf{\Lambda}^{(0)}) - \phi^* + \tfrac{\hat{\delta}_0}{\min\{1,\kappa(1-\rho)^2\}} + 1\right) + \frac{\Upsilon}{L\cdot\min\{1,\kappa(1-\rho)^2\}}\right)\cdot\frac{L\kappa^2}{\min\{1,\kappa(1-\rho)^2\}}\right).
\end{aligned}$$

Thus for the finite-sum setting, since $\Upsilon = 0$, we can produce an ε-stationary point as defined in Definition 3.1 by T iterations, with

$$T = \Theta\left(\frac{L\kappa^2}{\varepsilon^2\cdot\min\{1,\kappa(1-\rho)^2\}}\left(\phi(\bar{\mathbf{x}}^{(0)}, \mathbf{\Lambda}^{(0)}) - \phi^* + \frac{\hat{\delta}_0}{\min\{1,\kappa(1-\rho)^2\}} + 1\right)\right). \tag{D.9}$$

For the general case, $\Upsilon = \frac{\sigma^2}{\mathcal{S}_1}$, and an ε-stationary point can be produced with $\mathcal{S}_1 = \Theta\left(\frac{\sigma^2}{\varepsilon^2}\cdot\max\{\kappa^2, (1-\rho)^{-4}\}\right)$ and T in the same order as that in (D.9). For both general case and finite-sum case, we set $\mathcal{S}_2 = q = \lceil\sqrt{\mathcal{S}_1}\rceil$. Noticing the total number of communication is $T_c = T$ and the total number of sample gradients $T_s = (T - \lfloor\frac{T}{q}\rfloor)\mathcal{S}_2 + \lceil\frac{T}{q}\rceil\mathcal{S}_1$, we complete the proof. □

E Proofs of Lemmas and Corollary in Section 3.5

Proof. [Of Lemma 3.9] For (3.51), we use (D.4). When `VR`-tag == STORM in Algorithm 1, it holds

$$\begin{aligned}
&\left\|\mathbf{d}_{\mathbf{x},i}^{(t+1)} - \mathbf{d}_{\mathbf{x},i}^{(t)}\right\|_2^2 \le \left\|\mathbf{d}_i^{(t+1)} - \mathbf{d}_i^{(t)}\right\|_2^2 \\
=&\left\|G_i^{t+1}(\mathcal{B}_i^{(t+1)}) + (1-\beta)\left(\mathbf{d}_i^{(t)} - G_i^{(t)}(\mathcal{B}_i^{(t+1)})\right) - \mathbf{d}_i^{(t)}\right\|_2^2 \\
=&\left\|G_i^{t+1}(\mathcal{B}_i^{(t+1)}) - G_i^{(t)}(\mathcal{B}_i^{(t+1)}) + \beta\left(G_i^{(t)}(\mathcal{B}_i^{(t+1)}) - \nabla f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)})\right) + \beta\left(\nabla f_i(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}) - \mathbf{d}_i^{(t)}\right)\right\|_2^2.
\end{aligned} \tag{E.1}$$

Using Young's inequality and taking the expectation with respect to the samples $\mathcal{B}_i^{(t+1)}$ and then the full expectation yields

$$\mathbb{E}\left\|\mathbf{d}_{\mathbf{x},i}^{(t+1)} - \mathbf{d}_{\mathbf{x},i}^{(t)}\right\|_2^2 \le 3L^2\mathbb{E}\left\|\mathbf{x}_i^{(t+1)} - \mathbf{x}_i^{(t)}\right\|_2^2 + 3L^2\mathbb{E}\left\|\mathbf{y}_i^{(t+1)} - \mathbf{y}_i^{(t)}\right\|_2^2 + 3\beta^2\Upsilon_{t+1} + 3\beta^2\mathbb{E}\left\|\mathbf{r}_i^{(t)}\right\|_2^2, \tag{E.2}$$

where $(\mathbf{r}_i^{(t)})^\top$ is defined as the i-th row of $\mathbf{R}^{(t)}$ for all $t \ge 0$ and we have additionally used Assumption 3.3(ii). Taking the full expectation of (D.4) and applying (E.1) with (E.2) summed over $i = 1, \ldots, m$ yields (3.51). The proof of (3.52) follows from the same arguments of the proof of [28, Lemma C.9]. □

Proof. [Of Lemma 3.10] We first take the expectation of (3.16), apply (3.19), and plug in the values of c_1, c_2, and c_3 specified in (3.53) to have

$$\begin{aligned}
\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(t+1)}, \mathbf{\Lambda}^{(t+1)}) - \phi(\bar{\mathbf{x}}^{(t)}, \mathbf{\Lambda}^{(t)})\right] \le& -\frac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 - \frac{1}{2m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)} - \widetilde{\mathbf{X}}^{(t)}\right\|_F^2 \\
&- \left(\frac{1}{\eta_{\mathbf{\Lambda}}} - \frac{L_P}{2} - \frac{32\kappa^2 L}{\sqrt{\beta}}\right)\mathbb{E}\left\|\mathbf{\Lambda}^{(t+1)} - \mathbf{\Lambda}^{(t)}\right\|_F^2 + \frac{1}{2m}\left(L(\kappa+1) + \frac{\sqrt{\beta}}{32} + \frac{\rho^2}{\eta_{\mathbf{x}}}\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2 \\
&+ \frac{\sqrt{\beta}(L+1)}{64m\kappa^2}\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)} - \mathbf{Y}^{(t)}\right\|_F^2 + \frac{\sqrt{\beta}}{60mL_P}\mathbb{E}\left(\left\|\mathbf{R}^{(t)}\right\|_F^2 + \left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\right),
\end{aligned} \tag{E.3}$$

where the coefficient of $\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2$ is obtained by using $L_P = L\sqrt{4\kappa^2+1} \le L(2\kappa+1)$ and $\sqrt{\beta}<1$ to have

$$L\left(\kappa+1+\frac{32\kappa^2}{\sqrt{\beta}}\right)+L_P\left(1+\frac{60}{\sqrt{\beta}}\right) \le \frac{32\kappa^2 L+123\kappa L+62L}{\sqrt{\beta}} \le \frac{1}{2\eta_{\mathbf{x}}}.$$

Next, for any $\gamma_1,\gamma_2,\gamma_3,\gamma_4>0$, we add $\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2, \gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2, \gamma_3\mathbb{E}\left\|\mathbf{R}^{(t+1)}\right\|_F^2, \gamma_4\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2$ to both sides of (E.3) and upper bound the right-hand side using Lemmas 3.3, 3.6, and 3.9 to obtain

$$\begin{aligned}
&\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\phi(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right]+\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2+\gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2+\gamma_3\mathbb{E}\left\|\mathbf{R}^{(t+1)}\right\|_F^2+\gamma_4\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2\\
\le&-\frac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2-\frac{1}{2m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2+\left(\frac{3\gamma_2}{1-\rho}+2\gamma_3\right)m\beta^2\Upsilon_{t+1}\\
&-\left(\frac{1}{\eta_{\boldsymbol{\Lambda}}}-\frac{L_P}{2}-\frac{32\kappa^2L}{\sqrt{\beta}}-\frac{16\sqrt{2}\kappa^3m\gamma_4}{\sqrt{\beta}}\right)\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\\
&+\frac{1}{2m}\left(L(\kappa+1)+\frac{\sqrt{\beta}}{32}+\frac{\rho^2}{\eta_{\mathbf{x}}}+2m\rho\gamma_1\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\left(\rho\gamma_2+\frac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}+\frac{\sqrt{\beta}}{60mL_P}\right)\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\\
&+\left(\gamma_4\big(1-\frac{\sqrt{\beta}}{4\sqrt{2}\kappa}\big)+\frac{\sqrt{\beta}(L+1)}{64m\kappa^2}\right)\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2+\frac{16\sqrt{2}\kappa^3\gamma_4}{\sqrt{\beta}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2\\
&+\left(\frac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\mathbb{E}\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2+\left(\gamma_3(1-\beta)^2+\frac{3\beta^2\gamma_2}{1-\rho}+\frac{\sqrt{\beta}}{60mL_P}+\frac{\sqrt{\beta}\gamma_4}{\sqrt{2}L\mu}\right)\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2.
\end{aligned}\tag{E.4}$$

We apply $\left\|\mathbf{Z}^{(t+1)}-\mathbf{Z}^{(t)}\right\|_F^2=\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2+\left\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\right\|_F^2$ and (3.20) to (E.4) to have

$$\begin{aligned}
&\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\phi(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right]+\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2+\gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2\\
&+\gamma_3\mathbb{E}\left\|\mathbf{R}^{(t+1)}\right\|_F^2+\left(\gamma_4-\tfrac{\beta}{60\kappa^2}\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\right)\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2\\
\le&-\tfrac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2-\tfrac{1}{2m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2+\left(\tfrac{3\gamma_2}{1-\rho}+2\gamma_3\right)m\beta^2\Upsilon_{t+1}\\
&-\left(\tfrac{1}{\eta_{\boldsymbol{\Lambda}}}-\tfrac{L_P}{2}-\tfrac{32\kappa^2L}{\sqrt{\beta}}-\tfrac{16\sqrt{2}\kappa^3m\gamma_4}{\sqrt{\beta}}-\tfrac{m}{2}\left(\tfrac{\sqrt{\beta}}{\sqrt{2}}\sqrt{4\kappa^2+1}+30\kappa^2\right)\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\right)\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2\\
&+\tfrac{1}{2m}\left(L(\kappa+1)+\tfrac{\sqrt{\beta}}{32}+\tfrac{\rho^2}{\eta_{\mathbf{x}}}+2m\rho\gamma_1\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\left(\rho\gamma_2+\tfrac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}+\tfrac{\sqrt{\beta}}{60mL_P}\right)\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\\
&+\left(\gamma_4\big(1-\tfrac{\sqrt{\beta}}{4\sqrt{2}\kappa}\big)+\tfrac{\sqrt{\beta}(L+1)}{64m\kappa^2}\right)\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2+\tfrac{m\sqrt{\beta}}{\sqrt{2}L}\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\mathbb{E}\left[\hat{\delta}_t-\hat{\delta}_{t+1}\right]\\
&+\left(\tfrac{16\sqrt{2}\kappa^3\gamma_4}{\sqrt{\beta}}+\left(\tfrac{\sqrt{\beta}}{2\sqrt{2}}(\sqrt{4\kappa^2+1}+1)+15\kappa^2+1\right)\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\right)\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2\\
&+\left(\gamma_3(1-\beta)^2+\tfrac{3\beta^2\gamma_2}{1-\rho}+\tfrac{\sqrt{\beta}}{60mL_P}+\tfrac{\sqrt{\beta}\gamma_4}{\sqrt{2}L\mu}+\tfrac{\beta}{8L^2}\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\right)\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2.
\end{aligned}\tag{E.5}$$

Next, we utilize (3.7) to bound $\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\right\|_F^2$. Then we subtract $\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2, \gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2, \gamma_3\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2$ from both sides of (E.5) with

$$c_{\mathbf{x}}:=\tfrac{16\sqrt{2}\kappa^3\gamma_4}{\sqrt{\beta}}+\left(\tfrac{\sqrt{\beta}}{2\sqrt{2}}(\sqrt{4\kappa^2+1}+1)+15\kappa^2+1\right)\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\tag{E.6}$$

to have

$$\begin{aligned}
&\mathbb{E}\left[\phi(\bar{\mathbf{x}}^{(t+1)},\boldsymbol{\Lambda}^{(t+1)})-\phi(\bar{\mathbf{x}}^{(t)},\boldsymbol{\Lambda}^{(t)})\right]+\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t+1)}\right\|_F^2-\gamma_1\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2+\gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t+1)}\right\|_F^2-\gamma_2\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\\
&+\gamma_3\mathbb{E}\left\|\mathbf{R}^{(t+1)}\right\|_F^2-\gamma_3\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2+\left(\gamma_4-\tfrac{\beta}{60\kappa^2}\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\right)\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t+1)}-\mathbf{Y}^{(t+1)}\right\|_F^2\\
\le&-\tfrac{1}{4m\eta_{\mathbf{x}}}\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2-\left(\tfrac{1}{2m\eta_{\mathbf{x}}}-2c_{\mathbf{x}}\right)\mathbb{E}\left\|\mathbf{X}^{(t+1)}-\widetilde{\mathbf{X}}^{(t)}\right\|_F^2+\left(\tfrac{3\gamma_2}{1-\rho}+2\gamma_3\right)m\beta^2\Upsilon_{t+1}\\
&-\left(\tfrac{1}{\eta_{\boldsymbol{\Lambda}}}-\tfrac{L_P}{2}-\tfrac{32\kappa^2L}{\sqrt{\beta}}-\tfrac{16\sqrt{2}\kappa^3m\gamma_4}{\sqrt{\beta}}-\tfrac{m}{2}\left(\tfrac{\sqrt{\beta}}{\sqrt{2}}\sqrt{4\kappa^2+1}+30\kappa^2\right)\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\right)\mathbb{E}\left\|\boldsymbol{\Lambda}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}\right\|_F^2 \qquad \text{(E.7)}\\
&-\tfrac{1}{2m}\left(2m(1-\rho)\gamma_1-L(\kappa+1)-\tfrac{\sqrt{\beta}}{32}-\tfrac{\rho^2}{\eta_{\mathbf{x}}}-16mc_{\mathbf{x}}\right)\mathbb{E}\left\|\mathbf{X}_\perp^{(t)}\right\|_F^2-\left((1-\rho)\gamma_2-\tfrac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}-\tfrac{\sqrt{\beta}}{60mL_P}\right)\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(t)}\right\|_F^2\\
&+\left(\gamma_4\left(1-\tfrac{\sqrt{\beta}}{4\sqrt{2}\kappa}\right)+\tfrac{\sqrt{\beta}(L+1)}{64m\kappa^2}\right)\mathbb{E}\left\|\widetilde{\mathbf{Y}}^{(t)}-\mathbf{Y}^{(t)}\right\|_F^2+\tfrac{m\sqrt{\beta}}{\sqrt{2}L}\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\mathbb{E}\left[\hat{\delta}_t-\hat{\delta}_{t+1}\right]\\
&-\left((1-(1-\beta)^2)\gamma_3-\tfrac{3\beta^2\gamma_2}{1-\rho}-\tfrac{\sqrt{\beta}}{60mL_P}-\tfrac{\sqrt{\beta}\gamma_4}{\sqrt{2}L\mu}-\tfrac{\beta}{8L^2}\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)\right)\mathbb{E}\left\|\mathbf{R}^{(t)}\right\|_F^2.
\end{aligned}$$

Set

$$\gamma_1=\frac{2}{m(1-\rho)}\left((24\kappa^2+7\kappa+4)\left(\frac{16(L+1)}{\sqrt{\beta}}+\frac{8L}{\kappa(1-\rho)^2}\right)+L(\kappa+1)+\frac{\sqrt{\beta}}{32}+\frac{1}{\eta_{\mathbf{x}}}\right), \qquad \text{(E.8a)}$$

$$\gamma_2=\frac{2}{1-\rho}\left(\frac{\sqrt{\beta}}{60mL_P}+\frac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho}\right), \qquad \text{(E.8b)}$$

$$\gamma_3=\frac{2}{\beta}\left(\frac{\sqrt{\beta}}{60mL_P}+\frac{\sqrt{\beta}\gamma_4}{\sqrt{2}L\mu}+\frac{3\gamma_2\beta(\beta+1/8)}{1-\rho}\right), \qquad \text{(E.8c)}$$

$$\gamma_4=\frac{\sqrt{2}(L+1)}{8m\kappa}+\frac{4L^2\sqrt{\beta}}{15\sqrt{2}\kappa}\left(\frac{9\gamma_2}{2(1-\rho)}+\frac{12\gamma_2\beta}{1-\rho}+\frac{1}{15\sqrt{\beta}mL_P}\right). \qquad \text{(E.8d)}$$

By the definition of $\eta_{\mathbf{x}}$ and $\beta<1$, we can easily have $(24\kappa^2+7\kappa+4)\left(\frac{16(L+1)}{\sqrt{\beta}}+\frac{8L}{\kappa(1-\rho)^2}\right)+L(\kappa+1)+\frac{\sqrt{\beta}}{32}\le\frac{1}{\eta_{\mathbf{x}}}$ and thus $\gamma_1\le\frac{4}{m(1-\rho)\eta_{\mathbf{x}}}$. Then by the choice of γ_2, $L_P\ge2\kappa L$, and $\beta<1$, it follows that

$$\gamma_2\le\frac{1}{60m(1-\rho)\kappa L}+\frac{8\eta_{\mathbf{x}}}{m(1-\rho)^3}\le\frac{1}{40m(1-\rho)\kappa L}, \qquad \text{(E.9)}$$

where the last inequality uses $\eta_{\mathbf{x}}\le\frac{(1-\rho)^2}{960\kappa L}$. Hence by the choice of γ_4, $L_P\ge2\kappa L$, $\beta<1$, and (E.9), we additionally have

$$\gamma_4\le\frac{L+1}{5m\kappa}+\frac{\sqrt{\beta}L}{10m\kappa^2(1-\rho)^2}. \qquad \text{(E.10)}$$

Moreover, applying $\beta<1$, $L_P\ge2\kappa L$, (E.9), and (E.10) to (E.8c) yields $\gamma_3\le\frac{1+1/L}{2\sqrt{\beta}mL}+\frac{2}{5m\kappa L(1-\rho)^2}$. Thus by the upper bounds of γ_2 and γ_3 together with $1-\beta<1$ yields

$$\frac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\le\frac{L}{m(1-\rho)^2\kappa}+\frac{L+1}{\sqrt{\beta}m}. \qquad \text{(E.11)}$$

Hence, from the definition of $c_{\mathbf{x}}$ in (E.6) and (E.10), (E.11), it follows

$$\begin{aligned}
c_{\mathbf{x}}&\le\tfrac{16\sqrt{2}\kappa^3}{\sqrt{\beta}}\left(\tfrac{L+1}{5m\kappa}+\tfrac{\sqrt{\beta}L}{10m\kappa^2(1-\rho)^2}\right)+\left(\tfrac{\sqrt{\beta}}{2\sqrt{2}}(\sqrt{4\kappa^2+1}+1)+15\kappa^2+1\right)\left(\tfrac{L}{m(1-\rho)^2\kappa}+\tfrac{L+1}{\sqrt{\beta}m}\right)\\
&\le\tfrac{L+1}{m\sqrt{\beta}}\left(\tfrac{16\sqrt{2}\kappa^2}{5}+15\kappa^2+1+\tfrac{\kappa+1}{\sqrt{2}}\right)+\tfrac{L}{m(1-\rho)^2\kappa}\left(\tfrac{16\sqrt{2}\kappa^2}{10}+15\kappa^2+1+\tfrac{\kappa+1}{\sqrt{2}}\right)\le\tfrac{1}{8m\eta_{\mathbf{x}}}, \qquad \text{(E.12)}
\end{aligned}$$

where the last inequality follows from the definition of $\eta_{\mathbf{x}}$. In addition, by $\beta<1$, $\sqrt{4\kappa^2+1}\le2\kappa+1$, (E.10), and (E.11), we have

$$\tfrac{16\sqrt{2}\kappa^3m\gamma_4}{\sqrt{\beta}}+\tfrac{m}{2}\left(\tfrac{\sqrt{\beta}}{\sqrt{2}}\sqrt{4\kappa^2+1}+30\kappa^2\right)\left(\tfrac{3L^2\gamma_2}{1-\rho}+2L^2(1-\beta)^2\gamma_3\right)$$

$$\leq \frac{16\sqrt{2}\kappa^2(L+1)}{5\sqrt{\beta}} + \frac{8\sqrt{2}\kappa L}{5(1-\rho)^2} + (15\kappa^2+\kappa+1)\left(\frac{L}{(1-\rho)^2\kappa} + \frac{L+1}{\sqrt{\beta}}\right) \leq \frac{(L+1)(20\kappa^2+\kappa+1)}{\sqrt{\beta}} + \frac{L(18\kappa+2)}{(1-\rho)^2}.$$

By the above equation and the choice of $\eta_{\boldsymbol{\Lambda}}$, we have

$$\frac{1}{\eta_{\boldsymbol{\Lambda}}} - \frac{L_P}{2} - \frac{32\kappa^2 L}{\sqrt{\beta}} - \frac{16\sqrt{2}\kappa^3 m\gamma_4}{\sqrt{\beta}} - \frac{m}{2}\left(\frac{\sqrt{\beta}}{\sqrt{2}}\sqrt{4\kappa^2+1} + 30\kappa^2\right)\left(\frac{3L^2\gamma_2}{1-\rho} + 2L^2(1-\beta)^2\gamma_3\right) \geq \frac{1}{2\eta_{\boldsymbol{\Lambda}}}. \tag{E.13}$$

On the other hand, from (E.12) and the definition of γ_1, it follows

$$\frac{1}{2m}\left(2m(1-\rho)\gamma_1 - L(\kappa+1) - \frac{\sqrt{\beta}}{32} - \frac{\rho^2}{\eta_{\mathbf{x}}} - 16mc_{\mathbf{x}}\right) \geq \frac{1}{2m}\left(2m(1-\rho)\gamma_1 - \frac{m(1-\rho)\gamma_1}{2} - \frac{2}{\eta_{\mathbf{x}}}\right) \geq \frac{1}{6m\eta_{\mathbf{x}}}. \tag{E.14}$$

Additionally, by the definition of γ_2 and $\beta < 1$, (3.39) holds, and thus

$$\frac{(1-\rho)\gamma_2}{2} \geq \frac{\gamma_1\eta_{\mathbf{x}}^2}{1-\rho} \geq \frac{\eta_{\mathbf{x}}}{m(1-\rho)^2}. \tag{E.15}$$

Also, by the choice of γ_3 and γ_4 and $(1-\beta)^2 \leq 1$, it holds that

$$\begin{aligned}
&\gamma_4 - \frac{\beta}{60\kappa^2}\left(\frac{3L^2\gamma_2}{1-\rho} + 2L^2(1-\beta)^2\gamma_3\right) - \gamma_4\big(1 - \frac{\sqrt{\beta}}{4\sqrt{2}\kappa}\big) - \frac{\sqrt{\beta}(L+1)}{64m\kappa^2} \\
\geq& \frac{\sqrt{\beta}}{4\sqrt{2}\kappa}\gamma_4 - \frac{2\beta L^2}{60\kappa^2}\frac{2}{\beta}\left(\frac{\sqrt{\beta}}{60mL_P} + \frac{\sqrt{\beta}\gamma_4}{\sqrt{2}L\mu} + \frac{3\gamma_2\beta(\beta+1/8)}{1-\rho}\right) - \frac{3\beta L^2\gamma_2}{60\kappa^2(1-\rho)} - \frac{\sqrt{\beta}(L+1)}{64m\kappa^2} \\
=& \frac{11\sqrt{\beta}}{60\sqrt{2}\kappa}\left(\frac{\sqrt{2}(L+1)}{8m\kappa} + \frac{4L^2\sqrt{\beta}}{15\sqrt{2}\kappa}\left(\frac{9\gamma_2}{2(1-\rho)} + \frac{12\gamma_2\beta}{1-\rho} + \frac{1}{15\sqrt{\beta}mL_P}\right)\right) \\
&- \frac{2\beta L^2}{60\kappa^2}\frac{2}{\beta}\left(\frac{\sqrt{\beta}}{60mL_P} + \frac{3\gamma_2\beta(\beta+1/8)}{1-\rho}\right) - \frac{3\beta L^2\gamma_2}{60\kappa^2(1-\rho)} - \frac{\sqrt{\beta}(L+1)}{64m\kappa^2} \\
\geq& \frac{\sqrt{\beta}(L+1)}{160m\kappa^2}.
\end{aligned} \tag{E.16}$$

Furthermore, from $1-(1-\beta)^2 \geq \beta$ and the formula of γ_3, it holds that

$$\begin{aligned}
&(1-(1-\beta)^2)\gamma_3 - \frac{3\beta^2\gamma_2}{1-\rho} - \frac{\sqrt{\beta}}{60mL_P} - \frac{\sqrt{\beta}\gamma_4}{\sqrt{2}L\mu} - \frac{\beta}{8L^2}\left(\frac{3L^2\gamma_2}{1-\rho} + 2L^2(1-\beta)^2\gamma_3\right) \\
\geq& \frac{\sqrt{\beta}}{60mL_P} + \frac{\sqrt{\beta}\gamma_4}{\sqrt{2}L\mu} + \frac{6\gamma_2\beta(\beta+1/8)}{1-\rho} - \frac{\beta\gamma_3}{4} - \frac{3\gamma_2\beta(\beta+1/8)}{1-\rho} \\
=& \frac{\beta\gamma_3}{4} \geq \frac{\sqrt{\beta}\gamma_4}{2\sqrt{2}L\mu} \geq \frac{\sqrt{\beta}(L+1)}{16m\kappa L\mu} \geq \frac{\sqrt{\beta}}{16mL}.
\end{aligned} \tag{E.17}$$

Applying (E.13)-(E.17) to (E.7), summing over $t = 0$ to $T-1$, and finally using $\hat{\delta}_T \geq 0$ and the upper bounds on $\gamma_2, \gamma_3, \gamma_4$ completes the proof. □

Proof. [Of Corollary 2] By the choice of β in (3.59) and the upper bound of $\varepsilon \leq \sigma(1-\rho)^2$, it holds that $\sqrt{\beta} \leq (1-\rho)^2$. Thus a straight-forward comparison of the two fractions in (3.54a) and (3.54b) yields

$$\eta_{\mathbf{x}} = \Theta\left(\frac{\sqrt{\beta}}{\kappa^2 L}\right) = \Theta\left(\frac{\varepsilon}{\sigma\kappa^3 L}\right), \quad \eta_{\boldsymbol{\Lambda}} = \Theta\left(\frac{\varepsilon}{\sigma\kappa^3 L}\right).$$

Next we upper bound the right-hand side of (3.57). First, by the initialization assumption and $\sqrt{\beta} \leq (1-\rho)^2$, we have $\frac{1}{40m(1-\rho)\kappa L}\mathbb{E}\left\|\mathbf{V}_{\perp,\mathbf{x}}^{(0)}\right\|_F^2 \leq 1$, $\left(\frac{L+1}{5m\kappa} + \frac{\sqrt{\beta}L}{10m\kappa^2(1-\rho)^2}\right)\left\|\widetilde{\mathbf{Y}}^{(0)} - \mathbf{Y}^{(0)}\right\|_F^2 = \mathcal{O}(1)$, and

$$\left(\frac{1}{\sqrt{\beta}mL} + \frac{1}{4m(1-\rho)^2\kappa L}\right)\mathbb{E}\left\|\mathbf{R}^{(0)}\right\|_F^2 \leq \left(\frac{1}{\sqrt{\beta}mL} + \frac{1}{4m(1-\rho)^2\kappa L}\right)\frac{m\sigma^2}{\mathcal{S}_0} \leq 1,$$

where we have used the definition of $\mathcal{S}_0$. Second, by $\sqrt{\beta} \leq \frac{(1-\rho)^2}{8\sqrt{6}}$ and $L \geq 1$, it holds that $C_0 = \Theta\left(\phi(\bar{\mathbf{x}}^{(0)}, \boldsymbol{\Lambda}^{(0)}) - \phi^* - \hat{\delta}_0 + 1\right)$. Also by the choice of $\mathcal{S}_t = \mathcal{O}(1)$ for all $t \geq 1$, it holds that $\Upsilon_t = \mathcal{O}\left(\sigma^2\right)$ for all $t \geq 1$. Hence by (3.57), we have

$$\begin{aligned} &\mathbb{E}\left\|\frac{1}{\eta_{\mathbf{x}}}\left(\bar{\mathbf{x}}^{(\tau)} - \mathbf{prox}_{\eta_{\mathbf{x}} g}\left(\bar{\mathbf{x}}^{(\tau)} - \eta_{\mathbf{x}} \nabla P(\bar{\mathbf{x}}^{(\tau)}, \boldsymbol{\Lambda}^{(\tau)})\right)\right)\right\|_2^2 + \frac{L^2}{m}\mathbb{E}\left\|\mathbf{X}_{\perp}^{(\tau)}\right\|_F^2 \\ =&\mathcal{O}\left(\left[\frac{\phi(\bar{\mathbf{x}}^{(0)}, \boldsymbol{\Lambda}^{(0)}) - \phi^* - \hat{\delta}_0 + 1}{T} + \frac{1}{\sqrt{\beta}L}\beta^2\sigma^2\right] \cdot \frac{\sigma\kappa^3 L}{\varepsilon}\right) \end{aligned} \tag{E.18}$$

and similarly

$$\mathbb{E}\left\|\nabla_{\boldsymbol{\Lambda}} P(\bar{\mathbf{x}}^{(\tau)}, \boldsymbol{\Lambda}^{(\tau)})\right\|_F^2 = \mathcal{O}\left(\left[\frac{\phi(\bar{\mathbf{x}}^{(0)}, \boldsymbol{\Lambda}^{(0)}) - \phi^* + \hat{\delta}_0 + 1}{T} + \frac{1}{\sqrt{\beta}L}\beta^2\sigma^2\right] \cdot \frac{\sigma\kappa^3 L}{\varepsilon}\right). \tag{E.19}$$

Now by $\beta = \Theta\left(\frac{\varepsilon^2}{\sigma^2\kappa^2}\right)$, we have $\frac{\beta^2\sigma^2}{\sqrt{\beta}L} \cdot \frac{\sigma\kappa^3 L}{\varepsilon} = \Theta(\varepsilon^2)$, and the right-hand sides in (E.18) and (E.19) become $\mathcal{O}(T^{-1}\varepsilon^{-1}\sigma\kappa^3 L + \varepsilon^2)$. Hence $(\mathbf{X}^{(\tau)}, \boldsymbol{\Lambda}^{(\tau)})$ is an ε-stationary point when $T = \Theta\left(\frac{\sigma\kappa^3 L}{\varepsilon^3}\right)$. This completes the proof by noticing $T_c = T$ and $T_s = \Theta(T + \mathcal{S}_0)$. □

References

1. Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023. 3
2. L. Chen, H. Ye, and L. Luo. A simple and efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2212.02387*, 2022. 4
3. Z. Chen, S. Ma, and Y. Zhou. Accelerated proximal alternating gradient-descent-ascent for nonconvex minimax machine learning. *arXiv preprint arXiv:2112.11663*, 2021. 5
4. A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019. 4, 6
5. J. M. Danskin. The theory of max-min, with applications. *Siam Journal on Applied Mathematics*, 14:641–664, 1966. 8
6. A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014. 4
7. D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018. 8
8. C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 687–697, Red Hook, NY, USA, 2018. Curran Associates Inc. 4, 6, 30
9. P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 2
10. H. Gao. Decentralized stochastic gradient descent ascent for finite-sum minimax problems. *arXiv preprint arXiv:2212.02724*, 2022. 4, 17
11. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2
12. F. Huang, S. Gao, J. Pei, and H. Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022. 3, 5, 19
13. C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020. 5
14. N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6
15. A. Koloskova, T. Lin, and S. U. Stich. An improved analysis of gradient tracking for decentralized machine learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 6
16. A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 21

17. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 20
18. J. Li, L. Zhu, and A. M.-C. So. Global convergence rate analysis of nonsmooth nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2209.10825v2*, 2023. 5
19. X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5330–5340. Curran Associates, Inc., 2017. 2, 4
20. T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020. 5, 17
21. T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020. 5
22. M. Liu, W. Zhang, Y. Mroueh, X. Cui, J. Ross, T. Yang, and P. Das. A decentralized parallel algorithm for training generative adversarial nets. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11056–11070. Curran Associates, Inc., 2020. 2, 5, 20
23. Z. Liu, X. Zhang, S. Lu, and J. Liu. Precision: Decentralized constrained min-max learning with low communication and sample complexities. *arXiv preprint arXiv:2303.02532*, 2023. 4, 17
24. P. D. Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016. 6
25. S. Lu, X. Zhang, H. Sun, and M. Hong. Gnsd: a gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321, 2019. 6
26. Y. Lu and C. De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pages 7111–7123. PMLR, 2021. 3
27. L. Luo, H. Ye, Z. Huang, and T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20566–20577. Curran Associates, Inc., 2020. 3, 5
28. G. Mancino-Ball, S. Miao, Y. Xu, and J. Chen. Proximal stochastic recursive momentum methods for nonconvex composite decentralized optimization. *arXiv preprint arXiv:2211.11954*, 2022. 10, 19, 24, 30, 32
29. G. Mancino-Ball, Y. Xu, and J. Chen. A decentralized primal-dual framework for non-convex smooth consensus optimization. *IEEE Transactions on Signal Processing*, 71:525–538, 2023. 17
30. M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 09–15 Jun 2019. 21
31. H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2
32. A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27:2597 – 2633, 2017. 6
33. Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013. 9
34. L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pages 2613–2621. PMLR, 2017. 6
35. M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2
36. D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021. 5
37. G. Scutari and Y. Sun. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176:497–544, 2019. 17
38. W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25:944 – 966, 2015. 8
39. J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2015. 22
40. H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu. d^2: Decentralized training over decentralized data. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4848–4856, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 6
41. K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019. 5

42. I. Tsaknakis, M. Hong, and S. Liu. Decentralized min-max optimization: Formulations, algorithms and applications in network poisoning attack. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5755–5759, 2020. 3, 4
43. J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer. A survey on distributed machine learning. *ACM Comput. Surv.*, 53(2), mar 2020. 2
44. W. Xian, F. Huang, Y. Zhang, and H. Huang. A faster decentralized algorithm for nonconvex minimax problems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25865–25877. Curran Associates, Inc., 2021. 2, 4, 5, 20
45. H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 20
46. R. Xin, S. Das, U. A. Khan, and S. Kar. A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency. *arXiv preprint, arXiv:2110.01594*, 2021. 11
47. R. Xin, U. Khan, and S. Kar. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11459–11469. PMLR, 18–24 Jul 2021. 2, 19
48. Y. Xu. Decentralized gradient descent maximization method for composite nonconvex strongly-concave minimax problems. *arXiv preprint arXiv:2304.02441*, 2023. 3, 6, 8, 9, 20
49. Y. Xu and Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *Journal of Optimization Theory and Applications*, 196(1):266–297, 2023. 2, 6
50. Z. Xu, H. Zhang, Y. Xu, and G. Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, pages 1–72, 2023. 5
51. J. Yang, A. Orvieto, A. Lucchi, and N. He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR, 2022. 5
52. J. Zhang and K. You. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2020. 6
53. X. Zhang, N. Aybat, and M. Gurbuzbalaban. SAPD+: An accelerated stochastic method for nonconvex-concave minimax problems. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 5
54. X. Zhang, Z. Liu, J. Liu, Z. Zhu, and S. Lu. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18825–18838. Curran Associates, Inc., 2021. 2, 4, 8, 20
55. T. Zheng, L. Zhu, A. M.-C. So, J. Blanchet, and J. Li. Doubly smoothed gda: Global convergent algorithm for constrained nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2212.12978*, 2022. 8