Orthogonal Updates for the Win: Towards Accelerated Adaptive Minimax Optimization

Zhiwei Zhai[1], **Xinyu Wang**[2], **Wenjing Yan**[1]*, **Lei Ding**[3], **Ying-Jun Angela Zhang**[1]

[1]The Chinese University of Hong Kong [2]McGill University [3]University of Manitoba

zz024@ie.cuhk.edu.hk, xinyu.wang5@mail.mcgill.ca, wjyan@ie.cuhk.edu.hk,
lei.ding@umanitoba.ca, yjzhang@ie.cuhk.edu.hk

Abstract

Single-loop minimax methods are appealing for modern machine learning, but standard stochastic gradient updates can be unstable and oscillatory under tightly coupled primal–dual dynamics, making performance notoriously sensitive to stepsizes. To fundamentally improve stability, we develop an orthogonal update framework tailored to minimax optimization, which applies updates along orthogonalized directions to reduce update anisotropy. However, orthogonalization reshapes the optimization geometry and makes coordinating primal–dual stepsizes even more delicate. To overcome this challenge, we develop **AdaSGDA**, which equips orthogonal updates with an adaptive stepsize rule driven by accumulated gradient norms. This mechanism automatically balances coupled primal–dual progress under orthogonalization, yielding an effectively parameter-free method for nonconvex-strongly-concave minimax optimization. We prove that AdaSGDA avoids problem-dependent tuning while attaining the state-of-the-art $\mathcal{O}\left(T^{-1/4}\right)$ convergence rate. We further propose **AdaSGDA-VR**, which incorporates a carefully designed variance reduction scheme and achieves the faster $\mathcal{O}\left(T^{-1/3}\right)$ rate. Extensive experiments demonstrate that our methods deliver competitive performance compared to existing minimax optimization methods.

1 Introduction

With the rapid growth of machine learning research, many fundamental problems go beyond simple minimization. While classical training objectives can often be written as minimizing an empirical risk, modern learning pipelines frequently involve two competing objectives that naturally lead to minimax formulations. In these problems, one typically minimizes a model-side loss over the primal variable (e.g., network parameters), while maximizing over an adversary or a reweighting distribution that captures worst-case behaviors. Representative examples include generative adversarial networks (GAN) [Goodfellow et al., 2014, Stéphanovitch et al., 2023], robust optimization [Gulrajani et al., 2017, Mohri et al., 2019], adversarial training of neural networks [Wang et al., 2021], and fair machine learning [Madras et al., 2018, Shekhar et al., 2021], etc. These applications highlight the necessity of reliable minimax optimization methods as core building blocks in modern machine learning.

A standard baseline for minimax optimization is Gradient Descent Ascent (GDA) [Dem'yanov and Pevnyi, 1972, Nemirovski et al., 2009], which performs a descent step on the primal variable and an ascent step on the dual variable using stochastic gradients. However, reliable primal progress often requires the dual variable to closely track the inner maximizer, which typically necessitates multiple ascent steps or an explicit inner procedure before each primal update [Yang et al., 2022a]. Such inner-loop computations substantially increase cost and complicate implementation, making them a major obstacle to scalable minimax optimization. This has spurred interest in single-loop methods

*Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

[Yang et al., 2022b], which use only one stochastic gradient per iteration and are therefore easier to deploy. Yet, without an inner procedure, the dual variable may drift away from its instantaneous optimum, and the coupled dynamics become highly sensitive to time-scale separation. Consequently, performance can depend critically on the choice of stepsizes and their relative scaling to prevent instability, oscillation, or stagnation.

Existing theory indicates that optimal stepsize choices in minimax optimization depend on problem-specific quantities such as smoothness, strong concavity, and gradient noise levels [Zhang et al., 2021, Yang et al., 2020]. Even with adaptive methods such as AdaGrad [Ward et al., 2020], current minimax guarantees are only near-optimal, or they still require tuning guided by these problem parameters to keep the primal and dual updates on compatible time scales [Yang et al., 2022a, Li et al., 2022, Huang et al., 2024b]. Without such guidance, standard adaptive minimax updates often fail to properly balance the primal–dual progress, resulting in slow convergence and leading to near-optimal rates. This exposes a double bind: existing methods rarely achieve both optimal convergence rates and problem-parameter-free deployment at the same time.

To fundamentally address this limitation, we consider developing an orthogonal update framework for minimax optimization, which applies updates along orthogonalized directions to regularize the update geometry and improve stability. A recent example highlighting the promise of such updates is the Muon optimizer [Jordan et al., 2024a], which has shown strong empirical success in large language model training [Liu et al., 2025a]. Its key idea is orthogonalized momentum for matrix parameters, which distributes updates more evenly across directions and often leads to more stable progress. However, orthogonalization can also substantially reshape the optimization dynamics [Jordan et al., 2024b], making stepsize selection nontrivial even for standard minimization—and even more delicate in minimax settings where the primal and dual variables must be coordinated. Yet, the very mechanism that makes orthogonalized updates effective also obscures how their stepsizes should be chosen, especially under coupled minimax dynamics. This raises a fundamental question:

Can we design a single-loop, parameter-free minimax algorithm with orthogonal updates that achieves the optimal convergence rate, matching well-tuned counterparts?

1.1 Main Contribution

We answer this question affirmatively. We develop **AdaSGDA**, a single-loop algorithm for nonconvex-strongly-concave (NC–SC) minimax optimization that leverages the complementary strengths of orthogonal updates and adaptive stepsizes. AdaSGDA stabilizes the coupled primal–dual dynamics by updating along orthogonalized directions, while an accumulated-gradient-norm-based scaling rule automatically balances the relative progress of the two variables. As a result, AdaSGDA is problem-parameter-free and provably attains the optimal convergence rate. Our main contributions are summarized as follows:

- We propose **AdaSGDA**, the first *parameter-free* algorithm that attains the *optimal convergence rate* for minimax optimization. AdaSGDA is built upon orthogonalized momentum with adaptive scaling, but making these ingredients work for stochastic minimax is highly nontrivial: orthogonalization can substantially amplify stochastic noise when gradient matrices are ill conditioned [Zhang, 2025], making naive norm accumulation ineffective. Moreover, the coupled minimax dynamics require the primal and dual updates to be coordinated within a single-loop scheme. To address these issues, AdaSGDA introduces two key mechanisms: **(i) orthogonalization-calibrated adaptive stepsizes**, which adapt to historical gradient information while applying updates in a geometry that preserves orthogonality; and **(ii) hierarchical stepsize design** that respects the interdependence between the two variables while preserving the autonomy of adaptive stepsizes.
- We conduct a comprehensive theoretical analysis for our adaptive minimax scheme. This setting is technically challenging because adaptive scaling and orthogonalized directions make the effective geometry and stochastic error evolve over time, while the primal and dual updates remain tightly coupled. Establishing parameter-free convergence therefore calls for a new proof framework beyond existing minimax analyses. In our analysis, we show that AdaSGDA achieves a convergence rate of $\mathcal{O}(T^{-1/4})$, matching well-tuned methods, and that incorporating variance reduction further improves this rate to $\mathcal{O}(T^{-1/3})$. Notably, in contrast to existing parameter-free minimax approaches [Yang et al., 2022a, Li et al., 2022, Huang et al., 2024b], which typically obtain only near-optimal rates, our guarantees attain the best-known optimal rate.

Table 1: Comparison between different minimax optimization algorithms.
T denotes the number of iterations; $\widetilde{\mathcal{O}}(\cdot)$ suppresses logarithmic factors in the convergence bound; α and β are stepsize-decay exponents used in TiAda and D-AdaST, satisfying $0 < \beta < \alpha < 1$.

Algorithm	**Single-Loop**	**Parameter-Free**	**Optimal-Rate**[†]	**Convergence Rate**[◇]
GDA Jin et al. [2019]	✘	✘	✔	$\mathcal{O}\Big(\frac{1}{T^{1/4}}\Big)$
TTGDA Lin et al. [2025]	✔	✘	✔	$\mathcal{O}\Big(\frac{1}{T^{1/4}}\Big)$
NeAda Yang et al. [2022a]	✘	✔	✘	$\widetilde{\mathcal{O}}\Big(\frac{1}{T^{1/4}}\Big)$
TiAda Li et al. [2022]	✔	✔	✘	$\mathcal{O}\left(\frac{1}{T^{(1-\alpha)/2}} + \frac{1}{T^{\beta/2}}\right)$
D-AdaST Huang et al. [2024b]	✔	✔	✘	$\widetilde{\mathcal{O}}\left(\frac{1}{T^{(1-\alpha)/2}} + \frac{1}{T^{\beta/2}}\right)$
AdaSGDA (This paper)	✔	✔	✔	$\mathcal{O}\left(\frac{1}{T^{1/4}}\right)$
AdaSGDA-VR (This paper)	✔	✔	✔	$\mathcal{O}\left(\frac{1}{T^{1/3}}\right)$

[†] Optimal-Rate indicates whether the convergence rate matches the best-known scaling in T for the considered setting, without extra residual terms or logarithmic overheads (up to constant factors).
[◇] Convergence rate refers to an upper bound on the expected first-order stationarity measure averaged over T iterations.

- We evaluate our method on several benchmark tasks, including distributionally robust image classification and large language model pretraining, as well as GAN training, and show that our method performs competitive with existing well-tuned baselines. Moreover, our method is markedly less sensitive to the initial stepsize, remaining stable across a broad range of choices, which supports the effectiveness of our adaptive minimax approach.

1.2 Comparisons with Prior Approaches

We compare AdaSGDA with representative minimax methods in Table 1. Compared with non–parameter-free baselines, including non-loopless GDA [Jin et al., 2019] and loopless TTGDA [Lin et al., 2025], AdaSGDA matches their optimal rate $\mathcal{O}(T^{-1/4})$ while removing problem-parameter-based tuning. This matters in practice, since such parameters are rarely known and typically lead to substantial manual tuning. Moreover, AdaSGDA-VR further reduces stochastic errors via variance reduction and improves the rate to $\mathcal{O}(T^{-1/3})$, outperforming these well-tuned counterparts. Turning to adaptive methods, NeAda [Yang et al., 2022a] is parameter-free but nested and near-optimal with logarithmic overheads, while loopless TiAda [Li et al., 2022] and D-AdaST [Huang et al., 2024b] are also near-optimal due to residual terms. In contrast, AdaSGDA attains the optimal rate in a single loop by carefully controlling the primal–dual time scales via adaptive scaling and a hierarchical coordination rule. Moreover, we pair orthogonalized updates with orthogonalization-calibrated adaptivity, which stabilizes the induced matrix dynamics and prevents adversarial growth of update directions. This structure enables a sharper control of stochastic errors, leading to parameter-free guarantees that achieve the best-known optimal rate. As a result, our method is the first single-loop minimax method that is simultaneously parameter-free and rate-optimal. This represents a significant advancement over existing minimax methods, facilitating robust deployment across diverse environments.

2 Preliminaries

2.1 Problem Model

In this paper, we consider the following minimax optimization problem:

$$\min_{X\in\mathbb{R}^{m_x\times n_x}} \max_{Y\in\mathbb{R}^{m_y\times n_y}} f(X,Y) \triangleq \mathbb{E}_{\xi\sim P}[F(X,Y;\xi)]\,, \tag{1}$$

where P is an unknown distribution from which i.i.d. samples can be drawn, and ξ denotes a random variable. We assume that f is jointly L-smooth in (X,Y), possibly nonconvex in X, and μ-strongly concave in Y.

Notations. We denote L as the smoothness parameter and μ as the strong-concavity parameter, whose formal definitions are given in Assumptions 4.1 and 4.2. We assume access to a stochastic first-order oracle that returns $\big[\nabla_X F(X,Y;\xi),\ \nabla_Y F(X,Y;\xi)\big]$. For problem (1), we denote $Y^*(X) := \arg\max_{Y\in\mathbb{R}^{m_y\times n_y}} f(X,Y)$ as the solution of the inner maximization and $\Phi(X) := f\big(X, Y^*(X)\big)$

Algorithm 1 Adaptive Minimax Optimization with Orthogonal Updates (AdaSGDA)

Initialization: X_0, Y_0, $M_0^X = 0$, $M_0^Y = 0$, buffers $m_0^X = m_0^Y = c > 0$, hyperparameters $\eta_X, \eta_Y, \epsilon_X, \epsilon_Y, \gamma > 0$, and momentum parameter β.
for iteration $t = 0, 1, \cdots, T-1$ **do**
  Sample mini-batch of size b_t and compute stochastic gradients:
  $G_t^X = \nabla_X F(X_t, Y_t; \xi_t^X)$, $G_t^Y = \nabla_Y F(X_t, Y_t; \xi_t^Y)$.
  Momentum updates:
  $M_{t+1}^X = \beta M_t^X + (1-\beta) G_t^X$,
  $M_{t+1}^Y = \beta M_t^Y + (1-\beta) G_t^Y$.
  Accumulate the gradient norms:
  $[m_{t+1}^X]^2 = [m_t^X]^2 + \min\{\|G_t^X\|_F^2, \gamma^2\}$,
  $[m_{t+1}^Y]^2 = [m_t^Y]^2 + \min\{\|G_t^Y\|_F^2, \gamma^2\}$.
  Orthogonalize the momentum matrices:
  $O_{t+1}^X = \operatorname{argmin}_{O \in \mathbb{R}^{m_x \times n_x}} \|O - M_{t+1}^X\|_F$,
  s.t. $O^\top O = I_{n_x}$ if $m_x \geq n_x$, else $OO^\top = I_{m_x}$,
  $O_{t+1}^Y = \operatorname{argmin}_{O \in \mathbb{R}^{m_y \times n_y}} \|O - M_{t+1}^Y\|_F$,
  s.t. $O^\top O = I_{n_y}$ if $m_y \geq n_y$, else $OO^\top = I_{m_y}$.
  Define adaptive stepsizes:
  $\alpha_{t+1}^Y = \max\left\{\epsilon_Y,\ \eta_Y \frac{\min\{\|G_t^Y\|_F, \gamma\}}{m_{t+1}^Y}\right\}$,
  $\alpha_{t+1}^X = \max\left\{\epsilon_X,\ \eta_X \frac{\min\{\|G_t^X\|_F, \gamma\}}{\max\{m_{t+1}^X, m_{t+1}^Y\}}\right\}$.
  Update primal and dual variables:
  $Y_{t+1} = Y_t + \alpha_{t+1}^Y O_{t+1}^Y$,
  $X_{t+1} = X_t - \alpha_{t+1}^X O_{t+1}^X$.
end for

as the induced primal function. For a matrix $A \in \mathbb{R}^{m \times n}$, $\|A\|_F$ denotes the Frobenius norm, $\|A\|_*$ denotes the nuclear norm, and $\|A\|_{\text{op}}$ denotes the spectral (operator) norm. We will repeatedly use the standard bound $\|A\|_* \leq \sqrt{\operatorname{rank}(A)}\,\|A\|_F \leq \sqrt{r}\,\|A\|_F$, where r is a global rank upper bound for all matrices considered (e.g., $r := \min\{m, n\}$ for $A \in \mathbb{R}^{m \times n}$). In addition, for $A, B \in \mathbb{R}^{m \times n}$, we define $\langle A, B \rangle := \operatorname{Tr}(A^\top B)$, where $\operatorname{Tr}(\cdot)$ denotes the trace.

2.2 Orthogonal Updates

Recent progress on orthogonalized optimization, exemplified by Muon [Jordan et al., 2024b], advocates exploiting the intrinsic matrix structure of network parameters rather than treating them as long vectors. The key operation is to transform a momentum (or gradient) matrix $M_t \in \mathbb{R}^{m \times n}$ into an orthogonalized search direction O_t that preserves directional information while discarding magnitude effects. To streamline the exposition, we focus on the case $m \geq n$; the remaining case can be handled analogously. Concretely, Muon computes the closest matrix with orthonormal columns:

$$O_t \ := \ \arg \min_{O \in \mathbb{R}^{m \times n}:\ O^\top O = I_n} \|O - M_t\|_F. \tag{2}$$

By classical matrix theory, if $M_t = U_t \Sigma_t V_t^\top$ is the singular value decomposition (SVD), then the solution is the polar factor $O_t = U_t V_t^\top$. An equivalent characterization is $O_t \ = \ M_t (M_t^\top M_t)^{-1/2}$, which highlights that orthogonalization reduces to applying an inverse square root rather than recovering all singular components. Computing Eq. (2) exactly via SVD is often impractical for the large matrices encountered in modern deep networks. In practice, Muon instead approximates the polar factor by applying a few Newton–Schulz (NS) iterations to the inverse-square-root term in the polar decomposition [Bernstein and Newhouse, 2024, Higham, 2008, Björck and Bowie, 1971, Kovarik, 1970, Bernstein, 2025]. This yields an efficient and numerically stable approximation that is typically close to the SVD-based solution while being far less computationally demanding, thereby making orthogonalized updates feasible at scale. We adopt this orthogonalization scheme for implementation in our adaptive minimax algorithm.

3 Algorithm Development

In this section, we present the design of our adaptive minimax algorithm, which is driven by the need to address two key challenges:

1. **Time-scale adaptivity in single-loop minimax.** Unlike minimization, minimax dynamics require properly separated primal–dual time scales. As shown by Yang et al. [2022a], directly applying standard adaptive stepsizes to GDA—where each variable adapts only to the (moving) statistics of its own past gradients—can fail to converge unless the initial stepsize ratio between x and y is carefully calibrated. This implies that we must precisely control the relative progress of x and y to maintain the required time-scale separation and ensure stable convergence.
2. **Adaptivity under orthogonalized directions.** Orthogonal updates can substantially reshape the optimization dynamics [Jordan et al., 2024b]. The main difficulty is that orthogonalization is a nonlinear transformation of the gradient: two gradients that are close in Frobenius norm can lead to orthogonalized directions that differ much more, especially when the underlying gradient matrices are ill-conditioned [Zhang, 2025, Liu et al., 2025a]. Consequently, stochastic noise can have a disproportionately large impact on the update direction, and simply accumulating raw gradient norms (e.g., AdaGrad-Norm [Xie et al., 2020]) may not provide a reliable scale for stepsize adaptation in this setting.

We address these challenges by proposing **AdaSGDA**, a single-loop method with explicit time-scale adaptivity and orthogonalization-calibrated adaptive stepsizes. To further improve stochastic efficiency, we develop **AdaSGDA-VR** by incorporating variance reduction. We prove that both algorithms converge to an ϵ-stationary point of problem (1) with state-of-the-art rates.

3.1 Algorithm Development of AdaSGDA

We first present the development of our adaptive minimax algorithm with orthogonal updates, **AdaSGDA**; the detailed procedure is outlined in Algorithm 1. AdaSGDA is built on three key mechanisms: orthogonalized momentum, a hierarchical stepsize design, and orthogonalization-calibrated adaptive stepsizes. We elaborate on each mechanism below.

Orthogonalized Momentum. The momentum is incorporated to smooth stochastic fluctuations and to stabilize the optimization trajectory. At iteration t, we maintain momentum matrices for the primal and dual updates:

$$M_{t+1}^X = \beta M_t^X + (1-\beta)\nabla_X F(X_t, Y_t; \xi_t^X), \quad M_{t+1}^Y = \beta M_t^Y + (1-\beta)\nabla_Y F(X_t, Y_t; \xi_t^Y),$$

where $\beta \in (0,1)$ is the momentum coefficient, $\nabla_X F(X_t, Y_t; \xi_t^X)$ and $\nabla_Y F(X_t, Y_t; \xi_t^Y)$ are stochastic gradient estimators evaluated at the current iterate (X_t, Y_t) (based on randomness ξ_t^X and ξ_t^Y), and M_t^X, M_t^Y store the corresponding momentum matrices from the previous iteration.

Different from standard momentum methods that update parameters directly along M_{t+1}^X and M_{t+1}^Y, we subsequently orthogonalize these momentum matrices and use the orthogonalized directions for updating X and Y. In this way, the update direction leverages the accumulated matrix-structured signal in the momentum while being less sensitive to its magnitude, which leads to more stable orthogonal updates in our minimax framework.

Hierarchical Stepsize Design. A key objective of AdaSGDA is to avoid stepsize choices that depend on unknown problem constants. To this end, we use AdaGrad-type accumulators that track the magnitude of past stochastic gradients. In addition, we introduce a clamping constant $\gamma > 0$ and accumulate $\min\{\|\nabla_X F(X_t, Y_t; \xi_t^X)\|_F^2, \gamma^2\}$ and $\min\{\|\nabla_Y F(X_t, Y_t; \xi_t^Y)\|_F^2, \gamma^2\}$ instead of the raw squared gradient norms. This cap limits the contribution of each iteration to at most γ^2, so the accumulators m_t^X and m_t^Y grow in a more controlled way. Practically, clamping acts as a safeguard against occasional large stochastic gradients, preventing the adaptive denominators from being dominated by rare spikes and thereby stabilizing the resulting stepsizes and the primal–dual time-scale coordination.

Concretely, we define the accumulators for the dual and primal variables as

$$[m_{t+1}^Y]^2 = [m_t^Y]^2 + \min\left\{\left\|\nabla_Y F(X_t, Y_t; \xi_t^Y)\right\|_F^2, \gamma^2\right\}, \quad [m_{t+1}^X]^2 = [m_t^X]^2 + \min\left\{\left\|\nabla_X F(X_t, Y_t; \xi_t^X)\right\|_F^2, \gamma^2\right\}.$$

Based on these accumulators, we first set the adaptive scale for the dual variable as $\frac{\eta_Y}{m_{t+1}^Y}$, where $\eta_Y > 0$ is a control coefficient independent of problem parameters.

For the primal variable X, using the same form $\frac{\eta_X}{m_{t+1}^X}$ is generally insufficient in minimax optimization, because the primal update is sensitive to how accurately Y_t tracks the current maximizer $Y^*(X_t)$. If Y_t is still far from $Y^*(X_t)$, an overly aggressive primal step can destabilize the coupled dynamics. To account for this interdependence, we adopt a hierarchical scale for X: $\frac{\eta_X}{\max\{m_{t+1}^X, m_{t+1}^Y\}}$, where

$\eta_X > 0$ is also problem-parameter-free. This design automatically makes the primal step more conservative when the dual variable is still changing rapidly. It enforces the desired time-scale separation within a single loop and promotes stable convergence, without requiring any problem-parameter-based tuning.

Orthogonalization-Calibrated Adaptive Stepsizes. The hierarchical scales specify how the stepsizes should evolve over time to coordinate the primal–dual progress. We now turn to the remaining question: how should these scales be applied when the update directions are orthogonalized.

In AdaSGDA, we update X and Y along orthogonalized directions computed from the momentum matrices. Concretely, we compute the orthogonalized momentum directions O_{t+1}^X and O_{t+1}^Y to update X and Y, respectively, as

$$O_t^X = \operatorname*{argmin}_{O\in\mathbb{R}^{m_x\times n_x}} \{\|O-M_t^X\|_F : OO^\top = I_{m_x} \text{ or } O^\top O = I_{n_x}\},$$
$$O_t^Y = \operatorname*{argmin}_{O\in\mathbb{R}^{m_y\times n_y}} \{\|O-M_t^Y\|_F : OO^\top = I_{m_y} \text{ or } O^\top O = I_{n_y}\},$$

where the orthogonality constraint is chosen according to the matrix shape: for $O_t^X \in \mathbb{R}^{m_x\times n_x}$ we impose $(O_t^X)^\top O_t^X = I_{n_x}$ when $m_x \geq n_x$, and $O_t^X (O_t^X)^\top = I_{m_x}$ otherwise; the same convention applies to $O_t^Y \in \mathbb{R}^{m_y\times n_y}$. Moreover, since the orthogonalized directions satisfy $\|O_{t+1}^X\|_{\mathrm{op}} = \|O_{t+1}^Y\|_{\mathrm{op}} = 1$, we calibrate the effective stepsizes using the current stochastic gradient norms by scaling them with the clamped factors $\min\{\|\nabla_X F(X_t, Y_t; \xi_t^X)\|_F, \gamma\}$ and $\min\{\|\nabla_Y F(X_t, Y_t; \xi_t^Y)\|_F, \gamma\}$. This calibration makes the update magnitude shrink as the gradients shrink. In particular, near a stationary point where the gradients are small, the per-iteration updates naturally decay to zero, which is a desirable null-gradient consistency property in first-order optimization [Lau et al., 2025].

To avoid overly small steps caused by growing accumulators, we also impose lower bounds $\epsilon_X, \epsilon_Y > 0$ on the effective stepsizes. This ensures nontrivial progress and guarantees that AdaSGDA is at least as aggressive as using orthogonal updates with a small constant stepsize. Importantly, ϵ_X and ϵ_Y can be chosen without any problem-dependent constants.

Putting these components together, we define the calibrated stepsizes as

$$\alpha_{t+1}^X = \max\left\{\epsilon_X, \eta_X \frac{\min\{\|\nabla_X F(X_t, Y_t; \xi_t^X)\|_F, \gamma\}}{\max\{m_{t+1}^X, m_{t+1}^Y\}}\right\}, \alpha_{t+1}^Y = \max\left\{\epsilon_Y, \eta_Y \frac{\min\{\|\nabla_Y F(X_t, Y_t; \xi_t^Y)\|_F, \gamma\}}{m_{t+1}^Y}\right\}.$$

The resulting updates are $Y_{t+1} = Y_t + \alpha_{t+1}^Y O_{t+1}^Y$ and $X_{t+1} = X_t - \alpha_{t+1}^X O_{t+1}^X$.

Overall, this orthogonalization-calibrated design couples the hierarchical scaling with the orthogonal geometry, leading to stable single-loop minimax updates without problem-specific tuning.

3.2 Algorithm Development of AdaSGDA-VR

To further accelerate convergence under stochastic gradients, we integrate variance reduction into AdaSGDA and propose **AdaSGDA-VR** (Algorithm 2 in Section D). The key idea is to construct a momentum-based variance-reduced gradient surrogate that reuses information from the previous iterate, thereby reducing the variance of stochastic updates. Our design is inspired by the variance-reduction principles behind SPIDER [Fang et al., 2018], STORM [Cutkosky and Orabona, 2019], and SVRG-type updates [Zhou et al., 2020].

Specifically, AdaSGDA-VR augments the momentum updates with a variance-reduction correction computed from the same mini-batch evaluated at the current and previous iterates. For $t \geq 1$, the momentum matrices for the primal and dual variables are updated as

$$M_{t+1}^X = \beta M_t^X + (1-\beta)\nabla_X F(X_t, Y_t; \xi_t^X) + \beta\left(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X)\right),$$
$$M_{t+1}^Y = \beta M_t^Y + (1-\beta)\nabla_Y F(X_t, Y_t; \xi_t^Y) + \beta\left(\nabla_Y F(X_t, Y_t; \xi_t^Y) - \nabla_Y F(X_{t-1}, Y_{t-1}; \xi_t^Y)\right).$$

This structured correction is designed to cancel a large portion of the noise introduced by mini-batch sampling, thereby yielding a lower-variance update direction for subsequent orthogonalization and updates [Cutkosky and Orabona, 2019]. After forming the variance-reduced momentum matrices, AdaSGDA-VR applies the same orthogonalization procedure and calibrated stepsize rules as AdaSGDA. This results in a fully single-loop algorithm that remains parameter-free and time-scale adaptive, while achieving faster convergence by reducing stochastic errors.

4 Theoretical Analysis

4.1 Technical challenges

Establishing optimal convergence guarantees for single-loop, parameter-free minimax optimization with orthogonal updates poses several technical difficulties:

- **Coupled primal–dual dynamics.** The minimax scheme tightly couples the evolution of (X_t, Y_t): the primal progress depends on how well the dual variable tracks $Y^\star(X_t)$, while the dual update is driven by a moving target that changes with X_t. This dependence makes the analysis more delicate than in single-variable minimization.
- **Adaptive time-scale control under orthogonal geometry.** In a single-loop scheme, the stepsizes for X and Y cannot be analyzed separately: their history-dependent denominators must jointly enforce the required time-scale separation. Moreover, updates are applied along unit-magnitude orthogonalized directions rather than raw gradients, which changes the effective geometry and requires careful stepsize calibration. Establishing convergence in this coupled, stochastic two-time-scale setting calls for a refined analysis beyond prior minimax work.
- **Parameter-free stepsizes with optimal rates.** Achieving the best-known minimax rate typically relies on stepsizes tuned to problem-specific constants. In the parameter-free regime, the analysis must show that our adaptive stepsizes still realize the correct balance between descent in the primal objective and sufficient ascent in the dual variable, so that the overall error terms telescope to the optimal $\mathcal{O}(T^{-1/4})$ rate (and $\mathcal{O}(T^{-1/3})$ with variance reduction).

4.2 Assumptions

We consider the NC-SC setting of Problem (1) with the following assumptions that are commonly used in the existing works [Yang et al., 2022b,a, Li et al., 2022, Huang et al., 2024b, Lin et al., 2025, He et al., 2025].

Assumption 4.1 (Joint smoothness)**.** The stochastic objective function $F(X, Y; \xi)$ is jointly L-smooth ($L > 0$) in (X, Y). That is, for any $X_1, X_2 \in \mathbb{R}^{m_x \times n_x}$ and $Y_1, Y_2 \in \mathbb{R}^{m_y \times n_y}$, we have

$$\mathbb{E}_\xi[\|\nabla_X F(X_1, Y_1; \xi) - \nabla_X F(X_2, Y_2; \xi)\|_*^2] \le L^2 \left(\|X_1 - X_2\|_{\text{op}}^2 + \|Y_1 - Y_2\|_{\text{op}}^2\right),$$

$$\mathbb{E}_\xi[\|\nabla_Y F(X_1, Y_1; \xi) - \nabla_Y F(X_2, Y_2; \xi)\|_*^2] \le L^2 \left(\|X_1 - X_2\|_{\text{op}}^2 + \|Y_1 - Y_2\|_{\text{op}}^2\right).$$

Assumption 4.2 (μ-Strong concavity in Y)**.** Function $f(X, Y)$ is μ-strongly concave ($\mu > 0$) in Y, that is, for any $X \in \mathbb{R}^{m_x \times n_x}$ and $Y_1, Y_2 \in \mathbb{R}^{m_y \times n_y}$, we have

$$f(X, Y_1) \ge f(X, Y_2) + \langle \nabla_Y f(X, Y_1), Y_1 - Y_2 \rangle + \frac{\mu}{2}\|Y_1 - Y_2\|_F^2.$$

Assumption 4.3 (Stochastic gradients)**.** $\nabla_X F(X_t, Y_t; \xi_t^X)$ and $\nabla_Y F(X_t, Y_t; \xi_t^Y)$ are unbiased stochastic gradient estimators such that $\mathbb{E}[\nabla_X F(X_t, Y_t; \xi_t^X)] = \nabla_X f(X_t, Y_t)$ and $\mathbb{E}[\nabla_Y F(X_t, Y_t; \xi_t^Y)] = \nabla_Y f(X_t, Y_t)$, with a uniformly bounded variance:

$$\mathbb{E}\left[\|\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\|_F^2\right] \le \frac{\kappa^2}{b_t}, \mathbb{E}\left[\|\nabla_Y F(X_t, Y_t; \xi_t^Y) - \nabla_Y f(X_t, Y_t)\|_F^2\right] \le \frac{\kappa^2}{b_t},$$

where $b_t \ge 1$ is the batch size. Moreover, there exists $G > 0$ such that $\|\nabla_Z F(X_s, Y_s; \xi_t^Z)\|_F \le G$ almost surely for $Z \in \{X, Y\}$ and every stochastic oracle evaluation made by Algorithms 1 and 2 along their generated trajectories.

Remark 4.4. By the first-order optimality condition for the inner maximization defining $Y^\star(X)$, we have $\nabla_Y f(X, Y^\star(X)) = 0$. This property is important for AdaGrad-type stepsizes whose denominators accumulate past gradient norms. If the dual gradient does not vanish at $Y^\star(X)$, the accumulator may keep growing even near optimality, forcing the stepsize to keep shrinking and slowing convergence [Bach and Levy, 2019, Li et al., 2022]. In addition, bounded gradients (Assumption 4.3) are a standard assumption for controlling adaptive denominators in AdaGrad- and Adam-type analyses [Levy et al., 2018, Zou et al., 2019, Ivgi et al., 2023], and have also been commonly adopted in minimax optimization [Li et al., 2022, Huang et al., 2024b].

4.3 Theoretical Results

In this subsection, we present the main theoretical results for the proposed Algorithm 1 and Algorithm 2. As outlined in Lemma C.1, the convergence behavior is governed by the following key terms: (i) $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|M_t^X - \nabla\Phi(X_{t-1})\|_*\right]$, (ii) $\frac{\eta_X\sqrt{r}}{\epsilon_X T}\sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2}\right)+1}\sqrt{\sum_{t=1}^T \mathbb{E}\left[\|M_t^X - \nabla\Phi(X_{t-1})\|_F^2\right]}$, where $z_0 = \max\{m_0^X, m_0^Y\}$. Term (i)

is the average gradient estimation error measured in the nuclear norm $\|\cdot\|_*$, while term (ii) depends on the cumulative squared estimation error measured in the Frobenius norm $\|\cdot\|_F$.

Algorithm 1. Under the hyperparameter choices in Theorem 4.5, our analysis shows that term (i) is upper bounded by $\mathcal{O}\Big(T^{-\frac{1}{4}}\Big)$ and term (ii) is upper bounded by $\mathcal{O}\Big(T^{-\left(\frac{3}{8}+q\right)}\Big)$. Plugging these bounds into Lemma C.1 yields Theorem 4.5.

Theorem 4.5. *Under Assumptions 4.1 to 4.3, when we set* $1-\beta=\frac{1}{T^{\frac{1}{4}}}$, $\epsilon_X=\frac{1}{T^{\frac{3}{4}}}$, $\epsilon_Y=\frac{1}{T^{\frac{1}{2}}}$, $\eta_X=\frac{1}{T^{\frac{7}{8}+q}}$, $\eta_Y=\frac{1}{T^{\frac{5}{8}+q}}$, *where* $q\geq 0$ *is a fixed exponent offset used in the theoretical schedules, together with* $\gamma>0$, $z_0>0$, *and* $b_t\geq T^{\frac{1}{4}}$ *for all* $t\in\{1,\ldots,T\}$, *the iterates generated by Algorithm 1 satisfy:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Phi(X_t)\|_*]\leq\mathcal{O}\Bigg(\frac{\Delta+\kappa\sqrt{r}}{T^{\frac{1}{4}}}+\frac{GL\sqrt{r}(1+\lambda)+\kappa L\sqrt{r}+L^2r}{\mu T^{\frac{1}{4}}}\Bigg)+\mathcal{O}\Bigg(\frac{Lr^{\frac{3}{4}}\sqrt{\kappa G}+L^{\frac{3}{2}}r\sqrt{G}}{\mu T^{\frac{3}{8}+q}}\sqrt{\ln\Big(\frac{\gamma^2T}{z_0^2}\Big)+1}\Bigg),$$

where $\Delta:=\Phi(X_0)-\Phi_*$, $\Phi_*:=\min_X\Phi(X)$, *and* λ *denotes the Lipschitz constant of* $Y^*(\cdot)$ *as defined in Lemma B.1.*

where $\Delta:=\Phi(X_0)-\Phi_*$, $\Phi_*:=\min_X\Phi(X)$, λ *denotes the Lipschitz constant of* $Y^*(\cdot)$ *as defined in Lemma B.1, and* $q>0$ *is any fixed arbitrarily small constant used in the stepsize exponents; it only affects the higher-order term and does not change the leading* $\mathcal{O}(T^{-1/4})$ *rate.*

Algorithm 2. Under the hyperparameter choices in Theorem 4.6, we obtain that term (i) is upper bounded by $\mathcal{O}\Big(T^{-\frac{1}{3}}\Big)$ and term (ii) is upper bounded by $\mathcal{O}\Big(T^{-\left(\frac{1}{2}+q\right)}\Big)$. Substituting them into Lemma C.1 establishes Theorem 4.6.

Theorem 4.6. *Under Assumptions 4.1 to 4.3, when we set* $1-\beta=\frac{1}{T^{\frac{1}{3}}}$, $\epsilon_X=\frac{1}{T^{\frac{2}{3}}}$, $\epsilon_Y=\frac{1}{T^{\frac{1}{3}}}$, $\eta_X=\frac{1}{T^{\frac{2}{3}+q}}$, $\eta_Y=\frac{1}{T^{\frac{1}{2}+q}}$, $\gamma>0$, $z_0>0$, $b_t\geq T^{\frac{1}{3}}$ *for all* $t\in\{1,\cdots,T\}$, *the iterates generated by Algorithm 2 satisfy:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Phi(X_t)\|_*]\leq\mathcal{O}\Bigg(\frac{\Delta+rL+\kappa\sqrt{r}}{T^{\frac{1}{3}}}+\frac{GL\sqrt{r}(1+\lambda)+\kappa L\sqrt{r}+L^2r}{\mu T^{\frac{1}{3}}}\Bigg)+\mathcal{O}\Bigg(\frac{Lr}{T^{\frac{1}{2}+q}}\sqrt{\ln\Big(\frac{\gamma^2T}{z_0^2}\Big)+1}\Bigg).$$

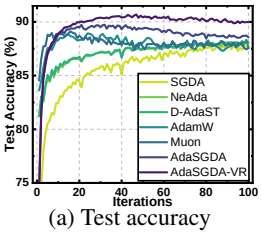
Remark 4.7. Our theoretical analysis establishes optimal convergence guarantees for adaptive minimax optimization with orthogonal updates. Existing adaptive minimax methods only provide near-optimal rates [Yang et al., 2022a, Li et al., 2022, Huang et al., 2024b], and attaining optimal rates often requires stepsize tuning based on problem-specific constants such as smoothness or strong concavity parameters (e.g., Jin et al., 2019, Lin et al., 2025). In sharp contrast, our orthogonal adaptive design is inherently parameter-free yet attains the optimal rate in Theorem 4.5. In addition, our carefully designed variance reduction scheme, together with a refined control of stochastic errors in the theoretical analysis, leads to the faster guarantee in Theorem 4.6, matching the best-known rate for NC-SC minimax optimization.

5 Numerical Experiments

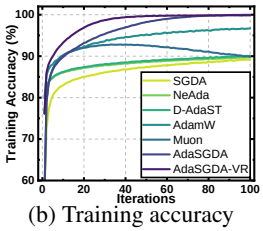
In this section, we evaluate **AdaSGDA** and **AdaSGDA-VR** against representative minimax baselines, including SGDA [Lin et al., 2025], NeAda [Yang et al., 2022a], D-AdaST [Huang et al., 2024b], AdamW [Loshchilov and Hutter, 2017], and Muon-based minimax method [Jordan et al., 2024a]. Our experiments cover two representative minimax tasks: (i) NC–SC distributional robustness optimization (DRO) [Sinha et al., 2017], and (ii) training the nonconvex–nonconcave (NC–NC) Wasserstein GAN with gradient penalty [Gulrajani et al., 2017]. Experiments are conducted on MNIST [LeCun et al., 2002], Fashion-MNIST (FMNIST) [Xiao et al., 2017], CIFAR-10 and CIFAR-100 [Krizhevsky et al., 2009], and on FineWeb [Penedo et al., 2024] using OPT-125M, OPT-350M, and OPT-1.3B. Additional results and configurations are provided in Section E.

5.1 Distributional Robustness Optimization

In this subsection, we consider DRO following the adversarial-perturbation formulation in Sinha et al. [2017]. The goal is to learn a model that is robust to worst-case perturbations of the training inputs. We treat the model parameters as the primal variable and the perturbations as the dual variable. Let $\{\xi_i\}_{i=1}^n$ be the training samples, and let V_i denote the corresponding input representation (e.g., a

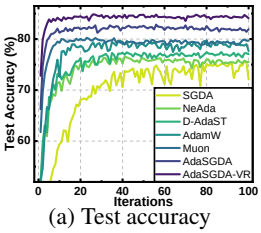


(a) Test accuracy

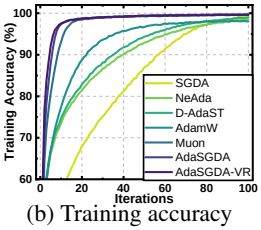


(b) Training accuracy

Figure 1: Performance comparison on FMNIST.

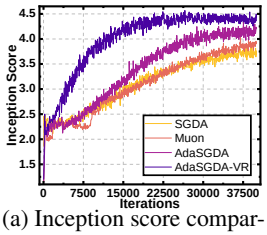


(a) Test accuracy

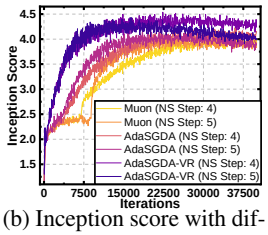


(b) Training accuracy

Figure 2: Performance comparison on CIFAR-10.

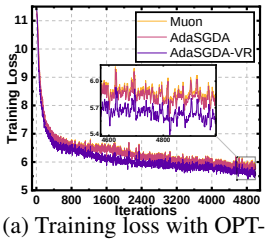


(a) Inception score comparison among baselines

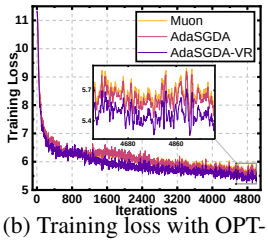


(b) Inception score with different NS steps

Figure 3: Inception score on WGAN-GP.



(a) Training loss with OPT-125M model



(b) Training loss with OPT-1.3B model

Figure 4: Training loss on the FineWeb dataset.

raw image for vision tasks, or an embedding sequence for language modeling). For each sample, we introduce a perturbed input Y_i and solve $\min_X \max_{Y=[Y_1,\dots,Y_n]} \frac{1}{n}\sum_{i=1}^n \ell(X;Y_i) - \rho\|Y_i - V_i\|_F^2$, where $\ell(X;Y_i)$ is the task loss evaluated at the perturbed input, and $\rho > 0$ controls the perturbation strength by penalizing deviations from V_i. The inner maximization searches for worst-case perturbations within a local neighborhood, while the outer minimization learns parameters that perform well under these worst-case inputs. When ρ is sufficiently large relative to the smoothness of $\ell(X;\cdot)$ in Y_i, the inner objective becomes strongly concave in Y, yielding an NC–SC minimax instance. We evaluate this DRO setup on MNIST, FMNIST, CIFAR-10/100, and FineWeb datasets.

In Figures 1 and 2, we report the training and test accuracy on FMNIST and CIFAR-10; additional results on MNIST and CIFAR-100 are provided in Section E. Overall, **AdaSGDA** achieves the best performance across all datasets and baselines. Moreover, **AdaSGDA-VR** further improves over AdaSGDA, as variance reduction reduces the stochastic gradient noise and leads to more stable and faster progress in practice. Notably, our methods also consistently outperform Muon. This advantage is attributed to our parameter-free design under orthogonal updates: the accumulated-gradient-norm-based stepsizes automatically adapt to the training dynamics, eliminating manual stepsize tuning while enabling convergence at optimal rates under orthogonalization. In addition, we evaluate on FineWeb using OPT-125M and OPT-1.3B in Figure 4. Our methods again outperform the Muon-based minimax optimization method, demonstrating strong performance in more challenging large-scale settings.

5.2 Generative Adversarial Networks

Another widely studied application of minimax optimization is generative adversarial networks. In GAN training, the discriminator (critic) learns to distinguish real from generated samples, while the generator is trained simultaneously to produce samples that can fool the critic, resulting in a coupled minimax game. In our experiments, we adopt the WGAN-GP objective [Gulrajani et al., 2017], which enforces the critic to be approximately 1-Lipschitz via a gradient penalty, and evaluate performance on the CIFAR-10 and CIFAR-100 datasets.

In Figure 3a, we plot the Inception Scores [Salimans et al., 2016] of different baselines. Our methods achieve the highest scores throughout training, demonstrating clear practical advantages on this GAN task. In Figure 3b, we further compare Muon and our methods under different numbers of NS steps, which affect the accuracy of the orthogonalization (i.e., the quality of the SVD/polar-factor approximation). Across all NS settings, our methods consistently outperform Muon, indicating that the performance gains persist even when the orthogonalization accuracy varies. We attribute this robustness to our adaptive design, where norm-accumulation-based stepsizes automatically adapt under orthogonal updates, maintaining strong performance across different NS configurations in line with our theoretical insights. Additional results on sensitivity of NS steps, stepsize choices, and the momentum coefficient β, as well as wall-clock time performance, are provided in Section E.

6 Conclusions and Limitations

We proposed **AdaSGDA**, a single-loop method with orthogonal updates for NC–SC minimax optimization. Specifically, our method introduces an adaptive update rule that removes problem-parameter-based stepsize tuning, while stabilizing the coupled primal–dual dynamics inherent to minimax optimization. Through rigorous theoretical analysis, we proved that AdaSGDA achieves the $\mathcal{O}(T^{-1/4})$ convergence guarantee. We further developed **AdaSGDA-VR** by integrating variance reduction, and showed that it attains the faster $\mathcal{O}(T^{-1/3})$ rate, achieving the best-known rate for NC–SC minimax optimization. Extensive experiments validate the effectiveness and scalability of our proposed methods. Nevertheless, despite the problem-parameter-free design of our methods, the theoretical guarantees still rely on horizon-dependent schedules and batch-size requirements. Reducing these dependencies remains an open direction. We aim to address these limitations in future work to further improve the practical flexibility and efficiency of our methods.

Acknowledgements

The work of Zhiwei Zhai, Wenjing Yan, and Ying-Jun Angela Zhang is supported in part by the General Research Fund (project numbers 14214122, 14202723, and 14207624), the Area of Excellence Scheme grant (project number AoE/E-601/22-R), and the NSFC/RGC Collaborative Research Scheme (project numbers CRS_HKUST603/22 and CRS_HKU702/24), all from the Research Grants Council of Hong Kong. The work of Xinyu Wang and Lei Ding is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2026-06465) and a University of Manitoba Start-Up Grant.

References

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on learning theory*, pages 164–194. PMLR, 2019.

Jeremy Bernstein. Deriving muon. `https://jeremybernste.in/writing/deriving-muon`, 2025. Accessed: 2025-07-01.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024. arXiv preprint arXiv:2409.20325 (if applicable); update with venue if published.

Åke Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Vladimir Fedorovich Dem'yanov and Aleksandr Borisovich Pevnyi. Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12(5):11–52, 1972.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Green9. Pytorch code for gan models. `https://github.com/Zeleni9/pytorch-wgan`, 2018. GitHub repository.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

Chuan He, Zhanwang Deng, and Zhaosong Lu. Low-rank orthogonalization for large-scale matrix optimization with applications to foundation model training. *arXiv preprint arXiv:2509.11983*, 2025.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.

Feihu Huang, Xinrui Wang, Junyi Li, and Songcan Chen. Adaptive federated minimax optimization with lower complexities. In *International Conference on Artificial Intelligence and Statistics*, pages 4663–4671. PMLR, 2024a.

Yan Huang, Xiang Li, Yipeng Shen, Niao He, and Jinming Xu. Achieving near-optimal convergence for distributed minimax optimization with adaptive stepsizes. *Advances in Neural Information Processing Systems*, 37:19740–19782, 2024b.

Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step size schedule. In *International conference on machine learning*, pages 14465–14499. PMLR, 2023.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Convergence analysis of adaptive gradient methods under refined smoothness and noise assumptions. *arXiv preprint arXiv:2406.04592*, 2024.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024a.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024b. Accessed: 2025-07-03.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Weiwei Kong and Renato DC Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021.

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Kfir Levy. Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30, 2017.

Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. *Advances in neural information processing systems*, 31, 2018.

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon and further. *arXiv e-prints*, pages arXiv–2502, 2025.

Xiang Li, Junchi Yang, and Niao He. Tiada: A time-scale adaptive algorithm for nonconvex minimax optimization. *arXiv preprint arXiv:2210.17478*, 2022.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International conference on machine learning*, pages 6083–6093. PMLR, 2020.

Tianyi Lin, Chi Jin, and Michael I Jordan. Two-timescale gradient descent ascent algorithms for nonconvex minimax optimization. *Journal of Machine Learning Research*, 26(11):1–45, 2025.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025a.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. Cosmos: A hybrid adaptive optimizer for memory-efficient training of llms. *arXiv preprint arXiv:2502.17410*, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019.

Naoki Marumo and Akiko Takeda. Parameter-free accelerated gradient descent for nonconvex minimization. *SIAM Journal on Optimization*, 34(2):2093–2120, 2024.

H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31 (4):2508–2538, 2021.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Randomized stochastic gradient descent ascent. In *International Conference on Artificial Intelligence and Statistics*, pages 2941–2969. PMLR, 2022.

Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34:24535–24544, 2021.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229, 2018.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Aman Sinha, Hongseok Namkoong, and John C Duchi. Certifiable distributional robustness with principled adversarial training. corr, abs/1710.10571. *arXiv preprint arXiv:1710.10571*, 2017.

Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Wasserstein gans are minimax optimal distribution estimators. *arXiv preprint arXiv:2311.18613*, 2023.

Tijmen Tieleman and G Hinton. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical report*, 2017.

Jingkang Wang, Tianyun Zhang, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, and Bo Li. Adversarial attack generation empowered by min-max optimization. *Advances in Neural Information Processing Systems*, 34:16020–16033, 2021.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear convergence of adaptive stochastic gradient descent. In *International conference on artificial intelligence and statistics*, pages 1475–1485. PMLR, 2020.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in neural information processing systems*, 33:1153–1165, 2020.

Junchi Yang, Xiang Li, and Niao He. Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. *Advances in Neural Information Processing Systems*, 35:11202–11216, 2022a.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International conference on artificial intelligence and statistics*, pages 5485–5517. PMLR, 2022b.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Zhiwei Zhai, Wenjing Yan, and Ying-Jun Zhang. Problem-parameter-free decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 38:144986–145053, 2026.

Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. Adagrad meets muon: Adaptive stepsizes for orthogonal updates. *arXiv preprint arXiv:2509.02981*, 2025.

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

Teng Zhang. Sharp perturbation bounds on the frobenius norm of subunitary and positive polar factor. *arXiv preprint arXiv:2507.14940*, 2025.

Renbo Zhao. A primal-dual smoothing framework for max-structured non-convex optimization. *Mathematics of operations research*, 49(3):1535–1565, 2024.

Taoli Zheng, Linglingzhi Zhu, Anthony Man-Cho So, José Blanchet, and Jiajin Li. Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization. *Advances in neural information processing systems*, 36:54075–54110, 2023.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of machine learning research*, 21(103):1–63, 2020.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135, 2019.

A Related Works

Minimax Optimization. Minimax problems are an important and tractable optimization class with broad interest in optimization and machine learning [Ostrovskii et al., 2021, Kong and Monteiro, 2021, Zhao, 2024]. A classical line of work studies stochastic algorithms with stationarity guarantees; for instance, Rafique et al. [2022] propose methods that rely on nested loops to approximately maximize the dual variable, which can be computationally costly. Motivated by large-scale training, many recent works develop single-loop schemes that enforce time-scale separation implicitly while keeping updates simple [Zheng et al., 2023, Lin et al., 2025]. While more deployable, these methods typically require stepsize choices based on problem-dependent constants (e.g., smoothness, strong concavity, or gradient-noise level), which are often unknown and thus require substantial manual tuning.

To mitigate dependence on problem constants, Yang et al. [2022a] propose a nested adaptive method that achieves parameter-free via an inner loop, and Li et al. [2022] propose a single-loop adaptive method using separated adaptive factors for the primal and dual stepsizes; adaptive minimax optimization in federated and decentralized settings has also been explored in [Huang et al., 2024a,b]. Despite this progress, existing parameter-free guarantees are typically near-optimal, as the rates incur residual terms or logarithmic penalties that prevent matching the best-known scaling with the iteration budget. Recently, orthogonal update techniques have emerged as an effective mechanism for stabilizing optimization [Jordan et al., 2024b]. However, their role in stochastic minimax optimization has not been systematically studied, and a rigorous convergence theory in this setting has yet to be established. This gap is nontrivial because orthogonalization decouples update magnitudes from gradient norms, and the resulting stepsize tuning becomes even more delicate in minimax optimization due to the coupled primal and dual dynamics. To the best of our knowledge, it remains open to design a single-loop, parameter-free minimax method with orthogonal updates that achieves the optimal convergence rate.

Orthogonal Update Techniques. Jordan et al. [2024a] introduced Muon, which orthogonalizes a momentum matrix using a few Newton–Schulz iterations, and reported strong empirical performance at large language model scale [Liu et al., 2025a]. On the theoretical side, Muon has been analyzed under an exact polar step computed via singular value decomposition [Shen et al., 2025], under Frobenius-smoothness with dimension-dependent constants [Li and Hong, 2025], and when the polar step is approximated by Newton–Schulz iterations [He et al., 2025]. However, these guarantees typically rely on stepsize choices that depend on problem-specific parameters to attain the best-known rates, which can hinder practical deployment when such parameters are unknown. Several adaptive variants of Muon have also been explored, including combining Muon with element-wise adaptivity [Si et al., 2025], SOAP [Liu et al., 2025b], and related adaptive mechanisms [Zhang et al., 2025].

While these hybrids often improve performance in practice, existing orthogonal-update methods and theory primarily focus on minimization settings. In contrast, orthogonalized updates for minimax optimization are still largely unexplored, and theoretical guidance is scarce regarding how to choose stepsizes or design parameter-free rules. These limitations motivate a more principled treatment of orthogonalized updates in minimax optimization that is both theoretically grounded and practically robust.

Adaptive Methods. Adaptive gradient methods have played a central role in modern optimization. Starting from AdaGrad [McMahan and Streeter, 2010, Duchi et al., 2011], which introduced coordinate-wise scaling in online convex optimization, a large family of adaptive optimizers has become standard in deep learning, including Adadelta [Zeiler, 2012], RMSprop [Tieleman and Hinton, 2017], and Adam [Kingma, 2014, Luo et al., 2019, Xie et al., 2024]. A particularly relevant line develops parameter-free adaptivity via normalized or self-tuned stepsizes, e.g., AdaNGD [Levy, 2017], AcceleGrad [Levy et al., 2018], and AdaGrad-Norm [Xie et al., 2020, Zhai et al., 2026], which reduce or eliminate reliance on problem constants in stepsize selection. Recent works further clarify the power and limits of adaptive stepsizes through lower bounds and refined designs, including norm-based objectives [Jiang et al., 2024], on-the-fly local curvature estimation [Malitsky and Mishchenko, 2019], and restart-based parameter-free acceleration [Marumo and Takeda, 2024].

Beyond element-wise adaptivity, another line of work exploits the matrix structure of gradients via matrix-aware preconditioning. Shampoo [Gupta et al., 2018] predates Muon and becomes equivalent to Muon when momentum and accumulation are removed. ASGO [An et al., 2025] proposes an adaptive one-sided preconditioner and reduces to Muon under the same simplifications. PolarGrad [Lau et al., 2025] provides a unifying view of matrix-aware preconditioned optimizers and

studies polar-decomposition updates that subsume Muon. More broadly, parameter-free adaptivity and geometry-modifying updates have distinct advantages: parameter-free adaptivity automatically calibrates stepsizes to stochasticity and evolving training dynamics, improving robustness and reducing manual tuning, whereas orthogonalization reshapes the update geometry, often enhancing stability and scalability in large-scale, high-dimensional training. However, developing a single-loop method for stochastic minimax optimization that addresses both aspects remains nontrivial, since the primal–dual coupling amplifies instability and makes stepsize choices particularly delicate. Establishing provable guarantees without relying on problem-dependent tuning under orthogonalized updates therefore remains an important open problem.

B Supporting Lemmas

Lemma B.1 (Smoothness of $\Phi(\cdot)$ and Lipschitzness of $Y^*(\cdot)$. Lemma 4.3 in Lin et al. [2020])**.** *Under Assumptions 4.1 and 4.2, we have that $Y^*(\cdot)$ is λ-Lipschitz with $\lambda = \frac{L}{\mu}$, and $\Phi(\cdot)$ is $L_\Phi = (L + \lambda L)$-smooth with $\nabla\Phi(X) = \nabla_X f(X, Y^*(X))$.*

Lemma B.2 (A logarithmic bound for cumulative ratios)**.** *Let $\{a_t\}_{t=1}^T$ be a sequence of nonnegative numbers with $a_1 > 0$. Define the cumulative sums $A_t := \sum_{i=1}^t a_i$, where $t = 1, \ldots, T$. Then it holds that*

$$\sum_{t=1}^{T} \frac{a_t}{A_t} \;\le\; 1 + \ln\!\left(\frac{A_T}{a_1}\right) \;=\; 1 + \ln\!\left(\frac{\sum_{t=1}^T a_t}{a_1}\right).$$

Proof. Note that $A_1 = a_1 > 0$ and A_t is nondecreasing since $a_t \ge 0$. For any $t \ge 2$, we write

$$\ln(A_t) - \ln(A_{t-1}) = \int_{A_{t-1}}^{A_t} \frac{1}{u}\, du.$$

Since $u \mapsto 1/u$ is decreasing on $(0, \infty)$, $u \in [A_{t-1}, A_t]$ implies $\frac{1}{u} \ge \frac{1}{A_t}$. Therefore,

$$\ln(A_t) - \ln(A_{t-1}) = \int_{A_{t-1}}^{A_t} \frac{1}{u}\, du \;\ge\; \int_{A_{t-1}}^{A_t} \frac{1}{A_t}\, du = \frac{A_t - A_{t-1}}{A_t} = \frac{a_t}{A_t}.$$

Therefore,

$$\sum_{t=1}^{T} \frac{a_t}{A_t} = \frac{a_1}{A_1} + \sum_{t=2}^{T} \frac{a_t}{A_t} \le 1 + \sum_{t=2}^{T} \big(\ln(A_t) - \ln(A_{t-1})\big) = 1 + \ln(A_T) - \ln(A_1),$$

which gives the desired bound since $A_1 = a_1$. □

C Proof of Theorem 4.5

Lemma C.1 (Stationarity Bound)**.** *Suppose Assumptions 4.1 to 4.3 hold, we have:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla\Phi(X_{t-1})\|_*] \\
&\le \frac{\Delta}{T\epsilon_x} + L_\Phi \epsilon_x + \frac{L_\Phi \eta_x^2}{T\epsilon_x}\left(\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1\right) \\
&\quad + \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}[\|M_t^x - \nabla\Phi(X_{t-1})\|_*] + \frac{2\sqrt{r}\eta_x}{T\epsilon_x}\sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1}\sqrt{\sum_{t=1}^{T} \mathbb{E}[\|M_t^x - \nabla\Phi(X_{t-1})\|_F^2]},
\end{aligned}$$

where $\Delta := \Phi(X_0) - \Phi_*$, $\Phi_* := \min_X \Phi(X)$, and $z_0 := \max\{m_0^x, m_0^y\}$.

Proof. Define the scalar function $g : [0,1] \to \mathbb{R}$ by $g(s) := \Phi\big(X_t + s(X_{t+1} - X_t)\big)$. By the fundamental theorem of calculus,

$$\Phi(X_{t+1}) - \Phi(X_t) = g(1) - g(0) = \int_0^1 g'(s)\, ds. \tag{3}$$

By the chain rule, we have

$$g'(s) = \left\langle \nabla\Phi\big(X_t + s(X_{t+1} - X_t)\big),\, X_{t+1} - X_t \right\rangle. \tag{4}$$

Substituting Eq. (4) into Eq. (3), we get

$$\begin{aligned}
&\Phi(X_{t+1}) - \Phi(X_t) \\
&= \langle \nabla\Phi(X_t), X_{t+1} - X_t \rangle + \int_0^1 \left\langle \nabla\Phi\big(X_t + s(X_{t+1} - X_t)\big) - \nabla\Phi(X_t),\, X_{t+1} - X_t \right\rangle ds.
\end{aligned}$$

Using Hölder's inequality $\langle A, B\rangle \le \|A\|_* \, \|B\|_{\mathrm{op}}$, we bound the integrand as

$$\begin{aligned}
&\left\langle \nabla\Phi\big(X_t + s(X_{t+1} - X_t)\big) - \nabla\Phi(X_t),\, X_{t+1} - X_t \right\rangle \\
&\le \left\| \nabla\Phi\big(X_t + s(X_{t+1} - X_t)\big) - \nabla\Phi(X_t) \right\|_* \, \|X_{t+1} - X_t\|_{\mathrm{op}}.
\end{aligned}$$

By the smoothness of Φ in Lemma B.1, we have

$$\left\| \nabla\Phi\big(X_t + s(X_{t+1} - X_t)\big) - \nabla\Phi(X_t) \right\|_* \le L_\Phi \|s(X_{t+1} - X_t)\|_{\mathrm{op}} = sL_\Phi \|X_{t+1} - X_t\|_{\mathrm{op}}.$$

Therefore,

$$\left\langle \nabla\Phi\big(X_t + s(X_{t+1} - X_t)\big) - \nabla\Phi(X_t),\, X_{t+1} - X_t \right\rangle \le sL_\Phi \|X_{t+1} - X_t\|_{\mathrm{op}}^2.$$

Plugging back and integrating yields

$$\begin{aligned}
\Phi(X_{t+1}) - \Phi(X_t) &\le \langle \nabla\Phi(X_t), X_{t+1} - X_t \rangle + \int_0^1 sL_\Phi \|X_{t+1} - X_t\|_{\mathrm{op}}^2 \, ds \\
&\le \langle \nabla\Phi(X_t), X_{t+1} - X_t \rangle + L_\Phi \|X_{t+1} - X_t\|_{\mathrm{op}}^2.
\end{aligned} \tag{5}$$

Then define $\theta_{t+1} := \frac{\min\{\|\nabla_X F(X_t, Y_t; \xi_t^X)\|_F, \gamma\}}{\max\{m_{t+1}^X, m_{t+1}^Y\}}$ and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | X_{t-1}]$ denote the conditional expectation given the previous iterates $X_0, \ldots, X_{t-1}$. By Eq. (5), we have

$$\begin{aligned}
&\mathbb{E}_t[\Phi(X_t) - \Phi(X_{t-1})] \\
&\le \mathbb{E}_t\left[-\langle \nabla\Phi(X_{t-1}), \max\{\epsilon_X, \eta_X\theta_t\} O_t^X \rangle\right] + L_\Phi \mathbb{E}_t\left[\max\{\epsilon_X, \eta_X\theta_t\}^2\right] \\
&\le \mathbb{E}_t\left[-\langle \nabla\Phi(X_{t-1}) - M_t^X, \max\{\epsilon_X, \eta_X\theta_t\} O_t^X \rangle\right] \\
&\quad - \mathbb{E}_t[\max\{\epsilon_X, \eta_X\theta_t\} \|M_t^X\|_*] + L_\Phi\epsilon_X^2 + L_\Phi\eta_X^2 \mathbb{E}_t[[\theta_t]^2] \\
&\overset{(a)}{\le} \mathbb{E}_t[\max\{\epsilon_X, \eta_X\theta_t\} \|\nabla\Phi(X_{t-1}) - M_t^X\|_*] - \mathbb{E}_t[\max\{\epsilon_X, \eta_X\theta_t\} \|M_t^X\|_*] \\
&\quad + L_\Phi\epsilon_X^2 + L_\Phi\eta_X^2 \mathbb{E}_t[[\theta_t]^2] \\
&\le \mathbb{E}_t[2\max\{\epsilon_X, \eta_X\theta_t\} \|\nabla\Phi(X_{t-1}) - M_t^X\|_*] - \mathbb{E}_t[\max\{\epsilon_X, \eta_X\theta_t\} \|\nabla\Phi(X_{t-1})\|_*] \\
&\quad + L_\Phi\epsilon_X^2 + L_\Phi\eta_X^2 \mathbb{E}_t[[\theta_t]^2],
\end{aligned}$$

where (a) is due to $-\langle A, B\rangle \le |\langle A, B\rangle| \le \|A\|_* \, \|B\|_{\mathrm{op}}$. Then by the law of total expectation,

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}[\max\{\epsilon_X, \eta_X\theta_t\} \|\nabla\Phi(X_{t-1})\|_*] \\
&\le \Phi(X_0) - \Phi_* + 2\sum_{t=1}^T \mathbb{E}[\max\{\epsilon_X, \eta_X\theta_t\} \|\nabla\Phi(X_{t-1}) - M_t^X\|_*] + L_\Phi\epsilon_X^2 T + L_\Phi\eta_X^2 \sum_{t=1}^T \mathbb{E}[\theta_t^2] \\
&\le \Phi(X_0) - \Phi_* + 2\epsilon_X \sum_{t=1}^T \mathbb{E}[\|E_t\|_*] + 2\eta_X \sum_{t=1}^T \mathbb{E}[\theta_t \|E_t\|_*] + L_\Phi\epsilon_X^2 T + L_\Phi\eta_X^2 \sum_{t=1}^T \mathbb{E}[\theta_t^2] \\
&\le \Phi(X_0) - \Phi_* + 2\epsilon_X \sum_{t=1}^T \mathbb{E}[\|E_t\|_*] + 2\eta_X \sum_{t=1}^T \sqrt{\mathbb{E}[\|E_t\|_*^2]} \sqrt{\mathbb{E}[\theta_t^2]} + L_\Phi\epsilon_X^2 T + L_\Phi\eta_X^2 \sum_{t=1}^T \mathbb{E}[\theta_t^2] \\
&\le \Phi(X_0) - \Phi_* + 2\epsilon_X \sum_{t=1}^T \mathbb{E}[\|E_t\|_*] + 2\eta_X \sqrt{\sum_{t=1}^T \mathbb{E}[\|E_t\|_*^2]} \sqrt{\sum_{t=1}^T \mathbb{E}[\theta_t^2]}
\end{aligned}$$

$$
\begin{aligned}
&+ L_\Phi \epsilon_x^2 T + L_\Phi \eta_x^2 \sum_{t=1}^{T} \mathbb{E}[\theta_t^2] \\
&\overset{(a)}{\leq} \Phi(X_0) - \Phi_* + 2\epsilon_x \sum_{t=1}^{T} \mathbb{E}[\|E_t\|_*] + 2\eta_x \sqrt{\sum_{t=1}^{T} \mathbb{E}[\|E_t\|_*^2]} \sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1} \\
&\quad + L_\Phi \epsilon_x^2 T + L_\Phi \eta_x^2 \left(\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1\right) \\
&\leq \Phi(X_0) - \Phi_* + 2\epsilon_x \sum_{t=1}^{T} \mathbb{E}[\|E_t\|_*] + 2\eta_x \sqrt{r \sum_{t=1}^{T} \mathbb{E}[\|E_t\|_F^2]} \sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1} \\
&\quad + L_\Phi \epsilon_x^2 T + L_\Phi \eta_x^2 \left(\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1\right),
\end{aligned}
$$

where $E_t := M_t^X - \nabla\Phi(X_{t-1})$ and $\Phi_* := \min_X \Phi(X)$. The last inequality follows from the inequality $\|A\|_* \leq \sqrt{r}\,\|A\|_F$. Moreover, (a) follows from Lemma B.2. Specifically, let $u_t = \min\{\|\nabla_X F(X_{t-1}, Y_{t-1}; \xi_{t-1}^X)\|_F, \gamma\}$ and $S_t = z_0^2 + \sum_{s=1}^{t} u_s^2$. Since $\max\{m_t^X, m_t^Y\}^2 \geq S_t$, we have $\theta_t^2 = \frac{u_t^2}{\max\{m_t^X, m_t^Y\}^2} \leq \frac{u_t^2}{S_t}$. Applying Lemma B.2 yields $\sum_{t=1}^{T} \theta_t^2 \leq \sum_{t=1}^{T} \frac{u_t^2}{S_t} \leq \ln\left(\frac{S_T}{S_0}\right) = \ln\left(1 + \frac{\sum_{t=1}^{T} u_t^2}{z_0^2}\right)$. Finally, using $u_t \leq \gamma$ so that $\sum_{t=1}^{T} u_t^2 \leq \gamma^2 T$ and $\ln(1+v) \leq \ln v + 1$ for $v > 0$, we obtain $\sum_{t=1}^{T} \theta_t^2 \leq \ln\left(1 + \frac{\gamma^2 T}{z_0^2}\right) \leq \ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1$. Then we can get

$$
\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla\Phi(X_{t-1})\|_*] \\
&\leq \frac{1}{T\epsilon_x} \sum_{t=1}^{T} \mathbb{E}[\max\{\epsilon_x, \eta_x \theta_t\} \|\nabla\Phi(X_{t-1})\|_*] \\
&\leq \frac{\Delta}{T\epsilon_x} + \frac{2}{T} \sum_{t=1}^{T} \mathbb{E}[\|E_t\|_*] + \frac{2\sqrt{r}\eta_x}{T\epsilon_x} \sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1} \sqrt{\sum_{t=1}^{T} \mathbb{E}[\|E_t\|_F^2]} \\
&\quad + L_\Phi \epsilon_x + \frac{L_\Phi \eta_x^2}{T\epsilon_x} \left(\ln\left(\frac{\gamma^2 T}{z_0^2}\right) + 1\right),
\end{aligned}
$$

where $\Delta := \Phi(X_0) - \Phi_*$. The proof is thus completed. □

We have the following results on the terms $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_*]$ and $\sum_{t=1}^{T} \mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_F^2]$ in Lemma C.1.

Lemma C.2. *Under Assumptions 4.1 to 4.3, for Algorithm 1, suppose the total iteration rounds is T, the term $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_*]$ is upper bounded by:*

$$
\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_*] \\
&\leq \frac{\kappa\sqrt{r}}{T(1-\beta)\sqrt{b_1}} + \frac{\kappa}{T} \sum_{t=0}^{T-1} \sqrt{\frac{r(1-\beta)}{b_t}} + \frac{\beta L_\Phi \epsilon_X}{1-\beta} + \frac{\beta \eta_X L_\Phi}{(1-\beta)\sqrt{T}} \sqrt{\ln\left(\frac{\gamma^2 T}{\left[\max\{m_0^X, m_0^Y\}\right]^2}\right) + 1} \\
&\quad + L\left(\frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}} \sqrt{\ln\left(\frac{\gamma^2 T}{\left[\max\{m_0^X, m_0^Y\}\right]^2}\right) + 1}\right. \\
&\quad \left. + \frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y} + \frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y} \ln\left(\frac{\gamma^2 T}{\left[\max\{m_0^X, m_0^Y\}\right]^2} + 1\right)\right.
\end{aligned}
$$

$$+\frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)}$$
$$+\frac{Lr\epsilon_Y}{2\mu}+\frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)+\frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)}$$
$$\cdot\left(\frac{3\kappa^2}{T(1-\beta)b_1}+\frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2+\frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right.$$
$$\left.\left.+r\epsilon_Y^2+\frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)+\frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\Bigg), \tag{6}$$

where $\Delta_0^Y:=\Phi(X_0)-f(X_0,Y_0)$.

Lemma C.3. *Under Assumptions 4.1 to 4.3, for Algorithm 1, suppose the total iteration rounds is T, the term $\sum_{t=1}^T\mathbb{E}[\|M_t^X-\nabla\Phi(X_{t-1})\|_F^2]$ can be upper bounded by:*

$$\sum_{t=1}^T\mathbb{E}[\|M_t^X-\nabla\Phi(X_{t-1})\|_F^2]$$
$$\leq\frac{4}{1-\beta^2}\frac{\kappa^2}{b_1}+\frac{4(1-\beta)}{1+\beta}\sum_{t=1}^T\frac{\kappa^2}{b_t}+\frac{4L_\Phi^2r\beta^2\epsilon_X^2T}{1-\beta^2}+\frac{4L_\Phi^2r\eta_X^2\beta^2}{1-\beta^2}T\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)$$
$$+4L^2T(1-\beta)\left[\frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y}+\frac{4Gr\epsilon_X^2T(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G\epsilon_X^2T}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2\epsilon_X^2T}{7\mu^4\epsilon_Y^2}\right.$$
$$+\left(\frac{4Gr\eta_X^2(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G\eta_X^2}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2\eta_X^2}{7\mu^4\epsilon_Y^2}\right)\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)$$
$$+\frac{G\epsilon_YT}{7\mu}+\frac{4GLr\epsilon_YT}{7\mu^2}+\left(\frac{G\eta_Y^2}{7\mu\epsilon_Y}+\frac{4\eta_Y^2GLr}{7\mu^2\epsilon_Y}\right)\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)$$
$$+\frac{16GT\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)}$$
$$\cdot\left(\frac{3\kappa^2}{T(1-\beta)b_1}+\frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2+\frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right.$$
$$\left.\left.\left.+r\epsilon_Y^2+\frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)+\frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\right], \tag{7}$$

where $\Delta_0^Y:=\Phi(X_0)-f(X_0,Y_0)$.

The proof of Lemma C.2 is presented in Section C.1 and the proof of Lemma C.3 is presented in Section C.2. In deriving these results, we will rely on the following lemma.

Lemma C.4. *Under Assumptions 4.1-4.3, for Algorithm 1, suppose the total iteration rounds is T, we have:*

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F]$$
$$\leq\frac{\Delta_0^Y}{T\mu\epsilon_Y}+\frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y}+\frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1}$$
$$+\frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y}+\frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)$$

$$
\begin{aligned}
&+\frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)}\\
&+\frac{Lr\epsilon_Y}{2\mu}+\frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)+\frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)}\\
&\cdot\left(\frac{3\kappa^2}{T(1-\beta)b_1}+\frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2+\frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right.\\
&\left.\left.+r\epsilon_Y^2+\frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)+\frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}},
\end{aligned}
\tag{8}
$$

and

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]\\
&\le\frac{8G\Delta_0^Y}{7\mu^2\epsilon_YT}+\frac{4Gr\epsilon_X^2(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G\epsilon_X^2}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2\epsilon_X^2}{7\mu^4\epsilon_Y^2}\\
&+\left(\frac{4Gr\eta_X^2(L_\Phi+L)}{7T\mu^2\epsilon_Y}+\frac{16r^2L^2G\eta_X^2}{7T\mu^3\epsilon_Y}+\frac{16rG^2L^2\eta_X^2}{7T\mu^4\epsilon_Y^2}\right)\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&+\frac{G\epsilon_Y}{7\mu}+\frac{4GLr\epsilon_Y}{7\mu^2}+\left(\frac{G\eta_Y^2}{7T\mu\epsilon_Y}+\frac{4\eta_Y^2GLr}{7T\mu^2\epsilon_Y}\right)\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\\
&+\frac{16G\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)}\\
&\cdot\left(\frac{3\kappa^2}{T(1-\beta)b_1}+\frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2+\frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right.\\
&\left.\left.+r\epsilon_Y^2+\frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)+\frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}},
\end{aligned}
\tag{9}
$$

where $\Delta_0^Y:=\Phi(X_0)-f(X_0,Y_0)$.

Proof. Let $\sigma_{t+1}:=\frac{\min\{\|\nabla_YF(X_t,Y_t;\xi_t^Y)\|_F,\gamma\}}{m_{t+1}^Y}$, $\Delta_t^Y:=\Phi(X_t)-f(X_t,Y_t)\ge0$, and $\bar{F}_{t+1}^Y:=M_{t+1}^Y-\nabla_Yf(X_t,Y_t)$. Then we have

$$
\begin{aligned}
\langle\nabla_Yf(X_t,Y_t),\,O_{t+1}^Y\rangle&=\langle M_{t+1}^Y,O_{t+1}^Y\rangle-\langle\bar{F}_{t+1}^Y,O_{t+1}^Y\rangle\\
&\overset{(a)}{=}\|M_{t+1}^Y\|_*-\langle\bar{F}_{t+1}^Y,O_{t+1}^Y\rangle\\
&\overset{(b)}{\ge}\|\nabla_Yf(X_t,Y_t)+\bar{F}_{t+1}^Y\|_*-\|\bar{F}_{t+1}^Y\|_*\\
&\overset{(c)}{\ge}\|\nabla_Yf(X_t,Y_t)\|_*-2\|\bar{F}_{t+1}^Y\|_*\\
&\overset{(d)}{\ge}\mu\|Y_t-Y^\star(X_t)\|_F-2\|\bar{F}_{t+1}^Y\|_*,
\end{aligned}
\tag{10}
$$

where (a) uses the orthogonalization rule $O_{t+1}^Y=\mathrm{Orth}(M_{t+1}^Y)$, defined as the Frobenius projection onto the Stiefel manifold $\mathrm{St}(m_y,n_y):=\{Q\in\mathbb{R}^{m_y\times n_y}:Q^\top Q=I_{n_y}\}$, namely

$$
O_{t+1}^Y\in\arg\min_{Q\in\mathrm{St}(m_y,n_y)}\|M_{t+1}^Y-Q\|_F^2.
$$

Since $\|Q\|_F^2=\mathrm{tr}(Q^\top Q)=n_y$ for all $Q\in\mathrm{St}(m_y,n_y)$, we have

$$
\|M_{t+1}^Y-Q\|_F^2=\|M_{t+1}^Y\|_F^2+n_y-2\langle M_{t+1}^Y,Q\rangle,
$$

and hence the above minimization is equivalent to the following inner-product maximization over $\mathrm{St}(m_y, n_y)$:

$$O_{t+1}^Y \in \arg\max_{Q \in \mathrm{St}(m_y,n_y)} \langle M_{t+1}^Y, Q \rangle.$$

Consequently,

$$\langle M_{t+1}^Y, O_{t+1}^Y \rangle = \|M_{t+1}^Y\|_* = \|\nabla_Y f(X_t, Y_t) + \bar{F}_{t+1}^Y\|_*.$$

Moreover, (b) follows from the Hölder inequality $|\langle \bar{F}_{t+1}^Y, O_{t+1}^Y \rangle| \le \|\bar{F}_{t+1}^Y\|_* \, \|O_{t+1}^Y\|_{\mathrm{op}}$, together with $\|O_{t+1}^Y\|_{\mathrm{op}} = 1$ for $O_{t+1}^Y \in \mathrm{St}(m_y, n_y)$, which yields

$$\langle \bar{F}_{t+1}^Y, O_{t+1}^Y \rangle \le |\langle \bar{F}_{t+1}^Y, O_{t+1}^Y \rangle| \le \|\bar{F}_{t+1}^Y\|_*.$$

Inequality (c) follows from the reverse triangle inequality for the nuclear norm, and (d) follows from Assumption 4.2, which implies $\|\nabla_Y f(X_t, Y_t)\|_F \ge \mu \|Y_t - Y^\star(X_t)\|_F$, together with the norm domination $\|\nabla_Y f(X_t, Y_t)\|_* \ge \|\nabla_Y f(X_t, Y_t)\|_F$. For notational simplicity, the above argument is presented for the case $m_y \ge n_y$; the complementary case follows by an analogous argument. In either case, $\|O_{t+1}^Y\|_F^2 = \min\{m_y, n_y\} \le r$ under the global rank bound r. Consequently, by L-smoothness, we obtain

$$\begin{aligned} f(X_t, Y_{t+1}) &\ge f(X_t, Y_t) + \max\{\epsilon_Y, \eta_Y \sigma_t\} \langle \nabla_Y f(X_t, Y_t), O_{t+1}^Y \rangle - \frac{L}{2} [\max\{\epsilon_Y, \eta_Y \sigma_t\}]^2 \|O_{t+1}^Y\|_F^2 \\ &= f(X_t, Y_t) + \max\{\epsilon_Y, \eta_Y \sigma_t\} \langle \nabla_Y f(X_t, Y_t), O_{t+1}^Y \rangle - \frac{Lr}{2} [\max\{\epsilon_Y, \eta_Y \sigma_t\}]^2. \quad (11) \end{aligned}$$

Equivalently, we have

$$\begin{aligned} &\Phi(X_t) - f(X_t, Y_{t+1}) \\ &\le \Phi(X_t) - f(X_t, Y_t) - \max\{\epsilon_Y, \eta_Y \sigma_t\} \langle \nabla_Y f(X_t, Y_t), O_{t+1}^Y \rangle + \frac{Lr}{2} (\max\{\epsilon_Y, \eta_Y \sigma_t\})^2. \quad (12) \end{aligned}$$

Plugging Eq. (10) into Eq. (12) yields

$$\begin{aligned} &\max\{\epsilon_Y, \eta_Y \sigma_t\} \, \mu \|Y_t - Y^\star(X_t)\|_F \\ &\le (\Phi(X_t) - f(X_t, Y_t)) - (\Phi(X_t) - f(X_t, Y_{t+1})) \\ &\quad + \frac{Lr}{2} (\max\{\epsilon_Y, \eta_Y \sigma_t\})^2 + 2 \max\{\epsilon_Y, \eta_Y \sigma_t\} \|\bar{F}_{t+1}^Y\|_*. \quad (13) \end{aligned}$$

Upper Bound for $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F]$. Using $\Delta_t^Y = \Phi(X_t) - f(X_t, Y_t)$ and inserting $\Phi(X_{t+1}) - f(X_{t+1}, Y_{t+1})$, we obtain

$$\begin{aligned} &\big(\Phi(X_t) - f(X_t, Y_t)\big) - \big(\Phi(X_t) - f(X_t, Y_{t+1})\big) \\ &= (\Delta_t^Y - \Delta_{t+1}^Y) + \big(\Phi(X_{t+1}) - \Phi(X_t)\big) - \big(f(X_{t+1}, Y_{t+1}) - f(X_t, Y_{t+1})\big). \quad (14) \end{aligned}$$

Summing Eq. (13) for $t = 0, \ldots, T-1$ and substituting Eq. (14) yields

$$\begin{aligned} &\mu \sum_{t=0}^{T-1} \max\{\epsilon_Y, \eta_Y \sigma_t\} \|Y_t - Y^\star(X_t)\|_F \\ &\le \sum_{t=0}^{T-1} (\Delta_t^Y - \Delta_{t+1}^Y) + \sum_{t=0}^{T-1} \Big[\Phi(X_{t+1}) - \Phi(X_t) - \big(f(X_{t+1}, Y_{t+1}) - f(X_t, Y_{t+1})\big)\Big] \\ &\quad + \frac{Lr}{2} \sum_{t=0}^{T-1} [\max\{\epsilon_Y, \eta_Y \sigma_t\}]^2 + 2 \sum_{t=0}^{T-1} \max\{\epsilon_Y, \eta_Y \sigma_t\} \|\bar{F}_{t+1}^Y\|_* \\ &\le \Delta_0^Y + \sum_{t=0}^{T-1} \Big[\Phi(X_{t+1}) - \Phi(X_t) - \big(f(X_{t+1}, Y_{t+1}) - f(X_t, Y_{t+1})\big)\Big] \\ &\quad + \frac{Lr}{2} \sum_{t=0}^{T-1} [\max\{\epsilon_Y, \eta_Y \sigma_t\}]^2 + 2 \sum_{t=0}^{T-1} \max\{\epsilon_Y, \eta_Y \sigma_t\} \|\bar{F}_{t+1}^Y\|_*, \quad (15) \end{aligned}$$

where the last inequality uses $-\Delta_T^Y \le 0$. We then decompose

$$\begin{aligned}
&\Phi(X_{t+1}) - \Phi(X_t) \\
&= f(X_{t+1}, Y^*(X_{t+1})) - f(X_t, Y^*(X_t)) \\
&= \underbrace{f(X_{t+1}, Y^*(X_t)) - f(X_t, Y^*(X_t))}_{(a)} + \underbrace{f(X_{t+1}, Y^*(X_{t+1})) - f(X_{t+1}, Y^*(X_t))}_{(b)}.
\end{aligned} \tag{16}$$

By Assumption 4.3, unbiasedness, and Jensen's inequality, we have $\|\nabla_Z f(X_t, Y_t)\|_F \le G$ for $Z \in \{X, Y\}$ almost surely. Moreover, strong concavity and $\nabla_Y f(X_t, Y^*(X_t)) = 0$ imply

$$\mu \|Y_t - Y^*(X_t)\|_F \le \|\nabla_Y f(X_t, Y_t)\|_F \le G.$$

For term (a), joint smoothness and the duality between the nuclear and operator norms give

$$\begin{aligned}
&f(X_{t+1}, Y^*(X_t)) - f(X_t, Y^*(X_t)) \\
&\le \left\langle \nabla_X f(X_t, Y^*(X_t)), X_{t+1} - X_t \right\rangle + \frac{L}{2}\|X_{t+1} - X_t\|_{\mathrm{op}}^2 \\
&\le \|\nabla_X f(X_t, Y^*(X_t))\|_* \|X_{t+1} - X_t\|_{\mathrm{op}} + \frac{L}{2}\|X_{t+1} - X_t\|_{\mathrm{op}}^2 \\
&\le \left(\sqrt{r}G + \frac{LG}{\mu}\right)\|X_{t+1} - X_t\|_{\mathrm{op}} + \frac{L}{2}\|X_{t+1} - X_t\|_{\mathrm{op}}^2,
\end{aligned} \tag{17}$$

where the last inequality follows from

$$\begin{aligned}
\|\nabla_X f(X_t, Y^*(X_t))\|_* &\le \|\nabla_X f(X_t, Y_t)\|_* + \|\nabla_X f(X_t, Y^*(X_t)) - \nabla_X f(X_t, Y_t)\|_* \\
&\le \sqrt{r}G + L\|Y^*(X_t) - Y_t\|_{\mathrm{op}} \\
&\le \sqrt{r}G + L\|Y^*(X_t) - Y_t\|_F \\
&\le \sqrt{r}G + \frac{LG}{\mu}.
\end{aligned}$$

Here, we use the global rank bound r and $\|A\|_* \le \sqrt{r}\|A\|_F$.

For term (b), applying smoothness at $(X_{t+1}, Y^*(X_{t+1}))$ and using $\nabla_Y f(X_{t+1}, Y^*(X_{t+1})) = 0$, we obtain

$$\begin{aligned}
f(X_{t+1}, Y^*(X_{t+1})) - f(X_{t+1}, Y^*(X_t)) &\le \frac{L}{2}\|Y^*(X_{t+1}) - Y^*(X_t)\|_{\mathrm{op}}^2 \\
&\le \frac{L\lambda^2}{2}\|X_{t+1} - X_t\|_{\mathrm{op}}^2,
\end{aligned} \tag{18}$$

where the last inequality follows from the λ-Lipschitz continuity of $Y^*(\cdot)$ in Lemma B.1. Combining Eqs. (16) to (18) yields

$$\Phi(X_{t+1}) - \Phi(X_t) \le \left(\sqrt{r}G + \frac{LG}{\mu}\right)\|X_{t+1} - X_t\|_{\mathrm{op}} + \frac{L}{2}(1 + \lambda^2)\|X_{t+1} - X_t\|_{\mathrm{op}}^2. \tag{19}$$

Next, joint smoothness implies

$$\begin{aligned}
\|\nabla_X f(X_t, Y_{t+1})\|_* &\le \|\nabla_X f(X_t, Y_t)\|_* + \|\nabla_X f(X_t, Y_{t+1}) - \nabla_X f(X_t, Y_t)\|_* \\
&\le \sqrt{r}G + L\|Y_{t+1} - Y_t\|_{\mathrm{op}}.
\end{aligned}$$

Consequently, another application of smoothness gives

$$\begin{aligned}
&-\left(f(X_{t+1}, Y_{t+1}) - f(X_t, Y_{t+1})\right) \\
&\le -\left\langle \nabla_X f(X_t, Y_{t+1}), X_{t+1} - X_t \right\rangle + \frac{L}{2}\|X_{t+1} - X_t\|_{\mathrm{op}}^2 \\
&\le \|\nabla_X f(X_t, Y_{t+1})\|_* \|X_{t+1} - X_t\|_{\mathrm{op}} + \frac{L}{2}\|X_{t+1} - X_t\|_{\mathrm{op}}^2 \\
&\le (\sqrt{r}G + L\|Y_{t+1} - Y_t\|_{\mathrm{op}})\|X_{t+1} - X_t\|_{\mathrm{op}} + \frac{L}{2}\|X_{t+1} - X_t\|_{\mathrm{op}}^2.
\end{aligned} \tag{20}$$

Adding Eqs. (19) and (20), we conclude that

$$\begin{aligned}
&\Phi(X_{t+1}) - \Phi(X_t) - (f(X_{t+1}, Y_{t+1}) - f(X_t, Y_{t+1})) \\
&\le G\sqrt{r}\,(2+\lambda)\,\|X_{t+1} - X_t\|_{\mathrm{op}} + L\|X_{t+1} - X_t\|_{\mathrm{op}}\|Y_{t+1} - Y_t\|_{\mathrm{op}} \\
&\quad + \frac{L}{2}(2+\lambda^2)\|X_{t+1} - X_t\|_{\mathrm{op}}^2. \qquad (21)
\end{aligned}$$

Since $\max\{\epsilon_Y, \eta_Y\sigma_t\} \ge \epsilon_Y > 0$, we have $\sum_{t=0}^{T-1} \max\{\epsilon_Y, \eta_Y\sigma_t\}\|Y_t - Y^\star(X_t)\|_F \ge \epsilon_Y \sum_{t=0}^{T-1} \|Y_t - Y^\star(X_t)\|_F$. Taking expectation in Eq. (15) and dividing both sides by $T\mu\epsilon_Y$ yields

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\sqrt{r}(2+\lambda)}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\|X_{t+1} - X_t\|_{\mathrm{op}} + \frac{L}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1} - X_t\|_{\mathrm{op}}\|Y_{t+1} - Y_t\|_{\mathrm{op}}] \\
&\quad + \frac{L(2+\lambda^2)}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1} - X_t\|_{\mathrm{op}}^2] + \frac{Lr}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2] \\
&\quad + \frac{2}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\big[\max\{\epsilon_Y, \eta_Y\sigma_t\}\|\bar{F}_{t+1}^Y\|_*\big] \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\sqrt{r}(2+\lambda)}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\max\{\epsilon_X, \eta_X\theta_t\}] + \frac{L}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\max\{\epsilon_X, \eta_X\theta_t\}\max\{\epsilon_Y, \eta_Y\sigma_t\}] \\
&\quad + \frac{L(2+\lambda^2)}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[(\max\{\epsilon_X, \eta_X\theta_t\})^2] + \frac{Lr}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2] \\
&\quad + \frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]}, \qquad (22)
\end{aligned}$$

where the last inequality follows by applying the Cauchy–Schwarz inequality, together with the bound $\|A\|_* \le \sqrt{r}\,\|A\|_F$.

Upper Bound for $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2]$. According to Assumptions 4.2 and 4.3, we have

$$\mu\|Y_t - Y^\star(X_t)\|_F \le \|\nabla_Y f(X_t, Y_t)\|_F \le G. \qquad (23)$$

Then we have $\mu\|Y_t - Y^\star(X_t)\|_F^2 \le G\|Y_t - Y^\star(X_t)\|_F$. Since $\max\{\epsilon_Y, \eta_Y\sigma_t\} \ge \epsilon_Y > 0$, we can get

$$\mu\|Y_t - Y^\star(X_t)\|_F^2 \le G\|Y_t - Y^\star(X_t)\|_F \le \frac{G}{\mu\epsilon_Y}\big(\max\{\epsilon_Y, \eta_Y\sigma_t\}\mu\|Y_t - Y^\star(X_t)\|_F\big). \qquad (24)$$

Combining Eq. (13) and Eq. (24), we have

$$\begin{aligned}
&\mu\|Y_t - Y^\star(X_t)\|_F^2 \\
&\le \frac{G}{\mu\epsilon_Y}\left[(\Phi(X_t) - f(X_t, Y_t)) - (\Phi(X_t) - f(X_t, Y_{t+1}))\right] + \frac{GLr}{2\mu\epsilon_Y}(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2 \\
&\quad + \frac{2G}{\mu\epsilon_Y}\max\{\epsilon_Y, \eta_Y\sigma_t\}\|\bar{F}_{t+1}^Y\|_*. \qquad (25)
\end{aligned}$$

Taking expectation and summing $t = 0, \ldots, T-1$ gives

$$\mu\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2]$$

$$\leq \frac{G}{\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[(\Phi(X_t)-f(X_t,Y_t))-(\Phi(X_t)-f(X_t,Y_{t+1}))]+\frac{GLr}{2\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]$$
$$+\frac{2G}{\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_Y,\eta_Y\sigma_t\}\|\bar{F}_{t+1}^Y\|_*\right]. \tag{26}$$

Since

$$\begin{aligned}&(\Phi(X_t)-f(X_t,Y_t))-(\Phi(X_t)-f(X_t,Y_{t+1}))\\&=(\Delta_t^Y-\Delta_{t+1}^Y)+\big(\Phi(X_{t+1})-\Phi(X_t)\big)-\big(f(X_{t+1},Y_{t+1})-f(X_t,Y_{t+1})\big),\end{aligned} \tag{27}$$

then

$$\begin{aligned}&\sum_{t=0}^{T-1}[(\Phi(X_t)-f(X_t,Y_t))-(\Phi(X_t)-f(X_t,Y_{t+1}))]\\&=\Delta_0^Y-\Delta_T^Y+\sum_{t=0}^{T-1}\left[\big(\Phi(X_{t+1})-\Phi(X_t)\big)-\big(f(X_{t+1},Y_{t+1})-f(X_t,Y_{t+1})\big)\right]\\&\overset{(a)}{\leq}\Delta_0^Y+\sum_{t=0}^{T-1}\left[\big(\Phi(X_{t+1})-\Phi(X_t)\big)-\big(f(X_{t+1},Y_{t+1})-f(X_t,Y_{t+1})\big)\right],\end{aligned} \tag{28}$$

where (a) is due to $\Delta_T^Y\geq 0$. By smoothness of Φ and $f(\cdot,Y)$, we have

$$\begin{aligned}&\big(\Phi(X_{t+1})-\Phi(X_t)\big)-\big(f(X_{t+1},Y_{t+1})-f(X_t,Y_{t+1})\big)\\&\leq\langle\nabla\Phi(X_t)-\nabla_X f(X_t,Y_{t+1}),X_{t+1}-X_t\rangle+\frac{L_\Phi+L}{2}\|X_{t+1}-X_t\|_F^2\\&\leq\|\nabla\Phi(X_t)-\nabla_X f(X_t,Y_{t+1})\|_F\,\|X_{t+1}-X_t\|_F+\frac{L_\Phi+L}{2}\|X_{t+1}-X_t\|_F^2\\&\leq L\|Y_{t+1}-Y^\star(X_t)\|_F\,\|X_{t+1}-X_t\|_F+\frac{L_\Phi+L}{2}\|X_{t+1}-X_t\|_F^2\\&\leq L\big(\|Y_t-Y^\star(X_t)\|_F+\max\{\epsilon_Y,\eta_Y\sigma_t\}\|O_{t+1}^Y\|_F\big)\|X_{t+1}-X_t\|_F+\frac{L_\Phi+L}{2}\|X_{t+1}-X_t\|_F^2\\&\overset{(a)}{\leq}L\|Y_t-Y^\star(X_t)\|_F\|X_{t+1}-X_t\|_F+L\sqrt{r}\,\max\{\epsilon_Y,\eta_Y\sigma_t\}\|X_{t+1}-X_t\|_F\\&\quad+\frac{L_\Phi+L}{2}\|X_{t+1}-X_t\|_F^2\\&\overset{(b)}{\leq}\frac{\epsilon_Y\mu^2}{8G}\|Y_t-Y^\star(X_t)\|_F^2+\frac{\mu}{8}(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\\&\quad+\left(\frac{L_\Phi+L}{2}+\frac{2rL^2}{\mu}+\frac{2L^2G}{\epsilon_Y\mu^2}\right)\|X_{t+1}-X_t\|_F^2.\end{aligned} \tag{29}$$

where (a) used $\|O_{t+1}^Y\|_F\leq\sqrt{r}$ under the global rank bound r and (b) comes from the Young's inequality: $L\|Y_t-Y^\star(X_t)\|_F\|X_{t+1}-X_t\|_F\leq\frac{\epsilon_Y\mu^2}{8G}\|Y_t-Y^\star(X_t)\|_F^2+\frac{2L^2G}{\epsilon_Y\mu^2}\|X_{t+1}-X_t\|_F^2$ and $L\sqrt{r}\,\max\{\epsilon_Y,\eta_Y\sigma_t\}\|X_{t+1}-X_t\|\leq\frac{\mu}{8}(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2+\frac{2rL^2}{\mu}\|X_{t+1}-X_t\|_F^2$.

Combining Eq. (28) and Eq. (29) gives

$$\begin{aligned}&\sum_{t=0}^{T-1}[(\Phi(X_t)-f(X_t,Y_t))-(\Phi(X_t)-f(X_t,Y_{t+1}))]\\&\leq\Delta_0^Y+\sum_{t=0}^{T-1}\mathbb{E}\left[\big(\Phi(X_{t+1})-\Phi(X_t)\big)-\big(f(X_{t+1},Y_{t+1})-f(X_t,Y_{t+1})\big)\right]\\&\leq\Delta_0^Y+\frac{\epsilon_Y\mu^2}{8G}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]+\frac{\mu}{8}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]\end{aligned}$$

$$+\left(\frac{L_\Phi+L}{2}+\frac{2rL^2}{\mu}+\frac{2L^2G}{\epsilon_Y\mu^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1}-X_t\|_F^2]. \tag{30}$$

Insert Eq. (30) into Eq. (26) to obtain

$$\begin{aligned}
&\mu\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]\\
&\le\frac{G\Delta_0^Y}{\mu\epsilon_Y}+\frac{\mu}{8}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]+\left(\frac{G}{8\epsilon_Y}+\frac{GLr}{2\mu\epsilon_Y}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]\\
&\quad+\left(\frac{G(L_\Phi+L)}{2\mu\epsilon_Y}+\frac{2rL^2G}{\mu^2\epsilon_Y}+\frac{2G^2L^2}{\mu^3\epsilon_Y^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1}-X_t\|_F^2]\\
&\quad+\frac{2G}{\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_Y,\eta_Y\sigma_t\}\|\bar F_{t+1}^Y\|_*\right].
\end{aligned}\tag{31}$$

Then we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]\\
&\le\frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y}+\left(\frac{G}{7\mu\epsilon_Y}+\frac{4GLr}{7\mu^2\epsilon_Y}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]\\
&\quad+\left(\frac{4G(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16rL^2G}{7\mu^3\epsilon_Y}+\frac{16G^2L^2}{7\mu^4\epsilon_Y^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1}-X_t\|_F^2]\\
&\quad+\frac{16G}{7\mu^2\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_Y,\eta_Y\sigma_t\}\|\bar F_{t+1}^Y\|_*\right].
\end{aligned}\tag{32}$$

Dividing by T and using Cauchy–Schwarz with $\|A\|_*\le\sqrt{r}\,\|A\|_F$ yields

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]\\
&\le\frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y T}+\left(\frac{G}{7\mu\epsilon_Y}+\frac{4GLr}{7\mu^2\epsilon_Y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]\\
&\quad+\left(\frac{4G(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16rL^2G}{7\mu^3\epsilon_Y}+\frac{16G^2L^2}{7\mu^4\epsilon_Y^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1}-X_t\|_F^2]\\
&\quad+\frac{16G}{7T\mu^2\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_Y,\eta_Y\sigma_t\}\|\bar F_{t+1}^Y\|_*\right]\\
&\le\frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y T}+\left(\frac{G}{7\mu\epsilon_Y}+\frac{4GLr}{7\mu^2\epsilon_Y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]\\
&\quad+\left(\frac{4Gr(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2}{7\mu^4\epsilon_Y^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_X,\eta_X\theta_t\})^2\right]\\
&\quad+\frac{16G\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar F_{t+1}^Y\|_F^2]}.
\end{aligned}\tag{33}$$

To proceed with the derivations following Eq. (22) and Eq. (33), we will need an upper bound on $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar F_{t+1}^Y\|_F^2]$. We therefore establish such a bound in the following.

Upper Bound for $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]$. We have $\bar{F}_{t+1}^Y = M_{t+1}^Y - \nabla_Y f(X_t, Y_t)$ satisfies

$$\begin{aligned}\bar{F}_{t+1}^Y =&\beta(M_t^Y - \nabla_Y f(X_{t-1}, Y_{t-1})) + \beta(\nabla_Y f(X_{t-1}, Y_{t-1}) - \nabla_Y f(X_t, Y_t))\\ &+ (1-\beta)(\nabla_Y F(X_t, Y_t; \xi_t^Y) - \nabla_Y f(X_t, Y_t)).\end{aligned} \tag{34}$$

Unrolling Eq. (34) yields

$$\begin{aligned}\bar{F}_{t+1}^Y =&\beta^t \bar{F}_1^Y + \sum_{\tau=0}^{t-1}\beta^\tau\Big(\beta\big(\nabla_Y f(X_{t-\tau-1}, Y_{t-\tau-1}) - \nabla_Y f(X_{t-\tau}, Y_{t-\tau})\big)\\ &+ (1-\beta)\big(\nabla_Y F(X_{t-\tau}, Y_{t-\tau}; \xi_{t-\tau}^Y) - \nabla_Y f(X_{t-\tau}, Y_{t-\tau})\big)\Big).\end{aligned} \tag{35}$$

Using $\|A+B+C\|_F^2 \le 3(\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2)$, the weighted Cauchy–Schwarz inequality $\Big\|\sum_{\tau=0}^{t-1}\beta^\tau V_\tau\Big\|_F^2 \le \Big(\sum_{\tau=0}^{t-1}\beta^\tau\Big)\Big(\sum_{\tau=0}^{t-1}\beta^\tau\|V_\tau\|_F^2\Big) \le \frac{1}{1-\beta}\sum_{\tau=0}^{t-1}\beta^\tau\|V_\tau\|_F^2$, and the mini-batch sampling property $\mathbb{E}[\nabla_Y F(X_t, Y_t; \xi_t^Y)] = \nabla_Y f(X_t, Y_t)$ with independent draws across t, we obtain for any $t \ge 0$,

$$\begin{aligned}\mathbb{E}\|\bar{F}_{t+1}^Y\|_F^2 \le 3\beta^{2t}\mathbb{E}\|\bar{F}_1^Y\|_F^2 + \frac{3\beta^2}{1-\beta}\sum_{\tau=0}^{t-1}\beta^\tau\,\mathbb{E}\big[\|\nabla_Y f(X_{t-\tau-1}, Y_{t-\tau-1}) - \nabla_Y f(X_{t-\tau}, Y_{t-\tau})\|_F^2\big]\\ + 3(1-\beta)^2\sum_{\tau=0}^{t-1}\beta^{2\tau}\,\mathbb{E}\big[\|\nabla_Y F(X_{t-\tau}, Y_{t-\tau}; \xi_{t-\tau}^Y) - \nabla_Y f(X_{t-\tau}, Y_{t-\tau})\|_F^2\big].\end{aligned} \tag{36}$$

Summing (36) over $t = 0, \dots, T-1$ and dividing by T, we use the bound $\sum_{t=0}^{T-1}\beta^{2t} \le \frac{1}{1-\beta^2}$ and the geometric-series rearrangement $\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^\tau a_{t-\tau} = \sum_{s=1}^{T-1}\Big(\sum_{j=0}^{T-1-s}\beta^j\Big)a_s \le \frac{1}{1-\beta}\sum_{s=1}^{T-1}a_s$, and similarly, $\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{2\tau}a_{t-\tau} = \sum_{s=1}^{T-1}\Big(\sum_{j=0}^{T-1-s}\beta^{2j}\Big)a_s \le \frac{1}{1-\beta^2}\sum_{s=1}^{T-1}a_s$. This yields

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\bar{F}_{t+1}^Y\|_F^2\\ &\le \frac{3}{T(1-\beta^2)}\mathbb{E}\|\bar{F}_1^Y\|_F^2 + \frac{3\beta^2}{T(1-\beta)^2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\nabla_Y f(X_{t-1}, Y_{t-1}) - \nabla_Y f(X_t, Y_t)\|_F^2\big]\\ &\quad + \frac{3(1-\beta)^2}{T(1-\beta^2)}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla_Y F(X_t, Y_t; \xi_t^Y) - \nabla_Y f(X_t, Y_t)\|_F^2\big]\\ &= \frac{3}{T(1-\beta^2)}\mathbb{E}\|\bar{F}_1^Y\|_F^2 + \frac{3\beta^2}{T(1-\beta)^2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\nabla_Y f(X_{t-1}, Y_{t-1}) - \nabla_Y f(X_t, Y_t)\|_F^2\big]\\ &\quad + \frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla_Y F(X_t, Y_t; \xi_t^Y) - \nabla_Y f(X_t, Y_t)\|_F^2\big]\\ &\overset{(a)}{\le} \frac{3}{T(1-\beta^2)}\mathbb{E}\|\bar{F}_1^Y\|_F^2 + \frac{6\beta^2L^2}{T(1-\beta)^2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|X_t - X_{t-1}\|_F^2 + \|Y_t - Y_{t-1}\|_F^2\big] + \frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\\ &\overset{(b)}{\le} \frac{3\kappa^2}{Tb_1(1-\beta^2)} + \frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\\ &\quad + \frac{6\beta^2L^2r}{(1-\beta)^2}\left(\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\big[(\max\{\epsilon_X, \eta_X\theta_t\})^2\big] + \frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\big[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2\big]\right)\\ &\overset{(c)}{\le} \frac{3\kappa^2}{Tb_1(1-\beta)} + \frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\end{aligned}$$

$$+\frac{6L^2 r}{(1-\beta)^2}\left(\epsilon_X^2+\epsilon_Y^2+\frac{\eta_X^2}{T}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\right), \tag{37}$$

where (a) follows from Assumption 4.1; (b) follows from Assumption 4.3 and the update rules $X_t - X_{t-1} = \max\{\epsilon_X, \eta_X\theta_t\}O_t^X$ and $Y_t - Y_{t-1} = \max\{\epsilon_Y, \eta_Y\sigma_t\}O_t^Y$, together with the bounds $\|O_t^X\|_F^2 \le r$ and $\|O_t^Y\|_F^2 \le r$, where r denotes the global rank upper bound as defined in Section 2.1; and (c) uses Lemma B.2.

By substituting Eq. (37) into Eq. (22), we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\sqrt{r}(2+\lambda)}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\max\{\epsilon_X,\eta_X\theta_t\}] + \frac{L}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\max\{\epsilon_X,\eta_X\theta_t\}\max\{\epsilon_Y,\eta_Y\sigma_t\}] \\
&\quad + \frac{L(2+\lambda^2)}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[(\max\{\epsilon_X,\eta_X\theta_t\})^2] + \frac{Lr}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2] \\
&\quad + \frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]} \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1} \\
&\quad + \frac{Lr\epsilon_Y}{2\mu} + \frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right) + \frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y} + \frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right) \\
&\quad + \frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\right)} \\
&\quad + \frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y,\eta_Y\sigma_t\})^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]} \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1} \\
&\quad + \frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y} + \frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right) \\
&\quad + \frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\right)} \\
&\quad + \frac{Lr\epsilon_Y}{2\mu} + \frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right) + \frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)} \\
&\quad \cdot\left(\frac{3\kappa^2}{T(1-\beta)b_1} + \frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2+\frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right. \\
&\quad \left.\left. + r\epsilon_Y^2+\frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\right) + \frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}.
\end{aligned} \tag{38}$$

Similarly, substituting Eq. (37) into Eq. (33), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2]$$

$$
\begin{aligned}
&\leq \frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y T} + \left(\frac{G}{7\mu\epsilon_Y} + \frac{4GLr}{7\mu^2\epsilon_Y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2\right] \\
&\quad + \left(\frac{4Gr(L_\Phi + L)}{7\mu^2\epsilon_Y} + \frac{16r^2L^2G}{7\mu^3\epsilon_Y} + \frac{16rG^2L^2}{7\mu^4\epsilon_Y^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_X, \eta_X\theta_t\})^2\right] \\
&\quad + \frac{16G}{7\mu^2\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]} \\
&\leq \frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y T} + \left(\frac{G}{7\mu\epsilon_Y} + \frac{4GLr}{7\mu^2\epsilon_Y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2\right] \\
&\quad + \left(\frac{4Gr(L_\Phi + L)}{7\mu^2\epsilon_Y} + \frac{16r^2L^2G}{7\mu^3\epsilon_Y} + \frac{16rG^2L^2}{7\mu^4\epsilon_Y^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_X, \eta_X\theta_t\})^2\right] \\
&\quad + \frac{16G\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2\right]} \\
&\quad \cdot \left(\frac{3\kappa^2}{T(1-\beta)b_1} + \frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2 + \frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} + 1\right)\right.\right. \\
&\quad \left.\left. + r\epsilon_Y^2 + \frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right)\right) + \frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}} \\
&\leq \frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y T} + \frac{4Gr\epsilon_X^2(L_\Phi + L)}{7\mu^2\epsilon_Y} + \frac{16r^2L^2G\epsilon_X^2}{7\mu^3\epsilon_Y} + \frac{16rG^2L^2\epsilon_X^2}{7\mu^4\epsilon_Y^2} \\
&\quad + \left(\frac{4Gr\eta_X^2(L_\Phi + L)}{7T\mu^2\epsilon_Y} + \frac{16r^2L^2G\eta_X^2}{7T\mu^3\epsilon_Y} + \frac{16rG^2L^2\eta_X^2}{7T\mu^4\epsilon_Y^2}\right)\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} + 1\right) \\
&\quad + \frac{G\epsilon_Y}{7\mu} + \frac{4GLr\epsilon_Y}{7\mu^2} + \left(\frac{G\eta_Y^2}{7T\mu\epsilon_Y} + \frac{4\eta_Y^2 GLr}{7T\mu^2\epsilon_Y}\right)\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right) \\
&\quad + \frac{16G\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\epsilon_Y^2 + \frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right)} \\
&\quad \cdot \left(\frac{3\kappa^2}{T(1-\beta)b_1} + \frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2 + \frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} + 1\right)\right.\right. \\
&\quad \left.\left. + r\epsilon_Y^2 + \frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right)\right) + \frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}. \qquad (39)
\end{aligned}
$$

The proof is thus completed. □

C.1 Proof of Lemma C.2

We first define $\tilde{E}_t := \nabla_X F(X_{t-1}, Y_{t-1}; \xi_{t-1}^X) - \nabla_X f(X_{t-1}, Y_{t-1})$ and $E_t := M_t^X - \nabla\Phi(X_{t-1})$ for $t \geq 1$. Then

$$
\begin{aligned}
M_{t+1}^X &= \beta M_t^X + (1-\beta)\nabla_X F(X_t, Y_t; \xi_t^X) \\
&= \beta(E_t + \nabla\Phi(X_{t-1})) + (1-\beta)\left(\tilde{E}_{t+1} + \nabla_X f(X_t, Y_t)\right) \\
&= \nabla_X f(X_t, Y_t)) + \beta(\nabla\Phi(X_{t-1}) - \nabla_X f(X_t, Y_t)) + \beta E_t + (1-\beta)\tilde{E}_{t+1}.
\end{aligned}
$$

Hence, for $t \geq 0$,

$$E_{t+1} = M_{t+1}^X - \nabla\Phi(X_t)$$

$$= \beta E_t + \beta\big(\nabla\Phi(X_{t-1}) - \nabla\Phi(X_t)\big) + (1-\beta)\tilde{E}_{t+1} + (1-\beta)\big(\nabla_X f(X_t, Y_t) - \nabla\Phi(X_t)\big).$$

A recursive formula can be derived as:

$$\begin{aligned} E_{t+1} = \beta^t E_1 + (1-\beta)\sum_{\tau=0}^{t-1}\beta^\tau \tilde{E}_{t+1-\tau} + \beta\sum_{\tau=0}^{t-1}\beta^\tau(\nabla\Phi(X_{t-\tau-1}) - \nabla\Phi(X_{t-\tau})) \\ + (1-\beta)\sum_{\tau=0}^{t-1}\beta^\tau\big(\nabla_X f(X_{t-\tau}, Y_{t-\tau}) - \nabla\Phi(X_{t-\tau})\big). \end{aligned} \tag{40}$$

Assume that, for each $t = 1, \ldots, T$, an independent mini-batch of size $b_t \geq 1$ is sampled. Then, by Lemma B.1, it follows that

$$\begin{aligned} &\mathbb{E}[\|E_{t+1}\|_*] \\ &\leq \beta^t\mathbb{E}[\|E_1\|_*] + (1-\beta)\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^\tau\tilde{E}_{t+1-\tau}\right\|_*\right] + \beta L_\Phi\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\|X_{t-\tau} - X_{t-\tau-1}\|_{\mathrm{op}}] \\ &\quad + (1-\beta)L\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_{\mathrm{op}}] \\ &\overset{(a)}{\leq} \beta^t\mathbb{E}[\|E_1\|_*] + (1-\beta)\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^\tau\tilde{E}_{t+1-\tau}\right\|_*\right] + \beta L_\Phi\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\max\{\epsilon_x, \eta_x\theta_{t-\tau}\}] \\ &\quad + (1-\beta)L\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F] \\ &\overset{(b)}{\leq} \frac{\beta^t\kappa\sqrt{r}}{\sqrt{b_1}} + (1-\beta)\sqrt{r}\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^\tau\tilde{E}_{t+1-\tau}\right\|_F\right] + \beta L_\Phi\epsilon_x\frac{1-\beta^t}{1-\beta} + \beta\eta_x L_\Phi\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\theta_{t-\tau}] \\ &\quad + (1-\beta)L\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F] \\ &\leq \frac{\beta^t\kappa\sqrt{r}}{\sqrt{b_1}} + (1-\beta)\sqrt{r}\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^\tau\tilde{E}_{t+1-\tau}\right\|_F^2\right]^{\frac{1}{2}} + \beta L_\Phi\epsilon_x\frac{1-\beta^t}{1-\beta} + \beta\eta_x L_\Phi\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\theta_{t-\tau}] \\ &\quad + (1-\beta)L\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F] \\ &\overset{(c)}{\leq} \frac{\beta^t\kappa\sqrt{r}}{\sqrt{b_1}} + (1-\beta)\frac{\kappa\sqrt{r}}{\sqrt{b_t}}\sqrt{\sum_{\tau=0}^{t-1}\beta^{2\tau}} + \beta L_\Phi\epsilon_x\frac{1-\beta^t}{1-\beta} + \beta\eta_x L_\Phi\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\theta_{t-\tau}] \\ &\quad + (1-\beta)L\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F] \\ &= \frac{\beta^t\kappa\sqrt{r}}{\sqrt{b_1}} + (1-\beta)\frac{\kappa\sqrt{r}}{\sqrt{b_t}}\sqrt{\frac{1-\beta^{2t}}{1-\beta^2}} + \beta L_\Phi\epsilon_x\frac{1-\beta^t}{1-\beta} + \beta\eta_x L_\Phi\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\theta_{t-\tau}] \\ &\quad + (1-\beta)L\sum_{\tau=0}^{t-1}\beta^\tau\mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F], \end{aligned}$$

where (a) uses $\|A\|_{\mathrm{op}} \leq \|A\|_F$, (b) uses Assumption 4.3 together with the inequality $\|A\|_* \leq \sqrt{r}\,\|A\|_F$, and (c) follows from Assumption 4.3. Therefore,

$$\sum_{t=0}^{T-1}\mathbb{E}[\|E_{t+1}\|_*]$$

$$
\begin{aligned}
&\leq \frac{\kappa\sqrt{r}}{(1-\beta)\sqrt{b_1}} + \sum_{t=0}^{T-1} \kappa \sqrt{\frac{r(1-\beta)}{b_t}} + \frac{T\beta L_\Phi \epsilon_X}{1-\beta} + \beta\eta_X L_\Phi \sum_{t=1}^{T}\sum_{\tau=0}^{t-1} \beta^\tau \mathbb{E}[\theta_{t-\tau}] \\
&\quad + (1-\beta)L \sum_{t=1}^{T}\sum_{\tau=0}^{t-1} \beta^\tau \mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F] \\
&\leq \frac{\kappa\sqrt{r}}{(1-\beta)\sqrt{b_1}} + \sum_{t=0}^{T-1} \kappa \sqrt{\frac{r(1-\beta)}{b_t}} + \frac{T\beta L_\Phi \epsilon_X}{1-\beta} + \frac{\beta\eta_X L_\Phi}{1-\beta} \sum_{t=1}^{T} \mathbb{E}[\theta_t] \\
&\quad + (1-\beta)L \sum_{t=1}^{T}\sum_{\tau=0}^{t-1} \beta^\tau \mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F] \\
&\leq \frac{\kappa\sqrt{r}}{(1-\beta)\sqrt{b_1}} + \sum_{t=0}^{T-1} \kappa \sqrt{\frac{r(1-\beta)}{b_t}} + \frac{T\beta L_\Phi \epsilon_X}{1-\beta} + \frac{\beta\eta_X L_\Phi}{1-\beta} \sqrt{T} \left(\sum_{t=1}^{T} \mathbb{E}[\theta_t^2]\right)^{\frac{1}{2}} \\
&\quad + (1-\beta)L \sum_{t=1}^{T}\sum_{\tau=0}^{t-1} \beta^\tau \mathbb{E}[\|Y_{t-\tau} - Y^*(X_{t-\tau})\|_F] \\
&\leq L \sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^*(X_t)\|_F] + \frac{T\beta L_\Phi \epsilon_X}{1-\beta} + \frac{\beta\eta_X L_\Phi}{1-\beta} \sqrt{T} \sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 1} \\
&\quad + \frac{\kappa\sqrt{r}}{(1-\beta)\sqrt{b_1}} + \sum_{t=0}^{T-1} \kappa \sqrt{\frac{r(1-\beta)}{b_t}},
\end{aligned}
$$

where the last inequality uses Lemma B.2. Then, according to Lemma C.4, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|E_{t+1}\|_*] \\
&\leq \frac{\kappa\sqrt{r}}{T(1-\beta)\sqrt{b_1}} + \frac{\kappa}{T} \sum_{t=0}^{T-1} \sqrt{\frac{r(1-\beta)}{b_t}} + \frac{\beta L_\Phi \epsilon_X}{1-\beta} + \frac{\beta\eta_X L_\Phi}{(1-\beta)\sqrt{T}} \sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 1} \\
&\quad + L\left(\frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}} \sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 1}\right. \\
&\quad + \frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y} + \frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y} \ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} + 1\right) \\
&\quad \left. + \frac{L}{\mu\epsilon_Y} \sqrt{\left(\epsilon_X^2 + \frac{\eta_X^2}{T} \ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} + 1\right)\right)\left(\epsilon_Y^2 + \frac{\eta_Y^2}{T} \ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right)\right)}\right) \\
&\quad + \frac{Lr\epsilon_Y}{2\mu} + \frac{Lr\eta_Y^2}{2T\mu\epsilon_Y} \ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right) + \frac{2\sqrt{r}}{\mu\epsilon_Y} \sqrt{\epsilon_Y^2 + \frac{\eta_Y^2}{T} \ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right)} \\
&\quad \cdot \left(\frac{3\kappa^2}{T(1-\beta)b_1} + \frac{6L^2}{(1-\beta)^2} \left(r\epsilon_X^2 + \frac{r\eta_X^2}{T} \ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} + 1\right)\right.\right. \\
&\quad \left.\left. + r\epsilon_Y^2 + \frac{r\eta_Y^2}{T} \ln\left(\frac{\gamma^2 T}{[m_0^Y]^2} + 1\right)\right) + \frac{3(1-\beta)}{T(1+\beta)} \sum_{t=0}^{T-1} \frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\Bigg).
\end{aligned} \tag{41}
$$

The proof for Lemma C.2 is thus completed.

C.2 Proof of Lemma C.3

According to Eq. (40), we have

$$
\begin{aligned}
&\mathbb{E}\left[\|E_{t+1}\|_F^2\right] \\
&\le 4\beta^{2t}\mathbb{E}\left[\|E_1\|_F^2\right] + 4(1-\beta)^2\sum_{\tau=0}^{t-1}\mathbb{E}\left[\left\|\beta^\tau \tilde{E}_{t+1-\tau}\right\|_F^2\right] \\
&\quad + 4\beta^2\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^\tau(\nabla\Phi(X_{t-\tau-1}) - \nabla\Phi(X_{t-\tau}))\right\|_F^2\right] \\
&\quad + 4(1-\beta)^2\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^\tau(\nabla\Phi(X_{t-\tau}) - \nabla_X f(X_{t-\tau}, Y_{t-\tau}))\right\|_F^2\right] \\
&\le \frac{4\beta^{2t}\kappa^2}{b_1} + 4(1-\beta)^2\frac{\kappa^2}{b_t}\sum_{\tau=0}^{t-1}\beta^{2\tau} + 4\beta^2\mathbb{E}\left[\left(\sum_{\tau=0}^{t-1}\beta^\tau\|\nabla\Phi(X_{t-\tau-1}) - \nabla\Phi(X_{t-\tau})\|_F\right)^2\right] \\
&\quad + 4(1-\beta)^2\mathbb{E}\left[\left(\sum_{\tau=0}^{t-1}\beta^\tau\left\|\nabla\Phi(X_{t-\tau}) - \nabla_X f(X_{t-\tau}, Y_{t-\tau})\right\|_F\right)^2\right].
\end{aligned}
$$

The last inequality follows from Assumption 4.3. Then, by the Cauchy–Schwarz inequality, Assumption 4.1, Lemma B.1, and the definition of the global rank bound r, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|E_{t+1}\|_F^2\right] \\
&\le \frac{4\beta^{2t}\kappa^2}{b_1} + \frac{4\kappa^2}{b_t}\frac{(1-\beta)(1-\beta^{2t})}{1+\beta} + 4L_\Phi^2 r\beta^2\epsilon_X^2\left(\frac{1-\beta^{2t}}{1-\beta^2}\right) \\
&\quad + 4L_\Phi^2 r\eta_X^2\beta^2\left(\sum_{\tau=0}^{t-1}\beta^{2\tau}\right)\mathbb{E}\left[\sum_{\tau=0}^{t-1}\theta_{t-\tau}^2\right] \\
&\quad + 4L^2(1-\beta)^2\left(\sum_{\tau=0}^{t-1}\beta^{2\tau}\right)\mathbb{E}\left[\sum_{\tau=0}^{t-1}\|Y^*(X_{t-\tau}) - Y_{t-\tau}\|_F^2\right] \\
&\le \frac{4\beta^{2t}\kappa^2}{b_1} + \frac{4\kappa^2}{b_t}\frac{(1-\beta)(1-\beta^{2t})}{1+\beta} + 4L_\Phi^2 r\beta^2\epsilon_X^2\left(\frac{1-\beta^{2t}}{1-\beta^2}\right) \\
&\quad + 4L_\Phi^2 r\eta_X^2\beta^2\left(\frac{1-\beta^{2t}}{1-\beta^2}\right)\mathbb{E}\left[\sum_{\tau=0}^{t-1}\theta_{t-\tau}^2\right] \\
&\quad + 4L^2\left(\frac{1-\beta}{1+\beta}\right)\mathbb{E}\left[\sum_{\tau=0}^{t-1}\|Y^*(X_{t-\tau}) - Y_{t-\tau}\|_F^2\right].
\end{aligned} \tag{42}
$$

Then, by applying Lemma 3.2 in Ward et al. [2020], we have

$$
\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\theta_{t-\tau}^2 = \sum_{t=0}^{T-1}\sum_{\tau=1}^{t}\theta_\tau^2 = \sum_{t=1}^{T}(T-t+1)\theta_t^2 \;\le\; T\sum_{t=1}^{T}\theta_t^2 \;\le\; T\ln\left(\frac{\gamma^2 T}{\left[\max\{m_0^X, m_0^Y\}\right]^2} + 1\right). \tag{43}
$$

Summing Eq. (42) over $t = 0, \ldots, T-1$ and combining the result with Eq. (43) and Lemma C.4, we obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left[\|E_{t+1}\|_F^2\right] \\
&\le \frac{4}{1-\beta^2}\frac{\kappa^2}{b_1} + \frac{4(1-\beta)}{1+\beta}\sum_{t=1}^{T}\frac{\kappa^2}{b_t} + \frac{4L_\Phi^2 r\beta^2\epsilon_X^2 T}{1-\beta^2} + \frac{4L_\Phi^2 r\eta_X^2\beta^2}{1-\beta^2}T\ln\left(\frac{\gamma^2 T}{\left[\max\{m_0^X, m_0^Y\}\right]^2} + 1\right)
\end{aligned}
$$

$$
\begin{aligned}
&+4L^2(1-\beta)\sum_{t=1}^{T}\sum_{\tau=0}^{t-1}\mathbb{E}[\|Y^*(X_{t-\tau})-Y_{t-\tau}\|_F^2]\\
&\le \frac{4}{1-\beta^2}\frac{\kappa^2}{b_1}+\frac{4(1-\beta)}{1+\beta}\sum_{t=1}^{T}\frac{\kappa^2}{b_t}+4L^2T(1-\beta)\sum_{t=1}^{T}\mathbb{E}\left[\|Y_t-Y^*(X_t)\|_F^2\right]\\
&\quad+\frac{4L_\Phi^2 r\beta^2\epsilon_X^2 T}{1-\beta^2}+\frac{4L_\Phi^2 r\eta_X^2\beta^2}{1-\beta^2}T\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&\le \frac{4}{1-\beta^2}\frac{\kappa^2}{b_1}+\frac{4(1-\beta)}{1+\beta}\sum_{t=1}^{T}\frac{\kappa^2}{b_t}+\frac{4L_\Phi^2 r\beta^2\epsilon_X^2 T}{1-\beta^2}+\frac{4L_\Phi^2 r\eta_X^2\beta^2}{1-\beta^2}T\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&\quad+4L^2T(1-\beta)\Bigg[\frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y}+\frac{4Gr\epsilon_X^2T(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G\epsilon_X^2T}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2\epsilon_X^2T}{7\mu^4\epsilon_Y^2}\\
&\quad+\left(\frac{4Gr\eta_X^2(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G\eta_X^2}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2\eta_X^2}{7\mu^4\epsilon_Y^2}\right)\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&\quad+\frac{G\epsilon_YT}{7\mu}+\frac{4GLr\epsilon_YT}{7\mu^2}+\left(\frac{G\eta_Y^2}{7\mu\epsilon_Y}+\frac{4\eta_Y^2GLr}{7\mu^2\epsilon_Y}\right)\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\\
&\quad+\frac{16GT\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)}\\
&\quad\cdot\left(\frac{3\kappa^2}{T(1-\beta)b_1}+\frac{6L^2}{(1-\beta)^2}\left(r\epsilon_X^2+\frac{r\eta_X^2}{T}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right.\\
&\quad\left.\left.+r\epsilon_Y^2+\frac{r\eta_Y^2}{T}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)+\frac{3(1-\beta)}{T(1+\beta)}\sum_{t=0}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\Bigg].
\end{aligned}
\tag{44}
$$

The proof for Lemma C.3 is thus completed.

Under the parameter choices $1-\beta=\frac{1}{T^{\frac{1}{4}}}$, $\epsilon_X=\frac{1}{T^{\frac{3}{4}}}$, $\epsilon_Y=\frac{1}{T^{\frac{1}{2}}}$, $\eta_X=\frac{1}{T^{\frac{7}{8}+q}}$, $\eta_Y=\frac{1}{T^{\frac{5}{8}+q}}$, $\gamma>0$, $m_0^X>0$, $m_0^Y>0$, and $b_t\ge T^{\frac{1}{4}}$ for $t\in\{1,\cdots,T\}$, it follows from Lemma C.2 that

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|E_{t+1}\|_*]\\
&\le\frac{\kappa\sqrt{r}}{T^{\frac{7}{8}}}+\frac{\kappa\sqrt{r}}{T^{\frac{1}{4}}}+\frac{L_\Phi}{\sqrt{T}}+\frac{L_\Phi}{T^{\frac{9}{8}+q}}\sqrt{\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1}\\
&\quad+L\left(\frac{\Delta_0^Y}{\sqrt{T}\mu}+\frac{G\sqrt{r}(2+\lambda)}{\mu T^{\frac{1}{4}}}+\frac{G\sqrt{r}(2+\lambda)}{\mu T^{\frac{7}{8}+q}}\sqrt{\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1}\right.\\
&\quad+\frac{L(2+\lambda^2)}{2\mu T}\left(1+\frac{1}{T^{5/4+2q}}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right)\\
&\quad+\frac{L}{\mu T^{3/4}}\sqrt{\left(1+\frac{1}{T^{5/4+2q}}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right)\left(1+\frac{1}{T^{5/4+2q}}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)\right)}\\
&\quad+\frac{Lr}{2\mu\sqrt{T}}+\frac{Lr}{2T^{\frac{7}{4}+2q}\mu}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)+\frac{2\sqrt{rT}}{\mu}\sqrt{\frac{1}{T}+\frac{1}{T^{\frac{9}{4}+2q}}\ln\left(\frac{\gamma^2T}{[m_0^Y]^2}+1\right)}\\
&\quad\cdot\left(\frac{3\kappa^2}{T}+6L^2\sqrt{T}\left(\frac{r}{T^{\frac{3}{2}}}+\frac{r}{T^{\frac{11}{4}+2q}}\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{r}{T}+\frac{r}{T^{\frac{9}{4}+2q}}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\Bigg)+\frac{3\kappa^2}{\sqrt{T}}\Bigg)^{\frac{1}{2}}\Bigg)\\
&\lesssim \frac{\kappa\sqrt{r}}{T^{\frac{1}{4}}}+\frac{LG\sqrt{r}(2+\lambda)+\kappa L\sqrt{r}+L^2 r}{\mu T^{\frac{1}{4}}}+\frac{L_\Phi}{\sqrt{T}}+\frac{\kappa L\sqrt{r}+L^2 r+L\Delta_0^Y}{\mu\sqrt{T}}. \qquad (45)
\end{aligned}
$$

Moreover, Lemma C.3 gives

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\big[\|E_{t+1}\|_F^2\big]\\
&\le 4\kappa^2+4\sqrt{T}\kappa^2+\frac{4L_\Phi^2 r}{T^{\frac{1}{4}}}+\frac{4L_\Phi^2 r}{T^{\frac{1}{2}+2q}}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&\quad+4L^2T^{\frac{3}{4}}\Bigg[\frac{8G\Delta_0^Y\sqrt{T}}{7\mu^2}+\frac{4Gr(L_\Phi+L)}{7\mu^2}+\frac{16r^2L^2G}{7\mu^3}+\frac{16rG^2L^2\sqrt{T}}{7\mu^4}\\
&\quad+\left(\frac{4Gr(L_\Phi+L)}{7\mu^2T^{\frac{5}{4}+2q}}+\frac{16r^2L^2G}{7\mu^3T^{\frac{5}{4}+2q}}+\frac{16rG^2L^2}{7\mu^4T^{\frac{3}{4}+2q}}\right)\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&\quad+\frac{G\sqrt{T}}{7\mu}+\frac{4GLr\sqrt{T}}{7\mu^2}+\left(\frac{G}{7\mu T^{\frac{3}{4}+2q}}+\frac{4GLr}{7\mu^2T^{\frac{3}{4}+2q}}\right)\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\\
&\quad+\frac{16T^{\frac{3}{2}}G\sqrt{r}}{7\mu^2}\sqrt{\frac{1}{T}+\frac{1}{T^{\frac{9}{4}+2q}}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)}\\
&\quad\cdot\left(\frac{3\kappa^2}{T}+6L^2\sqrt{T}\left(\frac{r}{T^{\frac{3}{2}}}+\frac{r}{T^{\frac{11}{4}+2q}}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right.\right.\\
&\quad\left.\left.+\frac{r}{T}+\frac{r}{T^{\frac{9}{4}+2q}}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\right)+\frac{3\kappa^2}{\sqrt{T}}\right)^{\frac{1}{2}}\Bigg]. \qquad (46)
\end{aligned}
$$

Hence, we obtain

$$
\begin{aligned}
&\frac{2\sqrt{r}}{T\epsilon_X}\left(\eta_X\sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1}\right)\sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|E_t\|_F^2]}\\
&\le\frac{2\sqrt{r}}{T^{\frac{9}{8}+q}}\left(\sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1}\right)\\
&\quad\cdot\Bigg[4\kappa^2+4\sqrt{T}\kappa^2+\frac{4L_\Phi^2 r}{T^{\frac{1}{4}}}+\frac{4L_\Phi^2 r}{T^{\frac{1}{2}+2q}}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&\quad+4L^2T^{\frac{3}{4}}\Bigg[\frac{8G\Delta_0^Y\sqrt{T}}{7\mu^2}+\frac{4Gr(L_\Phi+L)}{7\mu^2}+\frac{16r^2L^2G}{7\mu^3}+\frac{16rG^2L^2\sqrt{T}}{7\mu^4}\\
&\quad+\left(\frac{4Gr(L_\Phi+L)}{7\mu^2T^{\frac{5}{4}+2q}}+\frac{16r^2L^2G}{7\mu^3T^{\frac{5}{4}+2q}}+\frac{16rG^2L^2}{7\mu^4T^{\frac{3}{4}+2q}}\right)\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\\
&\quad+\frac{G\sqrt{T}}{7\mu}+\frac{4GLr\sqrt{T}}{7\mu^2}+\left(\frac{G}{7\mu T^{\frac{3}{4}+2q}}+\frac{4GLr}{7\mu^2T^{\frac{3}{4}+2q}}\right)\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\\
&\quad+\frac{16T^{\frac{3}{2}}G\sqrt{r}}{7\mu^2}\sqrt{\frac{1}{T}+\frac{1}{T^{\frac{9}{4}+2q}}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)}\\
&\quad\cdot\left(\frac{3\kappa^2}{T}+6L^2\sqrt{T}\left(\frac{r}{T^{\frac{3}{2}}}+\frac{r}{T^{\frac{11}{4}+2q}}\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}+1\right)\right)\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{r}{T}+\frac{r}{T^{\frac{9}{4}+2q}}\ln\Big(\frac{\gamma^2T}{[m_0^Y]^2}+1\Big)\Big)+\Big(\frac{3\kappa^2}{\sqrt{T}}\Big)^{\frac{1}{2}}\Big]\Big]^{\frac{1}{2}}\\
&\lesssim\Bigg[\frac{L\sqrt{r\Delta_0^Y G}+L^{\frac{3}{2}}r\sqrt{G}+L\sqrt{\kappa G}r^{\frac{3}{4}}+L^{\frac{3}{2}}r\sqrt{G}}{\mu T^{\frac{1}{2}+q}}+\frac{L^2rG}{\mu^2T^{\frac{1}{2}+q}}+\frac{L\sqrt{rG}}{\sqrt{\mu}T^{\frac{1}{2}+q}}\\
&+\frac{L\sqrt{\kappa G}r^{\frac{3}{4}}+L^{\frac{3}{2}}r\sqrt{G}}{\mu T^{\frac{3}{8}+q}}\Bigg]\sqrt{\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\Big)+1}. \qquad (47)
\end{aligned}
$$

Plugging Eq. (45) and Eq. (47) into Lemma C.1, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Phi(X_t)\|_*]\\
&\le\frac{\Delta}{T^{\frac{1}{4}}}+2\Bigg(\frac{\kappa\sqrt{r}}{T^{\frac{7}{8}}}+\frac{\kappa\sqrt{r}}{T^{\frac{1}{4}}}+\frac{L_\Phi}{\sqrt{T}}+\frac{L_\Phi}{T^{\frac{9}{8}+q}}\sqrt{\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\Big)+1}\\
&+L\Bigg(\frac{\Delta_0^Y}{\sqrt{T}\mu}+\frac{G\sqrt{r}(2+\lambda)}{\mu T^{\frac{1}{4}}}+\frac{G\sqrt{r}(2+\lambda)}{\mu T^{\frac{7}{8}+q}}\sqrt{\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\Big)+1}\\
&+\frac{L(2+\lambda^2)}{2\mu T}\Big(1+\frac{1}{T^{5/4+2q}}\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\Big)\Big)\\
&+\frac{L}{\mu T^{3/4}}\sqrt{\Big(1+\frac{1}{T^{5/4+2q}}\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\Big)\Big)\Big(1+\frac{1}{T^{5/4+2q}}\ln\Big(\frac{\gamma^2T}{[m_0^Y]^2}+1\Big)\Big)}\\
&+\frac{Lr}{2\mu\sqrt{T}}+\frac{Lr}{2T^{\frac{7}{4}+2q}\mu}\ln\Big(\frac{\gamma^2T}{[m_0^Y]^2}+1\Big)+\frac{2\sqrt{rT}}{\mu}\sqrt{\frac{1}{T}+\frac{1}{T^{\frac{9}{4}+2q}}\ln\Big(\frac{\gamma^2T}{[m_0^Y]^2}+1\Big)}\\
&\cdot\Bigg(\frac{3\kappa^2}{T}+6L^2\sqrt{T}\Bigg(\frac{r}{T^{\frac{3}{2}}}+\frac{r}{T^{\frac{11}{4}+2q}}\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\Big)\\
&+\frac{r}{T}+\frac{r}{T^{\frac{9}{4}+2q}}\ln\Big(\frac{\gamma^2T}{[m_0^Y]^2}+1\Big)\Big)+\Big(\frac{3\kappa^2}{\sqrt{T}}\Big)^{\frac{1}{2}}\Bigg)\Bigg)\\
&+\frac{2\sqrt{r}}{T^{\frac{9}{8}+q}}\Bigg(\sqrt{\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\Big)+1}\Bigg)\\
&\cdot\Bigg[4\kappa^2+4\sqrt{T}\kappa^2+\frac{4L_\Phi^2r}{T^{\frac{1}{4}}}+\frac{4L_\Phi^2r}{T^{\frac{1}{2}+2q}}\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\Big)\\
&+4L^2T^{\frac{3}{4}}\Bigg[\frac{8G\Delta_0^Y\sqrt{T}}{7\mu^2}+\frac{4Gr(L_\Phi+L)}{7\mu^2}+\frac{16r^2L^2G}{7\mu^3}+\frac{16rG^2L^2\sqrt{T}}{7\mu^4}\\
&+\Big(\frac{4Gr(L_\Phi+L)}{7\mu^2T^{\frac{5}{4}+2q}}+\frac{16r^2L^2G}{7\mu^3T^{\frac{5}{4}+2q}}+\frac{16rG^2L^2}{7\mu^4T^{\frac{3}{4}+2q}}\Big)\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\Big)\\
&+\frac{G\sqrt{T}}{7\mu}+\frac{4GLr\sqrt{T}}{7\mu^2}+\Big(\frac{G}{7\mu T^{\frac{3}{4}+2q}}+\frac{4GLr}{7\mu^2T^{\frac{3}{4}+2q}}\Big)\ln\Big(\frac{\gamma^2T}{[m_0^Y]^2}+1\Big)\\
&+\frac{16T^{\frac{3}{2}}G\sqrt{r}}{7\mu^2}\sqrt{\frac{1}{T}+\frac{1}{T^{\frac{9}{4}+2q}}\ln\Big(\frac{\gamma^2T}{[m_0^Y]^2}+1\Big)}\\
&\cdot\Bigg(\frac{3\kappa^2}{T}+6L^2\sqrt{T}\Bigg(\frac{r}{T^{\frac{3}{2}}}+\frac{r}{T^{\frac{11}{4}+2q}}\ln\Big(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}+1\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{r}{T}+\frac{r}{T^{\frac{9}{4}+2q}}\ln\left(\frac{\gamma^2 T}{[m_0^Y]^2}+1\right)\Bigg)+\frac{3\kappa^2}{\sqrt{T}}\Bigg)^{\frac{1}{2}}\Bigg]\Bigg]^{\frac{1}{2}}\\
&+\frac{L_\Phi}{T^{\frac{3}{4}}}+\frac{L_\Phi}{T^{\frac{5}{4}+2q}}\left(\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1\right)\\
\lesssim\ &\frac{\Delta+\kappa\sqrt{r}}{T^{\frac{1}{4}}}+\frac{LG\sqrt{r}(2+\lambda)+\kappa L\sqrt{r}+L^2 r}{\mu T^{\frac{1}{4}}}+\frac{L_\Phi}{\sqrt{T}}+\frac{\kappa L\sqrt{r}+L^2 r+L\Delta_0^Y}{\mu\sqrt{T}}\\
&+\Bigg[\frac{L\sqrt{r\Delta_0^Y G}+L^{\frac{3}{2}}r\sqrt{G}+L\sqrt{\kappa G}r^{\frac{3}{4}}+L^{\frac{3}{2}}r\sqrt{G}}{\mu T^{\frac{1}{2}+q}}+\frac{L^2 rG}{\mu^2 T^{\frac{1}{2}+q}}+\frac{L\sqrt{rG}}{\sqrt{\mu}T^{\frac{1}{2}+q}}\\
&+\frac{L\sqrt{\kappa G}r^{\frac{3}{4}}+L^{\frac{3}{2}}r\sqrt{G}}{\mu T^{\frac{3}{8}+q}}\Bigg]\sqrt{\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1}\\
\le\ &\mathcal{O}\left(\frac{\Delta+\kappa\sqrt{r}}{T^{\frac{1}{4}}}+\frac{LG\sqrt{r}(1+\lambda)+\kappa L\sqrt{r}+L^2 r}{\mu T^{\frac{1}{4}}}+\frac{Lr^{\frac{3}{4}}\sqrt{\kappa G}+L^{\frac{3}{2}}r\sqrt{G}}{\mu T^{\frac{3}{8}+q}}\sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2}\right)+1}\right),
\end{aligned}
\tag{48}
$$

where $\Delta := \Phi(X_0) - \Phi_*$ and $z_0 := \max\{m_0^X, m_0^Y\}$.

The proof for Theorem 4.5 is thus completed.

Algorithm 2 Adaptive Minimax Optimization with Variance-Reduced Orthogonal Updates (AdaSGDA-VR)

Initialization: X_0, Y_0, $X_{-1} = X_0$, $Y_{-1} = Y_0$, $M_0^X = 0$, $M_0^Y = 0$, buffers $m_0^X = m_0^Y = c > 0$, hyperparameters $\eta_X, \eta_Y, \epsilon_X, \epsilon_Y, \gamma > 0$, and momentum parameter β.
for iteration $t = 0, 1, \cdots, T-1$ **do**
  Sample a mini-batch of size b_t and compute the stochastic gradients: $\nabla_X F(X_t, Y_t; \xi_t^X)$ and $\nabla_Y F(X_t, Y_t; \xi_t^Y)$.
  Momentum updates:
  $M_{t+1}^X = \beta M_t^X + (1-\beta)\nabla_X F(X_t, Y_t; \xi_t^X) + \beta(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X))$,
  $M_{t+1}^Y = \beta M_t^Y + (1-\beta)\nabla_Y F(X_t, Y_t; \xi_t^Y) + \beta(\nabla_Y F(X_t, Y_t; \xi_t^Y) - \nabla_Y F(X_{t-1}, Y_{t-1}; \xi_t^Y))$.
  Accumulate the gradient norms:
  $[m_{t+1}^X]^2 = [m_t^X]^2 + \min\{\|\nabla_X F(X_t, Y_t; \xi_t^X)\|_F^2, \gamma^2\}$,
  $[m_{t+1}^Y]^2 = [m_t^Y]^2 + \min\{\|\nabla_Y F(X_t, Y_t; \xi_t^Y)\|_F^2, \gamma^2\}$.
  Orthogonalize the momentum matrices:
  $O_{t+1}^X = \operatorname{argmin}_{O \in \mathbb{R}^{m_x \times n_x}} \|O - M_{t+1}^X\|_F$, s.t. $O^\top O = I_{n_x}$ if $m_x \geq n_x$, else $OO^\top = I_{m_x}$,
  $O_{t+1}^Y = \operatorname{argmin}_{O \in \mathbb{R}^{m_y \times n_y}} \|O - M_{t+1}^Y\|_F$, s.t. $O^\top O = I_{n_y}$ if $m_y \geq n_y$, else $OO^\top = I_{m_y}$.
  Define adaptive stepsizes:
  $\alpha_{t+1}^Y = \max\left\{\epsilon_Y,\ \eta_Y \frac{\min\{\|\nabla_Y F(X_t, Y_t; \xi_t^Y)\|_F, \gamma\}}{m_{t+1}^Y}\right\}$,
  $\alpha_{t+1}^X = \max\left\{\epsilon_X,\ \eta_X \frac{\min\{\|\nabla_X F(X_t, Y_t; \xi_t^X)\|_F, \gamma\}}{\max\{m_{t+1}^X, m_{t+1}^Y\}}\right\}$.
  Update primal and dual variables:
  $Y_{t+1} = Y_t + \alpha_{t+1}^Y O_{t+1}^Y$,
  $X_{t+1} = X_t - \alpha_{t+1}^X O_{t+1}^X$.
end for

D Proof of Theorem 4.6

According to Lemma C.1, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Phi(X_{t-1})\|_*] \\
&\leq \frac{\Delta}{T\epsilon_X} + L_\Phi\epsilon_X + \frac{L_\Phi\eta_X^2}{T\epsilon_X}\left(\ln\left(\frac{\gamma^2 T}{z_0^2}\right)+1\right) \\
&\quad + \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_*] + \frac{2\sqrt{r}\eta_X}{T\epsilon_X}\sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2}\right)+1}\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_F^2]},
\end{aligned}$$

where $\Delta := \Phi(X_0) - \Phi_*$, $\Phi_* := \min_X \Phi(X)$, and $z_0 := \max\{m_0^X, m_0^Y\}$. Then we derive the following results on the terms $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_*]$ and $\sum_{t=1}^{T}\mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_F^2]$.

Lemma D.1. *Under Assumptions 4.1 to 4.3, for Algorithm 2, suppose the total iteration rounds is T, the term $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_*]$ is upper bounded by:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|M_t^X - \nabla\Phi(X_{t-1})\|_*] \\
&\leq \frac{\beta\sqrt{r}\,G}{T(1-\beta)} + \frac{\kappa\sqrt{r}}{T\sqrt{b_0}} + \sqrt{\frac{r}{1-\beta^2}}\left(\frac{2\beta^2L^2r\epsilon_X^2}{\min_{0\leq s<T} b_s} + \frac{2\beta^2L^2r\eta_X^2}{T\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)\right. \\
&\quad \left. + \frac{2\beta^2L^2r\epsilon_Y^2}{\min_{0\leq s<T} b_s} + \frac{2\beta^2L^2r\eta_Y^2}{T\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{2(1-\beta)^2}{T}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}} \\
&\quad + L\left[\frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)}\right.
\end{aligned}$$

$$
\begin{aligned}
&+\frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y}+\frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&+\frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)}\\
&+\frac{Lr\epsilon_Y}{2\mu}+\frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\\
&\cdot\left(\frac{3\beta^2G^2}{T(1-\beta)}+\frac{3(1-\beta)\kappa^2}{Tb_0}+3\epsilon_X^2L^2r+\frac{3\eta_X^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+3\epsilon_Y^2L^2r\right.\\
&\quad+\frac{3\eta_Y^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6\beta^2\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\\
&\quad+\frac{6\beta^2\eta_X^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+\frac{6\beta^2\epsilon_Y^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\\
&\quad\left.+\frac{6\beta^2\eta_Y^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\Bigg],
\end{aligned}
\tag{49}
$$

where $\Delta_0^Y:=\Phi(X_0)-f(X_0,Y_0)$.

Lemma D.2. *Under Assumptions 4.1 to 4.3, for Algorithm 2, suppose the total iteration rounds is T, the term $\sum_{t=1}^T\mathbb{E}[\|M_t^X-\nabla\Phi(X_{t-1})\|_F^2]$ can be upper bounded by:*

$$
\begin{aligned}
&\sum_{t=1}^T\mathbb{E}[\|M_t^X-\nabla\Phi(X_{t-1})\|_F^2]\\
&\le\frac{2\beta^2G^2}{1-\beta}+\frac{4L^2\Delta_0^Y}{\mu(1-\beta)}+\frac{(1-\beta)\kappa^2}{b_0}\\
&\quad+\left(\frac{3\beta^2L_\Phi^2r}{(1-\beta)^2}+\frac{3\beta^2L^2r}{(1-\beta)^2}+\frac{2\beta^2L^2r}{(1-\beta)\min_{0\le s<T}b_s}\right)\left(T\epsilon_X^2+\eta_X^2\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\\
&\quad+\left(\frac{3\beta^2L^2r}{(1-\beta)^2}+\frac{2\beta^2L^2r}{(1-\beta)\min_{0\le s<T}b_s}\right)\left(T\epsilon_Y^2+\eta_Y^2\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)+2(1-\beta)\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\\
&\quad+\frac{24GL^2\Delta_0^Y}{7\mu^2\epsilon_Y}+\frac{12GrL^2(L_\Phi+L)\epsilon_X^2T}{7\mu^2\epsilon_Y}+\frac{48r^2L^4G\epsilon_X^2T}{7\mu^3\epsilon_Y}+\frac{48rG^2L^4\epsilon_X^2T}{7\mu^4\epsilon_Y^2}\\
&\quad+\left(\frac{12GrL^2(L_\Phi+L)\eta_X^2}{7\mu^2\epsilon_Y}+\frac{48r^2L^4G\eta_X^2}{7\mu^3\epsilon_Y}+\frac{48rG^2L^4\eta_X^2}{7\mu^4\epsilon_Y^2}\right)\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&\quad+\frac{3GL^2\epsilon_YT}{7\mu}+\frac{12GL^3r\epsilon_YT}{7\mu^2}+\left(\frac{3GL^2\eta_Y^2}{7\mu\epsilon_Y}+\frac{12GL^3r\eta_Y^2}{7\mu^2\epsilon_Y}\right)\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\\
&\quad+\frac{48G\sqrt{r}\,L^2T}{7\mu^2\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\cdot\left(\frac{3\beta^2G^2}{T(1-\beta)}+\frac{3(1-\beta)\kappa^2}{Tb_0}+3\epsilon_X^2L^2r\right.\\
&\quad+\frac{3\eta_X^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+3\epsilon_Y^2L^2r+\frac{3\eta_Y^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\\
&\quad+\frac{6\beta^2\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}+\frac{6\beta^2\eta_X^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&\quad\left.+\frac{6\beta^2\epsilon_Y^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}+\frac{6\beta^2\eta_Y^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}},
\end{aligned}
\tag{50}
$$

where $\Delta_0^Y:=\Phi(X_0)-f(X_0,Y_0)$.

The proof of Lemma D.1 is presented in Section D.1 and the proof of Lemma D.2 is presented in Section D.2. In deriving these results, we will rely on the following lemma.

Lemma D.3. *Under Assumptions 4.1 to 4.3, for Algorithm 2, suppose the total iteration rounds is T, we have:*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&\leq \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)} \\
&\quad + \frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y} + \frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) \\
&\quad + \frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2 + \frac{\eta_X^2}{T}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)\right)\left(\epsilon_Y^2 + \frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right)\right)} \\
&\quad + \frac{Lr\epsilon_Y}{2\mu} + \frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2 + \frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right)} \\
&\quad \cdot \left(\frac{3\beta^2 G^2}{T(1-\beta)} + \frac{3(1-\beta)\kappa^2}{Tb_0} + 3\epsilon_X^2 L^2 r + \frac{3\eta_X^2 L^2 r}{T}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 3\epsilon_Y^2 L^2 r\right. \\
&\qquad + \frac{3\eta_Y^2 L^2 r}{T}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6\beta^2\epsilon_X^2 L^2 r}{(1-\beta^2)\min_{0\leq s<T} b_s} \\
&\qquad + \frac{6\beta^2\eta_X^2 L^2 r}{T(1-\beta^2)\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + \frac{6\beta^2\epsilon_Y^2 L^2 r}{(1-\beta^2)\min_{0\leq s<T} b_s} \\
&\qquad \left. + \frac{6\beta^2\eta_Y^2 L^2 r}{T(1-\beta^2)\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}, \qquad (51)
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2] \\
&\leq \frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y T} + \frac{4Gr\epsilon_X^2(L_\Phi + L)}{7\mu^2\epsilon_Y} + \frac{16r^2L^2G\epsilon_X^2}{7\mu^3\epsilon_Y} + \frac{16rG^2L^2\epsilon_X^2}{7\mu^4\epsilon_Y^2} \\
&\quad + \left(\frac{4Gr\eta_X^2(L_\Phi + L)}{7T\mu^2\epsilon_Y} + \frac{16r^2L^2G\eta_X^2}{7T\mu^3\epsilon_Y} + \frac{16rG^2L^2\eta_X^2}{7T\mu^4\epsilon_Y^2}\right)\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) \\
&\quad + \frac{G\epsilon_Y}{7\mu} + \frac{4GLr\epsilon_Y}{7\mu^2} + \left(\frac{G\eta_Y^2}{7T\mu\epsilon_Y} + \frac{4GLr\eta_Y^2}{7T\mu^2\epsilon_Y}\right)\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) \\
&\quad + \frac{16G\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\epsilon_Y^2 + \frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right)} \\
&\quad \cdot \left(\frac{3\beta^2 G^2}{T(1-\beta)} + \frac{3(1-\beta)\kappa^2}{Tb_0} + 3\epsilon_X^2 L^2 r + \frac{3\eta_X^2 L^2 r}{T}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 3\epsilon_Y^2 L^2 r\right. \\
&\qquad + \frac{3\eta_Y^2 L^2 r}{T}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6\beta^2\epsilon_X^2 L^2 r}{(1-\beta^2)\min_{0\leq s<T} b_s} \\
&\qquad + \frac{6\beta^2\eta_X^2 L^2 r}{T(1-\beta^2)\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + \frac{6\beta^2\epsilon_Y^2 L^2 r}{(1-\beta^2)\min_{0\leq s<T} b_s} \\
&\qquad \left. + \frac{6\beta^2\eta_Y^2 L^2 r}{T(1-\beta^2)\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}, \qquad (52)
\end{aligned}
$$

where $\Delta_0^Y := \Phi(X_0) - f(X_0, Y_0)$.

Proof. We first define $\sigma_{t+1} := \frac{\min\{\|\nabla_Y F(X_t,Y_t;\xi_t^Y)\|_F,\gamma\}}{m_{t+1}^Y}$ and $\Delta_t^Y := \Phi(X_t) - f(X_t,Y_t) \geq 0$. Then we use the tracker

$$M_{t+1}^Y = \beta M_t^Y + (1-\beta)\nabla_Y F(X_t,Y_t;\xi_t^Y) + \beta\big(\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y F(X_{t-1},Y_{t-1};\xi_t^Y)\big), \quad (53)$$

and define

$$\tilde{\Delta}_t^Y := \big(\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y f(X_t,Y_t)\big) - \beta\big(\nabla_Y F(X_{t-1},Y_{t-1};\xi_t^Y) - \nabla_Y f(X_{t-1},Y_{t-1})\big). \quad (54)$$

Then we have

$$\begin{aligned}
&M_{t+1}^Y \\
&= \beta M_t^Y + (1-\beta)\nabla_Y F(X_t,Y_t;\xi_t^Y) + \beta\big(\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y F(X_{t-1},Y_{t-1};\xi_t^Y)\big) \\
&= \beta M_t^Y + \nabla_Y f(X_t,Y_t) - \beta\nabla_Y f(X_{t-1},Y_{t-1}) \\
&\quad + \big(\big(\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y f(X_t,Y_t)\big) - \beta\big(\nabla_Y F(X_{t-1},Y_{t-1};\xi_t^Y) - \nabla_Y f(X_{t-1},Y_{t-1})\big)\big) \\
&= \beta M_t^Y + \nabla_Y f(X_t,Y_t) - \beta\nabla_Y f(X_{t-1},Y_{t-1}) + \tilde{\Delta}_t^Y. \quad (55)
\end{aligned}$$

Define $\bar{F}_{t+1}^Y := M_{t+1}^Y - \nabla_Y f(X_t,Y_t)$, so that $M_{t+1}^Y = \nabla_Y f(X_t,Y_t) + \bar{F}_{t+1}^Y$. Under the same orthogonalization property of O_{t+1}^Y and the same smoothness/concavity assumptions as in Lemma C.4, the subsequent analysis follows the same line of reasoning. We first provide an upper bound for $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]$.

Upper Bound for $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]$. Let us denote $\bar{F}_{t+1}^Y := M_{t+1}^Y - \nabla_Y f(X_t,Y_t)$. Then we can get, for $t \geq 1$,

$$\bar{F}_{t+1}^Y = \beta\big(M_t^Y - \nabla_Y f(X_{t-1},Y_{t-1})\big) + \tilde{\Delta}_t^Y = \beta\bar{F}_t^Y + \tilde{\Delta}_t^Y. \quad (56)$$

Iterating Eq. (56) gives, for any $t \geq 0$,

$$\bar{F}_{t+1}^Y = \beta^t \bar{F}_1^Y + \sum_{\tau=0}^{t-1}\beta^\tau\tilde{\Delta}_{t-\tau}^Y. \quad (57)$$

Then we have

$$\begin{aligned}
&\mathbb{E}\big[\|\tilde{\Delta}_t^Y\|_F^2\big] \\
&= \mathbb{E}\Big[\Big\|\big(\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y f(X_t,Y_t)\big) - \beta\big(\nabla_Y F(X_{t-1},Y_{t-1};\xi_t^Y) - \nabla_Y f(X_{t-1},Y_{t-1})\big)\Big\|_F^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-\beta)\big(\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y f(X_t,Y_t)\big) + \beta\Big(\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y F(X_{t-1},Y_{t-1};\xi_t^Y) \\
&\qquad - \nabla_Y f(X_t,Y_t) + \nabla_Y f(X_{t-1},Y_{t-1})\Big)\Big\|_F^2\Big] \\
&\leq 2(1-\beta)^2\mathbb{E}\Big[\big\|\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y f(X_t,Y_t)\big\|_F^2\Big] \\
&\quad + 2\beta^2\mathbb{E}\Big[\Big\|\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y F(X_{t-1},Y_{t-1};\xi_t^Y) - \nabla_Y f(X_t,Y_t) + \nabla_Y f(X_{t-1},Y_{t-1})\Big\|_F^2\Big] \\
&= 2(1-\beta)^2\mathbb{E}\Big[\big\|\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y f(X_t,Y_t)\big\|_F^2\Big] \\
&\quad + \frac{2\beta^2}{b_t}\mathbb{E}\Big[\Big\|\nabla_Y F(X_t,Y_t;\xi_{t,1}^Y) - \nabla_Y F(X_{t-1},Y_{t-1};\xi_{t,1}^Y) - \nabla_Y f(X_t,Y_t) + \nabla_Y f(X_{t-1},Y_{t-1})\Big\|_F^2\Big] \\
&\overset{(a)}{\leq} 2(1-\beta)^2\mathbb{E}\Big[\big\|\nabla_Y F(X_t,Y_t;\xi_t^Y) - \nabla_Y f(X_t,Y_t)\big\|_F^2\Big] \\
&\quad + \frac{2\beta^2}{b_t}\mathbb{E}\Big[\big\|\nabla_Y F(X_t,Y_t;\xi_{t,1}^Y) - \nabla_Y F(X_{t-1},Y_{t-1};\xi_{t,1}^Y)\big\|_F^2\Big] \\
&\overset{(b)}{\leq} \frac{2\beta^2L^2}{b_t}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\beta^2L^2}{b_t}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + \frac{2(1-\beta)^2\kappa^2}{b_t} \\
&\overset{(c)}{\leq} \frac{2\beta^2L^2r}{b_t}\mathbb{E}\left[\big(\max\{\epsilon_X,\eta_X\theta_t\}\big)^2\right] + \frac{2\beta^2L^2r}{b_t}\mathbb{E}\left[\big(\max\{\epsilon_Y,\eta_Y\sigma_t\}\big)^2\right] + \frac{2(1-\beta)^2\kappa^2}{b_t}. \quad (58)
\end{aligned}$$

Inequality (a) follows from the variance decomposition $\mathbb{E}[\|Z\|^2]-\|\mathbb{E}[Z]\|^2 = \mathrm{Var}(Z) \geq 0$, and hence $\mathbb{E}[\|Z\|^2] - \|\mathbb{E}[Z]\|^2 \leq \mathbb{E}[\|Z\|^2]$. This decomposition is applied conditionally on the history before sampling the current mini-batch, and then averaged. Inequality (b) follows from the mean-square smoothness in Assumption 4.1 and the variance bound in Assumption 4.3. Inequality (c) follows from the update forms $X_t - X_{t-1} = -\max\{\epsilon_X, \eta_X\theta_t\}O_t^X$ and $Y_t - Y_{t-1} = \max\{\epsilon_Y, \eta_Y\sigma_t\}O_t^Y$, together with $\|O_t^X\|_F^2 \leq r$ and $\|O_t^Y\|_F^2 \leq r$ under the global rank bound r.

Define $D_t := \sum_{\tau=0}^{t-1} \beta^\tau \big(\nabla_Y f(X_{t-\tau}, Y_{t-\tau}) - \nabla_Y f(X_{t-\tau-1}, Y_{t-\tau-1})\big)$ and $U_t := \sum_{\tau=0}^{t-1} \beta^\tau \tilde{\Delta}_{t-\tau}^Y$. From Eq. (57) and $\|A+B\|^2 \leq 2(\|A\|^2 + \|B\|^2)$, we have

$$\|\bar{F}_{t+1}^Y\|_F^2 \leq 2\beta^{2t}\|\bar{F}_1^Y\|_F^2 + 2\|U_t\|_F^2 \leq 3\beta^{2t}\|\bar{F}_1^Y\|_F^2 + 3(1-\beta)^2\|D_t\|_F^2 + 3\|U_t\|_F^2. \tag{59}$$

Summing the first term on the right-hand side of Eq. (59) over $t = 0, \ldots, T-1$ and taking expectations gives

$$\begin{aligned}
\sum_{t=0}^{T-1} 3\beta^{2t}\,\mathbb{E}[\|\bar{F}_1^Y\|_F^2] &\leq \frac{3}{1-\beta^2}\mathbb{E}[\|\bar{F}_1^Y\|_F^2] \\
&\leq \frac{3}{1-\beta}\mathbb{E}[\|\bar{F}_1^Y\|_F^2] \\
&\leq \frac{3\beta^2 G^2}{1-\beta} + \frac{3(1-\beta)\kappa^2}{b_0}.
\end{aligned} \tag{60}$$

Here the initialization $M_0^Y = 0$, $X_{-1} = X_0$, and $Y_{-1} = Y_0$ gives $M_1^Y = (1-\beta)\nabla_Y F(X_0, Y_0; \xi_0^Y)$. Consequently, conditional unbiasedness and Assumption 4.3 imply

$$\begin{aligned}
\mathbb{E}[\|\bar{F}_1^Y\|_F^2] &= \beta^2\|\nabla_Y f(X_0, Y_0)\|_F^2 + (1-\beta)^2\mathbb{E}\left[\|\nabla_Y F(X_0, Y_0; \xi_0^Y) - \nabla_Y f(X_0, Y_0)\|_F^2\right] \\
&\leq \beta^2 G^2 + \frac{(1-\beta)^2\kappa^2}{b_0}.
\end{aligned}$$

Since we have $D_t := \sum_{\tau=0}^{t-1} \beta^\tau \big(\nabla_Y f(X_{t-\tau}, Y_{t-\tau}) - \nabla_Y f(X_{t-\tau-1}, Y_{t-\tau-1})\big)$, by the weighted Cauchy–Schwarz inequality, $\left\|\sum_{\tau=0}^{t-1} \beta^\tau V_\tau\right\|_F^2 \leq \left(\sum_{\tau=0}^{t-1} \beta^\tau\right)\left(\sum_{\tau=0}^{t-1} \beta^\tau \|V_\tau\|_F^2\right) \leq \frac{1}{1-\beta}\sum_{\tau=0}^{t-1} \beta^\tau \|V_\tau\|_F^2$, we have

$$\mathbb{E}[\|D_t\|_F^2] \leq \frac{1}{1-\beta}\sum_{\tau=0}^{t-1} \beta^\tau\,\mathbb{E}\left[\|\nabla_Y f(X_{t-\tau}, Y_{t-\tau}) - \nabla_Y f(X_{t-\tau-1}, Y_{t-\tau-1})\|_F^2\right]. \tag{61}$$

Summing (61) over $t = 0, \ldots, T-1$ and swapping the order of summation yields

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}[\|D_t\|_F^2] &\leq \frac{1}{1-\beta}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1} \beta^\tau\,\mathbb{E}\left[\|\nabla_Y f(X_{t-\tau}, Y_{t-\tau}) - \nabla_Y f(X_{t-\tau-1}, Y_{t-\tau-1})\|_F^2\right] \\
&= \frac{1}{1-\beta}\sum_{t=1}^{T-1}\left(\sum_{j=0}^{T-1-t} \beta^j\right)\mathbb{E}\left[\|\nabla_Y f(X_t, Y_t) - \nabla_Y f(X_{t-1}, Y_{t-1})\|_F^2\right] \\
&\leq \frac{1}{(1-\beta)^2}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\nabla_Y f(X_t, Y_t) - \nabla_Y f(X_{t-1}, Y_{t-1})\|_F^2\right] \\
&\overset{(a)}{\leq} \frac{L^2}{(1-\beta)^2}\sum_{t=1}^{T-1}\mathbb{E}\left[\|X_t - X_{t-1}\|_F^2 + \|Y_t - Y_{t-1}\|_F^2\right] \\
&\overset{(b)}{\leq} \frac{L^2 r}{(1-\beta)^2}\sum_{t=1}^{T-1}\left(\mathbb{E}\left[\big(\max\{\epsilon_X, \eta_X\theta_t\}\big)^2\right] + \mathbb{E}\left[\big(\max\{\epsilon_Y, \eta_Y\sigma_t\}\big)^2\right]\right) \\
&\overset{(c)}{\leq} \frac{T\epsilon_X^2 L^2 r}{(1-\beta)^2} + \frac{\eta_X^2 L^2 r}{(1-\beta)^2}\ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)
\end{aligned}$$

$$+\frac{T\epsilon_Y^2 L^2 r}{(1-\beta)^2}+\frac{\eta_Y^2 L^2 r}{(1-\beta)^2}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right). \tag{62}$$

Here (a) follows from Assumption 4.1, and (b) uses $\|X_t - X_{t-1}\|_F^2 \le r(\max\{\epsilon_X, \eta_X\theta_t\})^2$ and $\|Y_t - Y_{t-1}\|_F^2 \le r(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2$ under the global rank bound r. For the initialization $m_0^X = m_0^Y = c > 0$, (c) uses Lemma B.2, applied to the sequence beginning with the positive initialization c^2, followed by the clipped squared gradient norms.

By conditional unbiasedness, $\tilde{\Delta}_t^Y$ has zero conditional mean given the history before sampling the current mini-batch. By martingale orthogonality, we have $\mathbb{E}[\|U_t\|_F^2] = \sum_{\tau=0}^{t-1}\beta^{2\tau}\mathbb{E}[\|\tilde{\Delta}_{t-\tau}^Y\|_F^2]$. Moreover, in what follows we invoke the global rank bound r to upper bound $\|X_t - X_{t-1}\|_F^2$ and $\|Y_t - Y_{t-1}\|_F^2$ below. Combining this identity with Lemma B.2 yields

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}[\|U_t\|_F^2]\\
&=\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{2\tau}\mathbb{E}\left[\|\tilde{\Delta}_{t-\tau}^Y\|_F^2\right]\\
&=\sum_{k=1}^{T-1}\left(\sum_{j=0}^{T-1-k}\beta^{2j}\right)\mathbb{E}\left[\|\tilde{\Delta}_k^Y\|_F^2\right]\\
&\le\frac{1}{1-\beta^2}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\tilde{\Delta}_t^Y\|_F^2\right]\\
&\le\frac{2\beta^2L^2}{1-\beta^2}\sum_{t=1}^{T-1}\frac{\mathbb{E}[\|X_t-X_{t-1}\|_F^2]}{b_t}+\frac{2\beta^2L^2}{1-\beta^2}\sum_{t=1}^{T-1}\frac{\mathbb{E}[\|Y_t-Y_{t-1}\|_F^2]}{b_t}\\
&\quad+\frac{2(1-\beta)^2}{1-\beta^2}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\\
&\le\frac{2\beta^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\sum_{t=1}^{T-1}\left(\mathbb{E}\left[\left(\max\{\epsilon_X,\eta_X\theta_t\}\right)^2\right]+\mathbb{E}\left[\left(\max\{\epsilon_Y,\eta_Y\sigma_t\}\right)^2\right]\right)\\
&\quad+\frac{2(1-\beta)^2}{1-\beta^2}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\\
&\le\frac{2\beta^2T\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}+\frac{2\beta^2\eta_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&\quad+\frac{2\beta^2T\epsilon_Y^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}+\frac{2\beta^2\eta_Y^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2(1-\beta)^2}{1-\beta^2}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}.
\end{aligned} \tag{63}$$

Summing Eq. (59) over $t = 0, \dots, T-1$ and dividing by T, and subsequently substituting Eq. (60), Eq. (62), and Eq. (63) into the resulting bound, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]\\
&\le\frac{3\mathbb{E}[\|\bar{F}_1^Y\|_F^2]}{T(1-\beta)}+3\epsilon_X^2L^2r+\frac{3\eta_X^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+3\epsilon_Y^2L^2r\\
&\quad+\frac{3\eta_Y^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6\beta^2\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\\
&\quad+\frac{6\beta^2\eta_X^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{6\beta^2\epsilon_Y^2 L^2 r}{(1-\beta^2)\min_{0\le s<T} b_s} + \frac{6\beta^2\eta_Y^2 L^2 r}{T(1-\beta^2)\min_{0\le s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t} \\
&\le \frac{3\beta^2 G^2}{T(1-\beta)} + \frac{3(1-\beta)\kappa^2}{Tb_0} + 3\epsilon_X^2 L^2 r + \frac{3\eta_X^2 L^2 r}{T}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 3\epsilon_Y^2 L^2 r \\
&\quad + \frac{3\eta_Y^2 L^2 r}{T}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6\beta^2\epsilon_X^2 L^2 r}{(1-\beta^2)\min_{0\le s<T} b_s} \\
&\quad + \frac{6\beta^2\eta_X^2 L^2 r}{T(1-\beta^2)\min_{0\le s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + \frac{6\beta^2\epsilon_Y^2 L^2 r}{(1-\beta^2)\min_{0\le s<T} b_s} \\
&\quad + \frac{6\beta^2\eta_Y^2 L^2 r}{T(1-\beta^2)\min_{0\le s<T} b_s}\ln\left(1+\frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}. \qquad (64)
\end{aligned}
$$

The last inequality follows from Assumption 4.3.

Upper Bound for $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F]$. Substituting Eq. (64) into Eq. (22) and applying the Cauchy–Schwarz inequality, the inequality $\|A\|_* \le \sqrt{r}\,\|A\|_F$, and Lemma B.1, we obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\sqrt{r}(2+\lambda)}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1}-X_t\|_{\mathrm{op}}] + \frac{L}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\|X_{t+1}-X_t\|_{\mathrm{op}}\|Y_{t+1}-Y_t\|_{\mathrm{op}}\right] \\
&\quad + \frac{L(2+\lambda^2)}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}[\|X_{t+1}-X_t\|_{\mathrm{op}}^2] + \frac{Lr}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\right)^2\right] \\
&\quad + \frac{2}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\|\bar{F}_{t+1}^Y\|_*\right] \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\sqrt{r}(2+\lambda)}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_X, \eta_X\theta_{t+1}\}\right] \\
&\quad + \frac{L}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_X, \eta_X\theta_{t+1}\}\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\right] + \frac{L(2+\lambda^2)}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_{t+1}\}\right)^2\right] \\
&\quad + \frac{Lr}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\right)^2\right] + \frac{2\sqrt{r}}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\|\bar{F}_{t+1}^Y\|_F\right] \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\sqrt{r}(2+\lambda)}{T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\max\{\epsilon_X, \eta_X\theta_{t+1}\}\right] \\
&\quad + \frac{L}{\mu\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_{t+1}\}\right)^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\right)^2\right]} \\
&\quad + \frac{L(2+\lambda^2)}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_{t+1}\}\right)^2\right] + \frac{Lr}{2T\mu\epsilon_Y}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\right)^2\right] \\
&\quad + \frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_{t+1}\}\right)^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]} \\
&\le \frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(1+\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y}+\frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&+\frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)}\\
&+\frac{Lr\epsilon_Y}{2\mu}+\frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]}\\
\le\ &\frac{\Delta_0^Y}{T\mu\epsilon_Y}+\frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y}+\frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\\
&+\frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y}+\frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&+\frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)}\\
&+\frac{Lr\epsilon_Y}{2\mu}+\frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\\
&\cdot\left(\frac{3\beta^2G^2}{T(1-\beta)}+\frac{3(1-\beta)\kappa^2}{Tb_0}+3\epsilon_X^2L^2r+\frac{3\eta_X^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+3\epsilon_Y^2L^2r\right.\\
&\quad+\frac{3\eta_Y^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6\beta^2\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\\
&\quad+\frac{6\beta^2\eta_X^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+\frac{6\beta^2\epsilon_Y^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\\
&\quad\left.+\frac{6\beta^2\eta_Y^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}. \qquad (65)
\end{aligned}
$$

Upper Bound for $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]$. Similarly, substituting Eq. (64) into Eq. (33), we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|Y_t-Y^\star(X_t)\|_F^2]\\
\le\ &\frac{8G\Delta_0^Y}{7\mu^2\epsilon_YT}+\left(\frac{G}{7\mu\epsilon_Y}+\frac{4GLr}{7\mu^2\epsilon_Y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y,\eta_Y\sigma_{t+1}\}\right)^2\right]\\
&+\left(\frac{4Gr(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2}{7\mu^4\epsilon_Y^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X,\eta_X\theta_{t+1}\}\right)^2\right]\\
&+\frac{16G\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y,\eta_Y\sigma_{t+1}\}\right)^2\right]}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{F}_{t+1}^Y\|_F^2]}\\
\overset{(a)}{\le}\ &\frac{8G\Delta_0^Y}{7\mu^2\epsilon_YT}+\left(\frac{G}{7\mu\epsilon_Y}+\frac{4GLr}{7\mu^2\epsilon_Y}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y,\eta_Y\sigma_{t+1}\}\right)^2\right]\\
&+\left(\frac{4Gr(L_\Phi+L)}{7\mu^2\epsilon_Y}+\frac{16r^2L^2G}{7\mu^3\epsilon_Y}+\frac{16rG^2L^2}{7\mu^4\epsilon_Y^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X,\eta_X\theta_{t+1}\}\right)^2\right]\\
&+\frac{16G\sqrt{r}}{7\mu^2\epsilon_Y}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y,\eta_Y\sigma_{t+1}\}\right)^2\right]}
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \left(\frac{3\beta^2 G^2}{T(1-\beta)} + \frac{3(1-\beta)\kappa^2}{Tb_0} + 3\epsilon_X^2 L^2 r + \frac{3\eta_X^2 L^2 r}{T} \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 3\epsilon_Y^2 L^2 r \right. \\
&\quad + \frac{3\eta_Y^2 L^2 r}{T} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6\beta^2 \epsilon_X^2 L^2 r}{(1-\beta^2)\min_{0\le s<T} b_s} \\
&\quad + \frac{6\beta^2 \eta_X^2 L^2 r}{T(1-\beta^2)\min_{0\le s<T} b_s} \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + \frac{6\beta^2 \epsilon_Y^2 L^2 r}{(1-\beta^2)\min_{0\le s<T} b_s} \\
&\quad \left. + \frac{6\beta^2 \eta_Y^2 L^2 r}{T(1-\beta^2)\min_{0\le s<T} b_s} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6(1-\beta)}{T(1+\beta)} \sum_{t=1}^{T-1} \frac{\kappa^2}{b_t} \right)^{\frac{1}{2}} \\
&\overset{(b)}{\le} \frac{8G\Delta_0^Y}{7\mu^2\epsilon_Y T} + \frac{4Gr\epsilon_X^2(L_\Phi + L)}{7\mu^2\epsilon_Y} + \frac{16r^2L^2G\epsilon_X^2}{7\mu^3\epsilon_Y} + \frac{16rG^2L^2\epsilon_X^2}{7\mu^4\epsilon_Y^2} \\
&\quad + \left(\frac{4Gr\eta_X^2(L_\Phi + L)}{7T\mu^2\epsilon_Y} + \frac{16r^2L^2G\eta_X^2}{7T\mu^3\epsilon_Y} + \frac{16rG^2L^2\eta_X^2}{7T\mu^4\epsilon_Y^2} \right) \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) \\
&\quad + \frac{G\epsilon_Y}{7\mu} + \frac{4GLr\epsilon_Y}{7\mu^2} + \left(\frac{G\eta_Y^2}{7T\mu\epsilon_Y} + \frac{4GLr\eta_Y^2}{7T\mu^2\epsilon_Y} \right) \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right) \\
&\quad + \frac{16G\sqrt{r}}{7\mu^2\epsilon_Y} \sqrt{\epsilon_Y^2 + \frac{\eta_Y^2}{T} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right)} \\
&\quad \cdot \left(\frac{3\beta^2 G^2}{T(1-\beta)} + \frac{3(1-\beta)\kappa^2}{Tb_0} + 3\epsilon_X^2 L^2 r + \frac{3\eta_X^2 L^2 r}{T} \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + 3\epsilon_Y^2 L^2 r \right. \\
&\quad + \frac{3\eta_Y^2 L^2 r}{T} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6\beta^2 \epsilon_X^2 L^2 r}{(1-\beta^2)\min_{0\le s<T} b_s} \\
&\quad + \frac{6\beta^2 \eta_X^2 L^2 r}{T(1-\beta^2)\min_{0\le s<T} b_s} \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) + \frac{6\beta^2 \epsilon_Y^2 L^2 r}{(1-\beta^2)\min_{0\le s<T} b_s} \\
&\quad \left. + \frac{6\beta^2 \eta_Y^2 L^2 r}{T(1-\beta^2)\min_{0\le s<T} b_s} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right) + \frac{6(1-\beta)}{T(1+\beta)} \sum_{t=1}^{T-1} \frac{\kappa^2}{b_t} \right)^{\frac{1}{2}}. \qquad (66)
\end{aligned}
$$

Here (a) uses Eq. (64). Inequality (b) uses the squared-step bounds above, which follow from Lemma B.2 with the positive initialization. The proof is thus completed. □

D.1 Proof of Lemma D.1

From the update rule, we have

$$
\begin{aligned}
&M_{t+1}^X \\
&= \beta M_t^X + (1-\beta)\nabla_X F(X_t, Y_t; \xi_t^X) + \beta\big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X)\big) \\
&= \beta M_t^X + \nabla_X f(X_t, Y_t) - \beta \nabla_X f(X_{t-1}, Y_{t-1}) \\
&\quad + \big(\big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) - \beta\big(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\big)\big) \\
&= \beta M_t^X + \nabla_X f(X_t, Y_t) - \beta \nabla_X f(X_{t-1}, Y_{t-1}) + \tilde{\Delta}_t, \qquad (67)
\end{aligned}
$$

where

$$
\tilde{\Delta}_t := \big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) - \beta\big(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\big).
$$

Let us define $\tilde{E}_{t+1} := M_{t+1}^X - \nabla_X f(X_t, Y_t)$. It then follows that, for $t \ge 1$,

$$
\tilde{E}_{t+1} = \beta\big(M_t^X - \nabla_X f(X_{t-1}, Y_{t-1})\big) + \tilde{\Delta}_t = \beta \tilde{E}_t + \tilde{\Delta}_t. \qquad (68)
$$

Applying Eq. (68) recursively yields, for any $t \ge 0$, $\tilde{E}_{t+1} = \beta^t \tilde{E}_1 + \sum_{\tau=0}^{t-1} \beta^\tau \tilde{\Delta}_{t-\tau}$. Using $\|A+B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$, we have

$\mathbb{E}[\|\tilde{\Delta}_t\|_F^2]$

$$
\begin{aligned}
&= \mathbb{E}\Big[\Big\|\big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) - \beta\big(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\big)\Big\|_F^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-\beta)\big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) \\
&\qquad + \beta\Big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_t, Y_t) + \nabla_X f(X_{t-1}, Y_{t-1})\Big)\Big\|_F^2\Big] \\
&\le 2(1-\beta)^2 \mathbb{E}\Big[\big\|\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big\|_F^2\Big] \\
&\quad + 2\beta^2 \mathbb{E}\Big[\Big\|\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_t, Y_t) + \nabla_X f(X_{t-1}, Y_{t-1})\Big\|_F^2\Big] \\
&= 2(1-\beta)^2 \mathbb{E}\Big[\big\|\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big\|_F^2\Big] \\
&\quad + \frac{2\beta^2}{b_t} \mathbb{E}\Big[\Big\|\nabla_X F(X_t, Y_t; \xi_{t,1}^X) - \nabla_X F(X_{t-1}, Y_{t-1}; \xi_{t,1}^X) - \nabla_X f(X_t, Y_t) + \nabla_X f(X_{t-1}, Y_{t-1})\Big\|_F^2\Big] \\
&\overset{(a)}{\le} 2(1-\beta)^2 \mathbb{E}\Big[\big\|\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big\|_F^2\Big] \\
&\quad + \frac{2\beta^2}{b_t} \mathbb{E}\Big[\big\|\nabla_X F(X_t, Y_t; \xi_{t,1}^X) - \nabla_X F(X_{t-1}, Y_{t-1}; \xi_{t,1}^X)\big\|_F^2\Big] \\
&\overset{(b)}{\le} \frac{2(1-\beta)^2\kappa^2}{b_t} + \frac{2\beta^2 L^2}{b_t} \mathbb{E}\Big[\|X_t - X_{t-1}\|_F^2 + \|Y_t - Y_{t-1}\|_F^2\Big] \\
&= \frac{2\beta^2 L^2}{b_t} \mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\beta^2 L^2}{b_t} \mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + \frac{2(1-\beta)^2\kappa^2}{b_t} \\
&\overset{(c)}{\le} \frac{2\beta^2 L^2 r}{b_t} \mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X \theta_t\}\right)^2\right] + \frac{2\beta^2 L^2 r}{b_t} \mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y \sigma_t\}\right)^2\right] + \frac{2(1-\beta)^2\kappa^2}{b_t},
\end{aligned} \tag{69}
$$

where (a) follows from $\mathbb{E}[\|Z\|^2] - \|\mathbb{E}[Z]\|^2 \le \mathbb{E}[\|Z\|^2]$, applied conditionally on the history before sampling the current mini-batch; (b) follows from the mean-square smoothness in Assumption 4.1 and the variance bound in Assumption 4.3; and (c) uses the update expressions together with the bounds $\|O_t^X\|_F^2 \le r$ and $\|O_t^Y\|_F^2 \le r$, which hold under the global rank bound r.

Moreover, $\{\tilde{\Delta}_{t+1}\}_{t\ge 0}$ forms a martingale difference sequence with respect to the corresponding pre-sampling histories, so that $\mathbb{E}\langle \tilde{\Delta}_{i+1}, \tilde{\Delta}_{j+1}\rangle = 0$ for $i \ne j$. Consequently, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1} \beta^\tau \tilde{\Delta}_{t-\tau}\right\|_*\right] \\
&\le \sqrt{r}\sqrt{\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1} \beta^\tau \tilde{\Delta}_{t-\tau}\right\|_F^2\right]} \\
&= \sqrt{r}\sqrt{\sum_{\tau=0}^{t-1} \beta^{2\tau} \mathbb{E}[\|\tilde{\Delta}_{t-\tau}\|_F^2]} \\
&\le \sqrt{r}\left(2\beta^2 L^2 r \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X \theta_{t-\tau}\}\right)^2\right] + 2\beta^2 L^2 r \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y \sigma_{t-\tau}\}\right)^2\right]\right. \\
&\qquad \left. + 2(1-\beta)^2\kappa^2 \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}}\right)^{\frac{1}{2}}.
\end{aligned} \tag{70}
$$

Then, by Jensen's inequality, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1} \beta^\tau \tilde{\Delta}_{t-\tau}\right\|_*\right]
$$

$$
\begin{aligned}
&\leq \sqrt{r}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^{\tau}\tilde{\Delta}_{t-\tau}\right\|_F^2\right]} \\
&= \sqrt{r}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{2\tau}\mathbb{E}[\|\tilde{\Delta}_{t-\tau}\|_F^2]} \\
&= \sqrt{r}\sqrt{\frac{1}{T}\sum_{k=1}^{T-1}\left(\sum_{j=0}^{T-1-k}\beta^{2j}\right)\mathbb{E}[\|\tilde{\Delta}_k\|_F^2]} \\
&\leq \sqrt{r}\sqrt{\frac{1}{T(1-\beta^2)}\sum_{t=1}^{T-1}\mathbb{E}[\|\tilde{\Delta}_t\|_F^2]} \\
&= \sqrt{\frac{r}{1-\beta^2}}\sqrt{\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}[\|\tilde{\Delta}_t\|_F^2]} \\
&\leq \sqrt{\frac{r}{1-\beta^2}}\left(\frac{2\beta^2L^2r}{T}\sum_{t=1}^{T-1}\frac{1}{b_t}\mathbb{E}\left[\left(\max\{\epsilon_X,\eta_X\theta_t\}\right)^2\right]\right. \\
&\quad \left.+\frac{2\beta^2L^2r}{T}\sum_{t=1}^{T-1}\frac{1}{b_t}\mathbb{E}\left[\left(\max\{\epsilon_Y,\eta_Y\sigma_t\}\right)^2\right]+\frac{2(1-\beta)^2}{T}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}} \\
&\leq \sqrt{\frac{r}{1-\beta^2}}\left(\frac{2\beta^2L^2r}{T\min_{0\leq s<T}b_s}\sum_{t=1}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X,\eta_X\theta_t\}\right)^2\right]\right. \\
&\quad \left.+\frac{2\beta^2L^2r}{T\min_{0\leq s<T}b_s}\sum_{t=1}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y,\eta_Y\sigma_t\}\right)^2\right]+\frac{2(1-\beta)^2}{T}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}} \\
&\leq \sqrt{\frac{r}{1-\beta^2}}\left(\frac{2\beta^2L^2r\epsilon_X^2}{\min_{0\leq s<T}b_s}+\frac{2\beta^2L^2r\eta_X^2}{T\min_{0\leq s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right. \\
&\quad \left.+\frac{2\beta^2L^2r\epsilon_Y^2}{\min_{0\leq s<T}b_s}+\frac{2\beta^2L^2r\eta_Y^2}{T\min_{0\leq s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2(1-\beta)^2}{T}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}. \qquad (71)
\end{aligned}
$$

The last inequality follows from Lemma B.2, applied to the sequence beginning with the positive initialization c^2 and followed by the clipped squared gradient norms, where $m_0^X = m_0^Y = c > 0$.

Then we have

$$
\begin{aligned}
&\mathbb{E}[\|\tilde{E}_{t+1}\|_*] \\
&= \mathbb{E}\left[\left\|\beta^t\tilde{E}_1+\sum_{\tau=0}^{t-1}\beta^{\tau}\tilde{\Delta}_{t-\tau}\right\|_*\right] \\
&\leq \beta^t\mathbb{E}[\|\tilde{E}_1\|_*]+\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^{\tau}\tilde{\Delta}_{t-\tau}\right\|_*\right] \\
&\leq \beta^t\mathbb{E}[\|\tilde{E}_1\|_*]+\sqrt{r}\sqrt{\mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1}\beta^{\tau}\tilde{\Delta}_{t-\tau}\right\|_F^2\right]} \\
&= \beta^t\mathbb{E}[\|\tilde{E}_1\|_*]+\sqrt{r}\sqrt{\sum_{\tau=0}^{t-1}\beta^{2\tau}\mathbb{E}[\|\tilde{\Delta}_{t-\tau}\|_F^2]}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \beta^t \mathbb{E}[\|\tilde{E}_1\|_*] + \sqrt{r}\left(2\beta^2 L^2 \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}[\|X_{t-\tau} - X_{t-\tau-1}\|_F^2]\right. \\
&\qquad \left. + 2\beta^2 L^2 \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}[\|Y_{t-\tau} - Y_{t-\tau-1}\|_F^2] + 2(1-\beta)^2\kappa^2 \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}}\right)^{\frac{1}{2}} \\
&\leq \beta^t \mathbb{E}[\|\tilde{E}_1\|_*] + \sqrt{r}\left(2\beta^2 L^2 r \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X \theta_{t-\tau}\}\right)^2\right]\right. \\
&\qquad \left. + 2\beta^2 L^2 r \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y \sigma_{t-\tau}\}\right)^2\right] + 2(1-\beta)^2\kappa^2 \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}}\right)^{\frac{1}{2}}. \qquad (72)
\end{aligned}
$$

We now consider $E_{t+1} = M_{t+1}^X - \nabla\Phi(X_t)$. By Assumption 4.1 and Lemma B.1, we obtain

$$
\begin{aligned}
\|E_{t+1}\|_* &= \|M_{t+1}^X - \nabla\Phi(X_t)\|_* \\
&\leq \|M_{t+1}^X - \nabla_X f(X_t, Y_t)\|_* + \|\nabla_X f(X_t, Y_t) - \nabla\Phi(X_t)\|_* \\
&= \underbrace{\|M_{t+1}^X - \nabla_X f(X_t, Y_t)\|_*}_{=\|\tilde{E}_{t+1}\|_*} + \|\nabla_X f(X_t, Y_t) - \nabla_X f(X_t, Y^\star(X_t))\|_* \\
&\leq \|\tilde{E}_{t+1}\|_* + L\|Y_t - Y^\star(X_t)\|_{\mathrm{op}}. \qquad (73)
\end{aligned}
$$

Using $\|A\|_{\mathrm{op}} \leq \|A\|_F$, taking expectations, and applying Eq. (72) yield the following bound:

$$
\begin{aligned}
&\mathbb{E}[\|E_{t+1}\|_*] \\
&\leq \mathbb{E}[\|\tilde{E}_{t+1}\|_*] + L\mathbb{E}[\|Y_t - Y^\star(X_t)\|_{\mathrm{op}}] \\
&\leq \mathbb{E}[\|\tilde{E}_{t+1}\|_*] + L\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&\leq \beta^t \mathbb{E}[\|\tilde{E}_1\|_*] + L\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] + \sqrt{r}\sqrt{\sum_{\tau=0}^{t-1} \beta^{2\tau} \mathbb{E}[\|\tilde{\Delta}_{t-\tau}\|_F^2]} \\
&\leq \beta^t \mathbb{E}[\|\tilde{E}_1\|_*] + L\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] + \sqrt{r}\left(2\beta^2 L^2 r \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X \theta_{t-\tau}\}\right)^2\right]\right. \\
&\qquad \left. + 2\beta^2 L^2 r \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}} \mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y \sigma_{t-\tau}\}\right)^2\right] + 2(1-\beta)^2\kappa^2 \sum_{\tau=0}^{t-1} \frac{\beta^{2\tau}}{b_{t-\tau}}\right)^{\frac{1}{2}}. \qquad (74)
\end{aligned}
$$

We next derive an upper bound on $\sum_{t=0}^{T-1} \mathbb{E}[\|E_{t+1}\|_*]$. Summing Eq. (73) over $t = 0, \ldots, T-1$, using the recursive expansion of $\tilde{E}_{t+1}$, and applying Eqs. (65) and (71), we obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E}[\|E_{t+1}\|_*] \\
&\leq \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{E}_{t+1}\|_*] + L\sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&\leq \sum_{t=0}^{T-1} \beta^t \mathbb{E}[\|\tilde{E}_1\|_*] + \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1} \beta^\tau \tilde{\Delta}_{t-\tau}\right\|_*\right] + L\sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&= \frac{1-\beta^T}{1-\beta} \mathbb{E}[\|\tilde{E}_1\|_*] + \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\sum_{\tau=0}^{t-1} \beta^\tau \tilde{\Delta}_{t-\tau}\right\|_*\right] + L\sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F] \\
&\leq \frac{\mathbb{E}[\|\tilde{E}_1\|_*]}{1-\beta} + T\sqrt{r}\sqrt{\frac{1}{T(1-\beta^2)} \sum_{t=1}^{T-1} \mathbb{E}[\|\tilde{\Delta}_t\|_F^2]} + L\sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{\mathbb{E}[\|\tilde{E}_1\|_*]}{1-\beta} + T\sqrt{\frac{r}{1-\beta^2}}\left(\frac{2\beta^2L^2r\epsilon_X^2}{\min_{0\leq s<T} b_s} + \frac{2\beta^2L^2r\eta_X^2}{T\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right.\\
&\quad\left. + \frac{2\beta^2L^2r\epsilon_Y^2}{\min_{0\leq s<T} b_s} + \frac{2\beta^2L^2r\eta_Y^2}{T\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right) + \frac{2(1-\beta)^2}{T}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\\
&\quad + LT\left[\frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\right.\\
&\quad + \frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y} + \frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&\quad \left. + \frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)}\right.\\
&\quad + \frac{Lr\epsilon_Y}{2\mu} + \frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right) + \frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\\
&\quad \cdot\left(\frac{3\beta^2G^2}{T(1-\beta)} + \frac{3(1-\beta)\kappa^2}{Tb_0} + 3\epsilon_X^2L^2r + \frac{3\eta_X^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right) + 3\epsilon_Y^2L^2r\right.\\
&\qquad + \frac{3\eta_Y^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right) + \frac{6\beta^2\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\leq s<T} b_s}\\
&\qquad + \frac{6\beta^2\eta_X^2L^2r}{T(1-\beta^2)\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right) + \frac{6\beta^2\epsilon_Y^2L^2r}{(1-\beta^2)\min_{0\leq s<T} b_s}\\
&\qquad \left.\left. + \frac{6\beta^2\eta_Y^2L^2r}{T(1-\beta^2)\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right) + \frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\right]. \qquad (75)
\end{aligned}
$$

With the initialization $M_0^X = 0$, $X_{-1} = X_0$, and $Y_{-1} = Y_0$, we have $M_1^X = (1-\beta)\nabla_X F(X_0, Y_0; \xi_0^X)$ and $\tilde{E}_1 = M_1^X - \nabla_X f(X_0, Y_0) = -\beta\nabla_X f(X_0, Y_0) + (1-\beta)\big(\nabla_X F(X_0, Y_0; \xi_0^X) - \nabla_X f(X_0, Y_0)\big)$. Consequently, conditional unbiasedness, $\|A\|_* \leq \sqrt{r}\,\|A\|_F$, Jensen's inequality, and Assumption 4.3 imply

$$
\begin{aligned}
\mathbb{E}[\|\tilde{E}_1\|_*] &\leq \beta\mathbb{E}[\|\nabla_X f(X_0, Y_0)\|_*] + (1-\beta)\mathbb{E}\left[\|\nabla_X F(X_0, Y_0; \xi_0^X) - \nabla_X f(X_0, Y_0)\|_*\right]\\
&\leq \beta\sqrt{r}\,\mathbb{E}[\|\nabla_X f(X_0, Y_0)\|_F] + (1-\beta)\sqrt{r}\sqrt{\mathbb{E}\left[\|\nabla_X F(X_0, Y_0; \xi_0^X) - \nabla_X f(X_0, Y_0)\|_F^2\right]}\\
&\leq \beta\sqrt{r}\,G + \frac{(1-\beta)\kappa\sqrt{r}}{\sqrt{b_0}}.
\end{aligned}
$$

Therefore, $\frac{\mathbb{E}[\|\tilde{E}_1\|_*]}{T(1-\beta)} \leq \frac{\beta\sqrt{r}\,G}{T(1-\beta)} + \frac{\kappa\sqrt{r}}{T\sqrt{b_0}}$. Dividing both sides of Eq. (75) by T and using the initialization bound above, we obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|E_{t+1}\|_*]\\
&\leq \frac{\beta\sqrt{r}\,G}{T(1-\beta)} + \frac{\kappa\sqrt{r}}{T\sqrt{b_0}} + \sqrt{\frac{r}{1-\beta^2}}\left(\frac{2\beta^2L^2r\epsilon_X^2}{\min_{0\leq s<T} b_s} + \frac{2\beta^2L^2r\eta_X^2}{T\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right.\\
&\quad\left. + \frac{2\beta^2L^2r\epsilon_Y^2}{\min_{0\leq s<T} b_s} + \frac{2\beta^2L^2r\eta_Y^2}{T\min_{0\leq s<T} b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right) + \frac{2(1-\beta)^2}{T}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\\
&\quad + L\left[\frac{\Delta_0^Y}{T\mu\epsilon_Y} + \frac{G\epsilon_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y} + \frac{G\eta_X\sqrt{r}(2+\lambda)}{\mu\epsilon_Y\sqrt{T}}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\right.\\
&\quad + \frac{L\epsilon_X^2(2+\lambda^2)}{2\mu\epsilon_Y} + \frac{L\eta_X^2(2+\lambda^2)}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{L}{\mu\epsilon_Y}\sqrt{\left(\epsilon_X^2+\frac{\eta_X^2}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\left(\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)}\\
&+\frac{Lr\epsilon_Y}{2\mu}+\frac{Lr\eta_Y^2}{2T\mu\epsilon_Y}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2\sqrt{r}}{\mu\epsilon_Y}\sqrt{\epsilon_Y^2+\frac{\eta_Y^2}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\\
&\cdot\left(\frac{3\beta^2G^2}{T(1-\beta)}+\frac{3(1-\beta)\kappa^2}{Tb_0}+3\epsilon_X^2L^2r+\frac{3\eta_X^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+3\epsilon_Y^2L^2r\right.\\
&\quad+\frac{3\eta_Y^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6\beta^2\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\\
&\quad+\frac{6\beta^2\eta_X^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+\frac{6\beta^2\epsilon_Y^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}\\
&\quad\left.+\frac{6\beta^2\eta_Y^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\right)^{\frac{1}{2}}\Bigg].
\end{aligned}
\tag{76}
$$

The proof for Lemma D.1 is thus completed.

D.2 Proof of Lemma D.2

Starting from the decomposition

$$
\begin{aligned}
&M_{t+1}^X\\
&=\beta M_t^X+\nabla_X f(X_t,Y_t)-\beta\nabla_X f(X_{t-1},Y_{t-1})\\
&\quad+\left[\left(\nabla_X F(X_t,Y_t;\xi_t^X)-\nabla_X f(X_t,Y_t)\right)-\beta\left(\nabla_X F(X_{t-1},Y_{t-1};\xi_t^X)-\nabla_X f(X_{t-1},Y_{t-1})\right)\right],
\end{aligned}
\tag{77}
$$

it follows that, for all $t\ge 1$,

$$
\begin{aligned}
&E_{t+1}\\
&=\beta\left(M_t^X-\nabla\Phi(X_{t-1})\right)\\
&\quad+\beta\left(\nabla\Phi(X_{t-1})-\nabla\Phi(X_t)\right)+\nabla_X f(X_t,Y_t)-\beta\nabla_X f(X_{t-1},Y_{t-1})-(1-\beta)\nabla\Phi(X_t)\\
&\quad+\left[\left(\nabla_X F(X_t,Y_t;\xi_t^X)-\nabla_X f(X_t,Y_t)\right)-\beta\left(\nabla_X F(X_{t-1},Y_{t-1};\xi_t^X)-\nabla_X f(X_{t-1},Y_{t-1})\right)\right]\\
&=\beta E_t+\beta\left(\nabla\Phi(X_{t-1})-\nabla\Phi(X_t)\right)\\
&\quad+\nabla_X f(X_t,Y_t)-\beta\nabla_X f(X_{t-1},Y_{t-1})-(1-\beta)\nabla\Phi(X_t)\\
&\quad+\left[\left(\nabla_X F(X_t,Y_t;\xi_t^X)-\nabla_X f(X_t,Y_t)\right)-\beta\left(\nabla_X F(X_{t-1},Y_{t-1};\xi_t^X)-\nabla_X f(X_{t-1},Y_{t-1})\right)\right].
\end{aligned}
\tag{78}
$$

Taking squared Frobenius norms on both sides of Eq. (78) and then taking expectations, we obtain

$$
\begin{aligned}
&\mathbb{E}[\|E_{t+1}\|_F^2]\\
&=\mathbb{E}\left[\left\|\beta E_t+\beta\left(\nabla\Phi(X_{t-1})-\nabla\Phi(X_t)\right)+\nabla_X f(X_t,Y_t)-\beta\nabla_X f(X_{t-1},Y_{t-1})-(1-\beta)\nabla\Phi(X_t)\right.\right.\\
&\qquad\left.\left.+\left(\nabla_X F(X_t,Y_t;\xi_t^X)-\nabla_X f(X_t,Y_t)\right)-\beta\left(\nabla_X F(X_{t-1},Y_{t-1};\xi_t^X)-\nabla_X f(X_{t-1},Y_{t-1})\right)\right\|_F^2\right]\\
&=\mathbb{E}\left[\left\|\beta E_t+\beta\left(\nabla\Phi(X_{t-1})-\nabla\Phi(X_t)\right)\right.\right.\\
&\qquad\left.\left.+\nabla_X f(X_t,Y_t)-\beta\nabla_X f(X_{t-1},Y_{t-1})-(1-\beta)\nabla\Phi(X_t)\right\|_F^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\mathbb{E}\Big\langle \beta E_t + \beta\big(\nabla\Phi(X_{t-1}) - \nabla\Phi(X_t)\big) \\
&\quad + \nabla_X f(X_t, Y_t) - \beta\nabla_X f(X_{t-1}, Y_{t-1}) - (1-\beta)\nabla\Phi(X_t), \\
&\quad \big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) - \beta\big(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\big)\Big\rangle \\
&+ \mathbb{E}\Bigg[\Bigg\|\big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) \\
&\quad - \beta\big(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\big)\Bigg\|_F^2\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\|\beta E_t + \beta\big(\nabla\Phi(X_{t-1}) - \nabla\Phi(X_t)\big) \\
&\quad + \nabla_X f(X_t, Y_t) - \beta\nabla_X f(X_{t-1}, Y_{t-1}) - (1-\beta)\nabla\Phi(X_t)\Bigg\|_F^2\Bigg] \\
&+ \mathbb{E}\Bigg[\Bigg\|\big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) \\
&\quad - \beta\big(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\big)\Bigg\|_F^2\Bigg] \\
&\le \beta\mathbb{E}[\|E_t\|_F^2] + \frac{1}{1-\beta}\mathbb{E}\Bigg[\Bigg\|\beta\big(\nabla\Phi(X_{t-1}) - \nabla\Phi(X_t)\big) \\
&\quad + \nabla_X f(X_t, Y_t) - \beta\nabla_X f(X_{t-1}, Y_{t-1}) - (1-\beta)\nabla\Phi(X_t)\Bigg\|_F^2\Bigg] \\
&+ \mathbb{E}\Bigg[\Bigg\|\big(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\big) \\
&\quad - \beta\big(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\big)\Bigg\|_F^2\Bigg]. \qquad (79)
\end{aligned}
$$

The last inequality follows from Young's inequality: $\|\beta A + B\|_F^2 \le \beta\|A\|_F^2 + \frac{1}{1-\beta}\|B\|_F^2$, where $\beta \in [0, 1)$.

To bound the second term in Eq. (79), note that

$$
\begin{aligned}
&\nabla_X f(X_t, Y_t) - \beta\nabla_X f(X_{t-1}, Y_{t-1}) - (1-\beta)\nabla\Phi(X_t) \\
&= (1-\beta)\big(\nabla_X f(X_t, Y_t) - \nabla\Phi(X_t)\big) + \beta\big(\nabla_X f(X_t, Y_t) - \nabla_X f(X_{t-1}, Y_{t-1})\big).
\end{aligned}
$$

Therefore, by the inequality $\|A + B + C\|_F^2 \le 3(\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2)$, we obtain

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Bigg\|\beta\big(\nabla\Phi(X_{t-1}) - \nabla\Phi(X_t)\big) + \nabla_X f(X_t, Y_t) - \beta\nabla_X f(X_{t-1}, Y_{t-1}) - (1-\beta)\nabla\Phi(X_t)\Bigg\|_F^2\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\|\beta\big(\nabla\Phi(X_{t-1}) - \nabla\Phi(X_t)\big) + (1-\beta)\big(\nabla_X f(X_t, Y_t) - \nabla\Phi(X_t)\big) \\
&\quad + \beta\big(\nabla_X f(X_t, Y_t) - \nabla_X f(X_{t-1}, Y_{t-1})\big)\Bigg\|_F^2\Bigg] \\
&\le 3\beta^2\mathbb{E}\left[\|\nabla\Phi(X_t) - \nabla\Phi(X_{t-1})\|_F^2\right] + 3(1-\beta)^2\mathbb{E}\left[\|\nabla_X f(X_t, Y_t) - \nabla\Phi(X_t)\|_F^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 3\beta^2 \mathbb{E}\left[\|\nabla_X f(X_t, Y_t) - \nabla_X f(X_{t-1}, Y_{t-1})\|_F^2\right] \\
&\leq 3\beta^2 L_\Phi^2 r \mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X \theta_t\}\right)^2\right] + 3(1-\beta)^2 L^2 \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2] \\
&\quad + 3\beta^2 L^2 r \left(\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X \theta_t\}\right)^2\right] + \mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y \sigma_t\}\right)^2\right]\right). \qquad (80)
\end{aligned}
$$

The last inequality follows from Assumption 4.1, Lemma B.1, and the global rank bound r, which yields $\|X_t - X_{t-1}\|_F^2 \leq r(\max\{\epsilon_X, \eta_X\theta_t\})^2$ and $\|Y_t - Y_{t-1}\|_F^2 \leq r(\max\{\epsilon_Y, \eta_Y\sigma_t\})^2$.

Next, according to Eq. (69), the third term in Eq. (79) can be upper bounded by

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\left(\nabla_X F(X_t, Y_t; \xi_t^X) - \nabla_X f(X_t, Y_t)\right) - \beta\left(\nabla_X F(X_{t-1}, Y_{t-1}; \xi_t^X) - \nabla_X f(X_{t-1}, Y_{t-1})\right)\right\|_F^2\right] \\
&\leq \frac{2\beta^2 L^2}{b_t}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\beta^2 L^2}{b_t}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + \frac{2(1-\beta)^2\kappa^2}{b_t} \\
&\leq \frac{2\beta^2 L^2 r}{b_t}\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_t\}\right)^2\right] + \frac{2\beta^2 L^2 r}{b_t}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_t\}\right)^2\right] + \frac{2(1-\beta)^2\kappa^2}{b_t}. \qquad (81)
\end{aligned}
$$

Summing Eq. (79) over $t = 1, \ldots, T-1$, keeping the initial term separately, and substituting Eq. (80) and Eq. (81), we obtain

$$
\begin{aligned}
&(1-\beta)\sum_{t=0}^{T-1}\mathbb{E}[\|E_{t+1}\|_F^2] \\
&\leq \mathbb{E}[\|E_1\|_F^2] - \beta\mathbb{E}[\|E_T\|_F^2] + \left(\frac{3\beta^2 L_\Phi^2 r}{1-\beta} + \frac{3\beta^2 L^2 r}{1-\beta}\right)\sum_{t=1}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_t\}\right)^2\right] \\
&\quad + \frac{3\beta^2 L^2 r}{1-\beta}\sum_{t=1}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_t\}\right)^2\right] + 2\beta^2 L^2 r\sum_{t=1}^{T-1}\frac{1}{b_t}\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_t\}\right)^2\right] \\
&\quad + 2\beta^2 L^2 r\sum_{t=1}^{T-1}\frac{1}{b_t}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_t\}\right)^2\right] \\
&\quad + 3(1-\beta)L^2\sum_{t=1}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2] + 2(1-\beta)^2\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}.
\end{aligned}
$$

Dividing by $1-\beta$, using $b_t \geq \min_{0\leq s<T} b_s$, and applying Lemma B.2 with the positive initialization yield

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}[\|E_{t+1}\|_F^2] \\
&\leq \frac{\mathbb{E}[\|E_1\|_F^2]}{1-\beta} - \frac{\beta}{1-\beta}\mathbb{E}[\|E_T\|_F^2] + \sum_{t=1}^{T-1}\left(\frac{3\beta^2 L_\Phi^2 r}{(1-\beta)^2} + \frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)b_t}\right)\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_t\}\right)^2\right] \\
&\quad + \sum_{t=1}^{T-1}\left(\frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)b_t}\right)\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_t\}\right)^2\right] \\
&\quad + 3L^2\sum_{t=1}^{T-1}\mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2] + 2(1-\beta)\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t} \\
&\leq \frac{\mathbb{E}[\|E_1\|_F^2]}{1-\beta} + \left(\frac{3\beta^2 L_\Phi^2 r}{(1-\beta)^2} + \frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)\min_{0\leq s<T} b_s}\right)\sum_{t=1}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_X, \eta_X\theta_t\}\right)^2\right] \\
&\quad + \left(\frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)\min_{0\leq s<T} b_s}\right)\sum_{t=1}^{T-1}\mathbb{E}\left[\left(\max\{\epsilon_Y, \eta_Y\sigma_t\}\right)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 3L^2 \sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2] + 2(1-\beta) \sum_{t=1}^{T-1} \frac{\kappa^2}{b_t} \\
&\leq \frac{\mathbb{E}[\|E_1\|_F^2]}{1-\beta} + 3L^2 \sum_{t=0}^{T-1} \mathbb{E}[\|Y_t - Y^\star(X_t)\|_F^2] + 2(1-\beta) \sum_{t=1}^{T-1} \frac{\kappa^2}{b_t} \\
&\quad + \left(\frac{3\beta^2 L_\Phi^2 r}{(1-\beta)^2} + \frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)\min_{0\leq s<T} b_s} \right) \left(T\epsilon_X^2 + \eta_X^2 \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)\right) \\
&\quad + \left(\frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)\min_{0\leq s<T} b_s} \right) \left(T\epsilon_Y^2 + \eta_Y^2 \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right)\right). \qquad (82)
\end{aligned}
$$

We next bound the initialization error. For fixed initial iterates, the initialization $M_0^X = 0$, $X_{-1} = X_0$, and $Y_{-1} = Y_0$ gives $E_1 = (1-\beta)\nabla_X F(X_0, Y_0; \xi_0^X) - \nabla\Phi(X_0)$. By strong concavity and inner optimality, $\frac{\mu}{2}\|Y_0 - Y^\star(X_0)\|_F^2 \leq \Phi(X_0) - f(X_0, Y_0) = \Delta_0^Y$. Moreover, Assumption 4.1 and Lemma B.1 imply

$$\|\nabla_X f(X_0, Y_0) - \nabla\Phi(X_0)\|_F^2 = \|\nabla_X f(X_0, Y_0) - \nabla_X f(X_0, Y^\star(X_0))\|_F^2 \leq L^2 \|Y_0 - Y^\star(X_0)\|_F^2.$$

Using conditional unbiasedness of the initial batch and Assumption 4.3, we obtain

$$
\begin{aligned}
\mathbb{E}[\|E_1\|_F^2] &= \|(1-\beta)\nabla_X f(X_0, Y_0) - \nabla\Phi(X_0)\|_F^2 \\
&\quad + (1-\beta)^2 \mathbb{E}\left[\|\nabla_X F(X_0, Y_0; \xi_0^X) - \nabla_X f(X_0, Y_0)\|_F^2\right] \\
&= \left\|-\beta\nabla_X f(X_0, Y_0) + \left(\nabla_X f(X_0, Y_0) - \nabla\Phi(X_0)\right)\right\|_F^2 \\
&\quad + (1-\beta)^2 \mathbb{E}\left[\|\nabla_X F(X_0, Y_0; \xi_0^X) - \nabla_X f(X_0, Y_0)\|_F^2\right] \\
&\leq 2\beta^2 \|\nabla_X f(X_0, Y_0)\|_F^2 + 2\|\nabla_X f(X_0, Y_0) - \nabla\Phi(X_0)\|_F^2 \\
&\quad + (1-\beta)^2 \mathbb{E}\left[\|\nabla_X F(X_0, Y_0; \xi_0^X) - \nabla_X f(X_0, Y_0)\|_F^2\right] \\
&\leq 2\beta^2 G^2 + 2L^2 \|Y_0 - Y^\star(X_0)\|_F^2 + \frac{(1-\beta)^2\kappa^2}{b_0} \\
&\leq 2\beta^2 G^2 + \frac{4L^2\Delta_0^Y}{\mu} + \frac{(1-\beta)^2\kappa^2}{b_0}.
\end{aligned}
$$

Then, applying Assumption 4.3 and invoking Eq. (66), we obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E}[\|E_{t+1}\|_F^2] \\
&\leq \frac{2\beta^2 G^2}{1-\beta} + \frac{4L^2\Delta_0^Y}{\mu(1-\beta)} + \frac{(1-\beta)\kappa^2}{b_0} \\
&\quad + \left(\frac{3\beta^2 L_\Phi^2 r}{(1-\beta)^2} + \frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)\min_{0\leq s<T} b_s} \right) \left(T\epsilon_X^2 + \eta_X^2 \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right)\right) \\
&\quad + \left(\frac{3\beta^2 L^2 r}{(1-\beta)^2} + \frac{2\beta^2 L^2 r}{(1-\beta)\min_{0\leq s<T} b_s} \right) \left(T\epsilon_Y^2 + \eta_Y^2 \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right)\right) + 2(1-\beta) \sum_{t=1}^{T-1} \frac{\kappa^2}{b_t} \\
&\quad + \frac{24GL^2\Delta_0^Y}{7\mu^2\epsilon_Y} + \frac{12GrL^2(L_\Phi + L)\epsilon_X^2 T}{7\mu^2\epsilon_Y} + \frac{48r^2L^4G\epsilon_X^2 T}{7\mu^3\epsilon_Y} + \frac{48rG^2L^4\epsilon_X^2 T}{7\mu^4\epsilon_Y^2} \\
&\quad + \left(\frac{12GrL^2(L_\Phi + L)\eta_X^2}{7\mu^2\epsilon_Y} + \frac{48r^2L^4G\eta_X^2}{7\mu^3\epsilon_Y} + \frac{48rG^2L^4\eta_X^2}{7\mu^4\epsilon_Y^2} \right) \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2}\right) \\
&\quad + \frac{3GL^2\epsilon_Y T}{7\mu} + \frac{12GL^3 r\epsilon_Y T}{7\mu^2} + \left(\frac{3GL^2\eta_Y^2}{7\mu\epsilon_Y} + \frac{12GL^3 r\eta_Y^2}{7\mu^2\epsilon_Y} \right) \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right) \\
&\quad + \frac{48G\sqrt{r}\, L^2 T}{7\mu^2\epsilon_Y} \sqrt{\epsilon_Y^2 + \frac{\eta_Y^2}{T} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2}\right)} \cdot \left(\frac{3\beta^2 G^2}{T(1-\beta)} + \frac{3(1-\beta)\kappa^2}{Tb_0} + 3\epsilon_X^2 L^2 r \right.
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{3\eta_X^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+3\epsilon_Y^2L^2r+\frac{3\eta_Y^2L^2r}{T}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\\
&+\frac{6\beta^2\epsilon_X^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}+\frac{6\beta^2\eta_X^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&+\frac{6\beta^2\epsilon_Y^2L^2r}{(1-\beta^2)\min_{0\le s<T}b_s}+\frac{6\beta^2\eta_Y^2L^2r}{T(1-\beta^2)\min_{0\le s<T}b_s}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{6(1-\beta)}{T(1+\beta)}\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\Bigg)^{\frac{1}{2}}.
\end{aligned}
\tag{83}
$$

The proof for Lemma D.2 is thus completed.

If we set $1-\beta=T^{-\frac{1}{3}}$, $\epsilon_X=T^{-\frac{2}{3}}$, $\epsilon_Y=T^{-\frac{1}{3}}$, $\eta_X=T^{-(\frac{2}{3}+q)}$, $\eta_Y=T^{-(\frac{1}{2}+q)}$, where $q\ge 0$ is fixed, together with $T\ge 2$, $\gamma>0$, $m_0^X=m_0^Y>0$, and $b_t\ge T^{\frac{1}{3}}$ for every $t\in\{0,\dots,T-1\}$, then $1-\beta^2=T^{-\frac{1}{3}}(1+\beta)$, $\frac{\beta^2}{1+\beta}\le\frac{1}{2}$, and $\frac{1}{1+\beta}\le 1$. Moreover, $\frac{1}{b_0}\le T^{-\frac{1}{3}}$, $\frac{1}{\min_{0\le s<T}b_s}\le T^{-\frac{1}{3}}$, and $\sum_{t=1}^{T-1}\frac{\kappa^2}{b_t}\le(T-1)\kappa^2T^{-\frac{1}{3}}\le\kappa^2T^{\frac{2}{3}}$. Therefore, substituting the schedules into Eq. (76), we obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|E_{t+1}\|_*]\\
&\le\frac{\sqrt{r}\,G}{T^{\frac{2}{3}}}+\frac{\kappa\sqrt{r}}{T^{\frac{7}{6}}}+\Bigg(\frac{r^2L^2+2r\kappa^2}{T^{\frac{2}{3}}}+\frac{r^2L^2}{T^{\frac{4}{3}}}+\frac{r^2L^2}{T^{\frac{7}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&\quad+\frac{r^2L^2}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\Bigg)^{\frac{1}{2}}+\frac{L\Delta_0^Y}{\mu T^{\frac{2}{3}}}+\frac{GL\sqrt{r}(2+\lambda)}{\mu T^{\frac{1}{3}}}\\
&\quad+\frac{GL\sqrt{r}(2+\lambda)}{\mu T^{\frac{5}{6}+q}}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\\
&\quad+\frac{L^2(2+\lambda^2)}{2\mu T}+\frac{L^2(2+\lambda^2)}{2\mu T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&\quad+\frac{L^2T^{\frac{1}{3}}}{\mu}\sqrt{\left(\frac{1}{T^{\frac{4}{3}}}+\frac{1}{T^{\frac{7}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\left(\frac{1}{T^{\frac{2}{3}}}+\frac{1}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)}\\
&\quad+\frac{L^2r}{2\mu T^{\frac{1}{3}}}+\frac{L^2r}{2\mu T^{\frac{5}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2L\sqrt{r}\,T^{\frac{1}{3}}}{\mu}\sqrt{\frac{1}{T^{\frac{2}{3}}}+\frac{1}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\\
&\quad\cdot\Bigg(\frac{3G^2+6L^2r+6\kappa^2}{T^{\frac{2}{3}}}+\frac{6L^2r}{T^{\frac{4}{3}}}+\frac{3\kappa^2}{T^{\frac{5}{3}}}+\frac{6L^2r}{T^{\frac{7}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&\qquad+\frac{6L^2r}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\Bigg)^{\frac{1}{2}}\\
&\lesssim\frac{rL+\kappa\sqrt{r}}{T^{\frac{1}{3}}}+\frac{L\sqrt{r}G(1+\lambda)+L^2r+\kappa L\sqrt{r}}{\mu T^{\frac{1}{3}}}+\frac{\sqrt{r}\,G}{T^{\frac{2}{3}}}+\frac{L\Delta_0^Y+L^2}{\mu T^{\frac{2}{3}}}.
\end{aligned}
\tag{84}
$$

Similarly, for Lemma D.2, we have

$$
\begin{aligned}
&\frac{2\sqrt{r}\,\eta_X}{T\epsilon_X}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|E_{t+1}\|_F^2]}\\
&=\frac{2\sqrt{r}}{T^{1+q}}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|E_{t+1}\|_F^2]}\\
&\le\frac{2\sqrt{r}}{T^{1+q}}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\cdot\Bigg[3rL^2T+\left(\frac{3GL^2}{7\mu}+\frac{12GL^3r}{7\mu^2}\right)T^{\frac{2}{3}}
\end{aligned}
$$

$$
\begin{aligned}
&+\left(2G^2+2\kappa^2+3rL_\Phi^2+5rL^2+\frac{4L^2\Delta_0^Y}{\mu}+\frac{24GL^2\Delta_0^Y}{7\mu^2}+\frac{48rG^2L^4}{7\mu^4}\right)T^{\frac{1}{3}}\\
&+\frac{2rL^2}{T^{\frac{1}{3}}}+\frac{\kappa^2}{T^{\frac{2}{3}}}+\frac{12GrL^2(L_\Phi+L)}{7\mu^2}+\frac{48r^2GL^4}{7\mu^3}+\left(\frac{3r(L_\Phi^2+L^2)+\frac{48rG^2L^4}{7\mu^4}}{T^{\frac{2}{3}+2q}}+\frac{2rL^2}{T^{\frac{4}{3}+2q}}\right.\\
&\left.+\frac{\frac{12GrL^2(L_\Phi+L)}{7\mu^2}+\frac{48r^2GL^4}{7\mu^3}}{T^{1+2q}}\right)\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&+\left(\frac{3rL^2}{T^{\frac{1}{3}+2q}}+\frac{2rL^2}{T^{1+2q}}+\frac{\frac{3GL^2}{7\mu}+\frac{12GL^3r}{7\mu^2}}{T^{\frac{2}{3}+2q}}\right)\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\\
&+\frac{48G\sqrt{r}\,L^2T^{\frac{4}{3}}}{7\mu^2}\sqrt{\frac{1}{T^{\frac{2}{3}}}+\frac{1}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\\
&\cdot\left(\frac{3G^2+6L^2r+6\kappa^2}{T^{\frac{2}{3}}}+\frac{6L^2r}{T^{\frac{4}{3}}}+\frac{3\kappa^2}{T^{\frac{5}{3}}}+\frac{6L^2r}{T^{\frac{7}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right.\\
&\left.+\frac{6L^2r}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)^{\frac{1}{2}}\Bigg]^{\frac{1}{2}}\\
&\lesssim\left[\frac{rL}{T^{\frac{1}{2}+q}}+\frac{GL\left(\mu+rL+\sqrt{r}\sqrt{G^2+L^2r+\kappa^2}\right)}{\mu^2T^{\frac{5}{6}+q}}\right]\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}. \qquad (85)
\end{aligned}
$$

Substituting Eq. (84) and Eq. (85) into Lemma C.1, we obtain

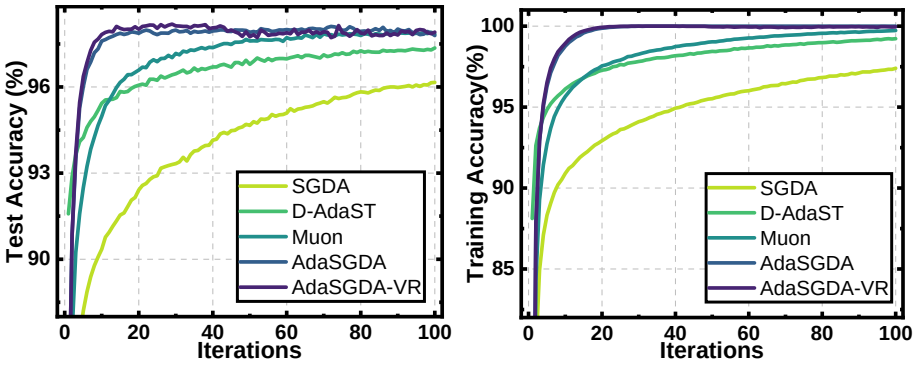
$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Phi(X_t)\|_*]\\
&\le\frac{\Delta}{T^{\frac{1}{3}}}+2\left(\frac{\sqrt{r}\,G}{T^{\frac{2}{3}}}+\frac{\kappa\sqrt{r}}{T^{\frac{7}{6}}}+\left(\frac{r^2L^2+2r\kappa^2}{T^{\frac{2}{3}}}+\frac{r^2L^2}{T^{\frac{4}{3}}}+\frac{r^2L^2}{T^{\frac{7}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right.\right.\\
&\left.\left.+\frac{r^2L^2}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)^{\frac{1}{2}}+\frac{L\Delta_0^Y}{\mu T^{\frac{2}{3}}}+\frac{GL\sqrt{r}(2+\lambda)}{\mu T^{\frac{1}{3}}}\right.\\
&+\frac{GL\sqrt{r}(2+\lambda)}{\mu T^{\frac{5}{6}+q}}\sqrt{\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)}\\
&+\frac{L^2(2+\lambda^2)}{2\mu T}+\frac{L^2(2+\lambda^2)}{2\mu T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\\
&+\frac{L^2T^{\frac{1}{3}}}{\mu}\sqrt{\left(\frac{1}{T^{\frac{4}{3}}}+\frac{1}{T^{\frac{7}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right)\left(\frac{1}{T^{\frac{2}{3}}}+\frac{1}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)}\\
&+\frac{L^2r}{2\mu T^{\frac{1}{3}}}+\frac{L^2r}{2\mu T^{\frac{5}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)+\frac{2L\sqrt{r}\,T^{\frac{1}{3}}}{\mu}\sqrt{\frac{1}{T^{\frac{2}{3}}}+\frac{1}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)}\\
&\cdot\left(\frac{3G^2+6L^2r+6\kappa^2}{T^{\frac{2}{3}}}+\frac{6L^2r}{T^{\frac{4}{3}}}+\frac{3\kappa^2}{T^{\frac{5}{3}}}+\frac{6L^2r}{T^{\frac{7}{3}+2q}}\ln\left(1+\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)\right.\\
&\left.\left.+\frac{6L^2r}{T^{2+2q}}\ln\left(1+\frac{\gamma^2T}{[m_0^Y]^2}\right)\right)^{\frac{1}{2}}\right)+\frac{2\sqrt{r}}{T^{1+q}}\left(\sqrt{\ln\left(\frac{\gamma^2T}{[\max\{m_0^X,m_0^Y\}]^2}\right)+1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \Bigg[3rL^2T + \left(\frac{3GL^2}{7\mu} + \frac{12GL^3r}{7\mu^2} \right) T^{\frac{2}{3}} \\
&+ \left(2G^2 + 2\kappa^2 + 3rL_\Phi^2 + 5rL^2 + \frac{4L^2\Delta_0^Y}{\mu} + \frac{24GL^2\Delta_0^Y}{7\mu^2} + \frac{48rG^2L^4}{7\mu^4} \right) T^{\frac{1}{3}} \\
&+ \frac{2rL^2}{T^{\frac{1}{3}}} + \frac{\kappa^2}{T^{\frac{2}{3}}} + \frac{12GrL^2(L_\Phi + L)}{7\mu^2} + \frac{48r^2GL^4}{7\mu^3} + \left(\frac{3r(L_\Phi^2 + L^2) + \frac{48rG^2L^4}{7\mu^4}}{T^{\frac{2}{3}+2q}} + \frac{2rL^2}{T^{\frac{4}{3}+2q}} \right. \\
&\left. + \frac{\frac{12GrL^2(L_\Phi + L)}{7\mu^2} + \frac{48r^2GL^4}{7\mu^3}}{T^{1+2q}} \right) \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} \right) \\
&+ \left(\frac{3rL^2}{T^{\frac{1}{3}+2q}} + \frac{2rL^2}{T^{1+2q}} + \frac{\frac{3GL^2}{7\mu} + \frac{12GL^3r}{7\mu^2}}{T^{\frac{2}{3}+2q}} \right) \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2} \right) \\
&+ \frac{48G\sqrt{r}\, L^2 T^{\frac{4}{3}}}{7\mu^2} \sqrt{\frac{1}{T^{\frac{2}{3}}} + \frac{1}{T^{2+2q}} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2} \right)} \\
&\cdot \left(\frac{3G^2 + 6L^2r + 6\kappa^2}{T^{\frac{2}{3}}} + \frac{6L^2r}{T^{\frac{4}{3}}} + \frac{3\kappa^2}{T^{\frac{5}{3}}} + \frac{6L^2r}{T^{\frac{7}{3}+2q}} \ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} \right) \right. \\
&\left. + \frac{6L^2r}{T^{2+2q}} \ln\left(1 + \frac{\gamma^2 T}{[m_0^Y]^2} \right) \right)^{\frac{1}{2}} \Bigg]^{\frac{1}{2}} + \frac{L_\Phi}{T^{\frac{2}{3}}} + \frac{L_\Phi}{T^{\frac{5}{3}+2q}} \left(\ln\left(\frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} \right) + 1 \right) \\
\lesssim\ & \frac{\Delta + rL + \kappa\sqrt{r}}{T^{\frac{1}{3}}} + \frac{L\sqrt{r}G(1+\lambda) + L^2r + \kappa L\sqrt{r}}{\mu T^{\frac{1}{3}}} + \frac{L_\Phi + \sqrt{r}\, G}{T^{\frac{2}{3}}} + \frac{L\Delta_0^Y + L^2}{\mu T^{\frac{2}{3}}} \\
&+ \left[\frac{rL}{T^{\frac{1}{2}+q}} + \frac{GL\left(\mu + rL + \sqrt{r}\sqrt{G^2 + L^2r + \kappa^2}\right)}{\mu^2 T^{\frac{5}{6}+q}} \right] \sqrt{\ln\left(1 + \frac{\gamma^2 T}{[\max\{m_0^X, m_0^Y\}]^2} \right)} \\
\leq\ & \mathcal{O}\left(\frac{\Delta + rL + \kappa\sqrt{r}}{T^{\frac{1}{3}}} + \frac{GL\sqrt{r}(1+\lambda) + rL^2 + \kappa L\sqrt{r}}{\mu T^{\frac{1}{3}}} + \frac{rL}{T^{\frac{1}{2}+q}} \sqrt{\ln\left(\frac{\gamma^2 T}{z_0^2} \right) + 1} \right), \qquad (86)
\end{aligned}
$$

where $\Delta := \Phi(X_0) - \Phi_*$ and $z_0 := \max\{m_0^X, m_0^Y\}$.

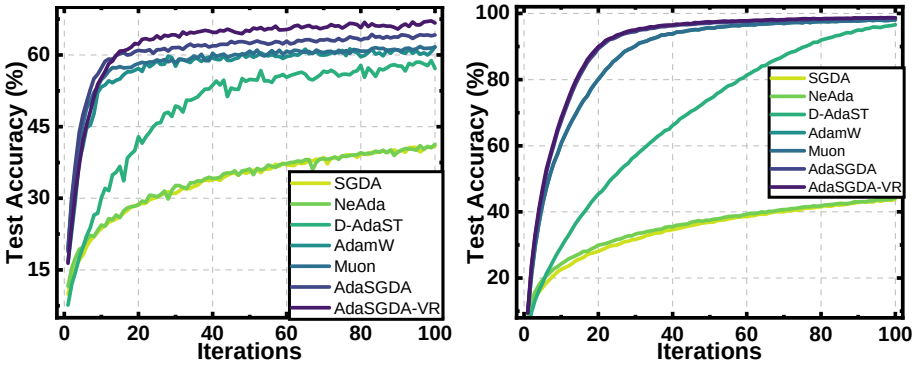
The proof for Theorem 4.6 is thus completed.

E Additional Experiments



(a) Test accuracy (b) Training accuracy

Figure 5: Test and training accuracy on the MNIST dataset.



(a) Test accuracy (b) Training accuracy

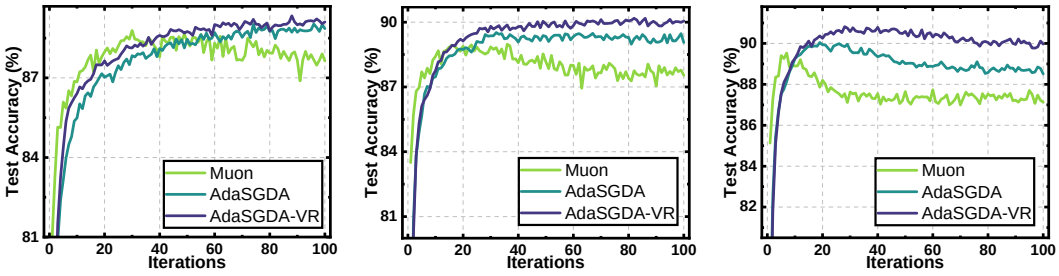
Figure 6: Test and training accuracy on the CIFAR-100 dataset.

E.1 Experimental Details

This section summarizes the experimental setups and hyperparameter choices used throughout our experiments. For our methods, hyperparameter selection is guided by the theoretical analysis in Theorems 4.5 and 4.6. Throughout all experiments, the additional exponent offset is fixed to $q = 0$. We fix γ and the initial accumulators m_0^X, m_0^Y to 0.9 in all experiments and do not tune them. For both Muon and our methods, the number of Newton–Schulz (NS) iterations is set to 3 by default. For all other baselines, we tune hyperparameters to report their best performance. In particular, we first locate a reasonable stepsize scale by searching over $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$, and then perform a finer grid search around the best scale. For Muon, we additionally tune the momentum coefficient over $\{0.1, 0.3, 0.5, 0.7, 0.9, 0.95\}$ and report the best result.

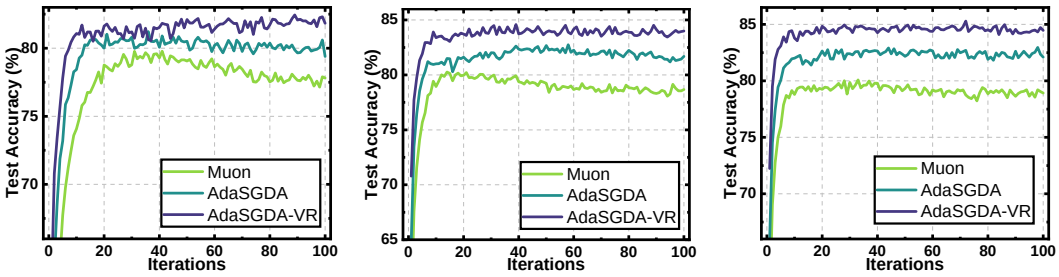
Distributional Robustness Optimization. For the DRO experiments, we follow the same setup as Sebbouh et al. [2022] and Sinha et al. [2017], and fix the robustness parameter to $\rho = 1.3$. We use different model architectures across datasets: an MLP (0.5M parameters) for MNIST [LeCun et al., 2002] and FMNIST [Xiao et al., 2017], CifarNet (2M) for CIFAR-10 [Krizhevsky et al., 2009], ResNet-34 (21.8M) for CIFAR-100 [Krizhevsky et al., 2009], and three Transformer language models for FineWeb [Penedo et al., 2024], namely OPT-125M, OPT-350M, and OPT-1.3B.

Generative Adversarial Network. For the GAN experiments, we use an implementation adapted from Green9 [2018]. The discriminator is a four-layer convolutional neural network (CNN) and the generator is a four-layer transposed-convolution CNN. Following the experimental protocol in



(a) `Newton-Schulz` steps: 1 (b) `Newton-Schulz` steps: 2 (c) `Newton-Schulz` steps: 5

Figure 7: Test accuracy on the FMNIST dataset with different `Newton-Schulz` steps.



(a) `Newton-Schulz` steps: 1 (b) `Newton-Schulz` steps: 2 (c) `Newton-Schulz` steps: 5

Figure 8: Test accuracy on the CIFAR-10 dataset with different `Newton-Schulz` steps.

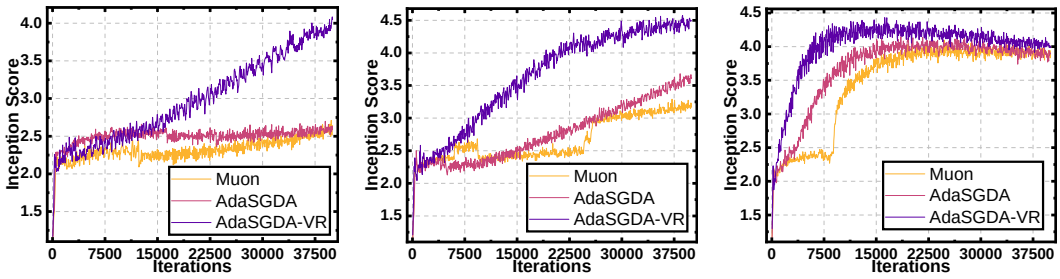
Daskalakis et al. [2017], we set the latent dimension to 50 and use a gradient-penalty weight of 10^{-4}. Inception Score is computed by feeding $8{,}000$ generated samples into a pretrained Inception network.

E.2 Additional Results on MNIST and CIFAR-100 Datasets

As shown in Figures 5 and 6, which report both training and test accuracies on the MNIST and CIFAR-100 datasets, respectively, **AdaSGDA** and **AdaSGDA-VR** significantly outperform SGDA and D-AdaST, which rely on conventional gradient directions. Compared with the orthogonal-update baseline Muon, our proposed methods also achieve stronger performance. This improvement stems from our parameter-free design: stepsizes are automatically adjusted based on accumulated gradient norms, calibrating the update magnitude throughout training and avoiding brittle manually tuned schedules. Notably, the advantage of our methods becomes more pronounced as the datasets and models grow more challenging (e.g., CIFAR-100 with a deeper network). In these settings, orthogonal updates can provide stability but also make the effective update scale more sensitive to the evolving geometry; our adaptive stepsize rule compensates for this effect and maintains well-balanced progress throughout training. Finally, we emphasize that all baselines require substantial manual tuning to reach their best performance, whereas our methods can be run with theory-guided hyperparameters, greatly simplifying deployment in practice.

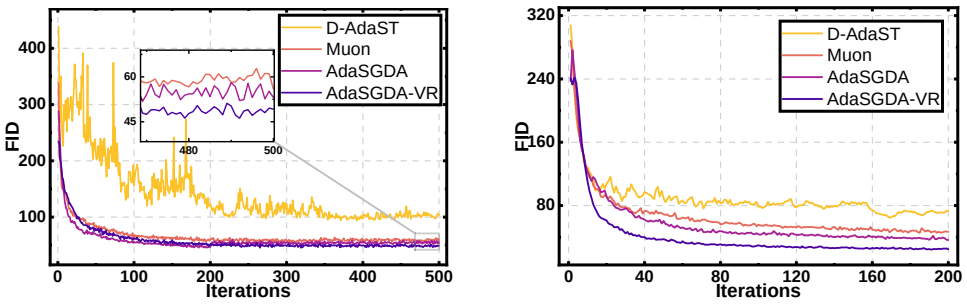
E.3 Sensitivity Analysis to Newton–Schulz Steps

Beyond the default setting of 3 NS steps, we vary the number of NS steps in $\{1, 2, 5\}$ and evaluate on the FMNIST and CIFAR-10 datasets, as shown in Figures 7 and 8, respectively. The NS iterations determine the orthogonalization accuracy, i.e., how well the update direction approximates the orthogonal (polar) factor; increasing NS steps typically yields a more accurate orthogonal direction, which can improve stability and training effectiveness. As shown in the results, our methods consistently outperform Muon under all NS choices. Moreover, increasing the number of NS steps generally improves performance for all methods, as orthogonalization becomes more accurate. Notably, our methods exhibit robust behavior across different NS steps and achieve consistently strong results, whereas Muon shows performance fluctuations when the NS steps are not well chosen. This further demonstrates the effectiveness of our adaptive design under orthogonal updates.



(a) `Newton-Schulz` steps: 1 (b) `Newton-Schulz` steps: 2 (c) `Newton-Schulz` steps: 5

Figure 9: Inception Score on WGAN-GP with different `Newton-Schulz` steps.



(a) CIFAR-10 dataset (b) CIFAR-100 dataset

Figure 10: FID on WGAN-GP with CIFAR-10 and CIFAR-100 datasets.

E.4 Additional Results on WGAN-GP

In Figure 9, we report the Inception Score under different numbers of NS steps. When NS steps are small (e.g., 1 or 2), the orthogonalization is less accurate, which can distort the update direction and amplify training instability. In this regime, AdaSGDA-VR shows a clear advantage over Muon, achieving higher inception scores with faster improvements over training. More broadly, AdaSGDA and AdaSGDA-VR consistently outperform Muon across all NS choices, while Muon exhibits visible performance fluctuations when NS steps are not well chosen. These results highlight the robustness of our adaptive orthogonal-update design, which remains stable even when the orthogonalization accuracy varies. We further report wall-clock time performance in Table 2. Our methods achieve better time-to-quality than all baselines: AdaSGDA-VR attains the highest inception score at every time budget, and AdaSGDA also consistently improves over SGDA and Muon. This demonstrates that our gains are not only in iteration complexity but also translate into practical efficiency in terms of wall-clock time.

Following [Shmelkov et al., 2018], we additionally evaluate WGAN-GP using the Fréchet Inception Distance (FID), which compares the feature statistics of generated and real images and is a standard metric for GAN evaluation. We report the iteration-wise FID curves on CIFAR-10 and CIFAR-100 in Figure 10. On CIFAR-10, both AdaSGDA and AdaSGDA-VR consistently achieve lower FID than Muon and D-AdaST throughout training, while also exhibiting noticeably more stable trajectories; in particular, D-AdaST shows substantial fluctuations, whereas our methods remain much more stable over the entire course of training. On the more challenging CIFAR-100 task, the same trend becomes even more pronounced: AdaSGDA and especially AdaSGDA-VR maintain a clearer advantage over the baselines, indicating that our adaptive orthogonal methods produce samples that are closer to the real data distribution and that the benefit of adaptivity becomes stronger in harder generative settings. We further report wall-clock FID in Tables 3 and 4. The same ranking is preserved under equal time budgets on both datasets: AdaSGDA-VR consistently achieves the best time-to-quality tradeoff, and AdaSGDA also remains uniformly better than Muon and D-AdaST. Therefore, the improvement of our methods is not only reflected by iteration-based comparisons, but also by stronger practical efficiency and better generative quality in real running time.

Table 2: Wall-clock time inception score of WGAN-GP for different algorithms.

	Wall-clock time inception score						
Algorithm	**600 sec**	**1200 sec**	**1800 sec**	**2400 sec**	**3000 sec**	**3600 sec**	**4200 sec**
SGDA	2.2465	2.3326	2.3479	2.3512	2.3546	2.3787	2.4095
Muon	2.0232	2.2709	2.3299	2.3385	2.3552	2.3711	2.3752
AdaSGDA	2.2542	2.3227	2.3751	2.4293	2.4438	2.4487	2.4758
AdaSGDA-VR	2.3062	2.3452	2.4475	2.4748	2.4832	2.5507	2.6897

Table 3: Wall-clock time FID of WGAN-GP on CIFAR-10 dataset for different algorithms.

	Wall-clock time FID						
Algorithm	**600 sec**	**1200 sec**	**1800 sec**	**2400 sec**	**3000 sec**	**3600 sec**	**4200 sec**
D-AdaST	271.75	242.48	226.89	199.81	171.55	160.87	134.67
Muon	167.84	128.28	107.90	97.04	92.60	84.79	79.24
AdaSGDA	164.24	127.06	105.06	96.09	89.58	83.41	75.94
AdaSGDA-VR	131.39	106.77	92.88	86.52	79.56	74.00	66.86

Table 4: Wall-clock time FID of WGAN-GP on CIFAR-100 dataset for different algorithms.

	Wall-clock time FID						
Algorithm	**600 sec**	**1200 sec**	**1800 sec**	**2400 sec**	**3000 sec**	**3600 sec**	**4200 sec**
D-AdaST	237.42	167.50	147.39	124.18	111.55	98.31	89.68
Muon	198.96	129.08	114.99	100.96	91.21	82.71	76.52
AdaSGDA	163.15	122.94	106.38	96.62	86.25	76.97	71.60
AdaSGDA-VR	158.19	115.35	97.81	89.03	81.25	71.14	66.70

E.5 Additional Results on the FineWeb Dataset

We further evaluate our methods on the FineWeb dataset under an embedding-level adversarial DRO formulation. For each training sample (token sequence) S_i, let $V_i = E_X(S_i)$ denote its input embedding sequence. We introduce an adversarial variable Y_i of the same shape, representing a perturbed embedding, and consider the following minimax objective with a quadratic penalty: $\min_X \max_{Y_1,\dots,Y_n} \frac{1}{n}\sum_{i=1}^n \ell(X;Y_i,S_i) - \rho\|Y_i - V_i\|_F^2$, where $\ell(X;Y_i,S_i)$ is the next-token prediction loss when using Y_i as the input embedding, and $\rho > 0$ controls the perturbation strength. This formulation trains the model against adversarial embedding perturbations in a local neighborhood around V_i, while discouraging overly large deviations through the quadratic regularization.

In Figure 11, we report the training loss and validation loss for OPT-350M. Both AdaSGDA and AdaSGDA-VR achieve lower losses than Muon under the same DRO training setup, demonstrating consistently better empirical performance. In Figure 12, we further compare validation loss on FineWeb for OPT-125M and OPT-1.3B. Across model scales, all three methods exhibit stable training trends, while the benefits of orthogonal updates become more pronounced as model size increases. This behavior can be attributed to the fact that, as model size increases, gradients and momentum directions tend to become more ill-conditioned and highly correlated across layers and parameters, so successive updates can interfere with each other and amplify unstable components in minimax training. Orthogonalized updates help alleviate this issue by reshaping the update directions to reduce such interference, thereby enabling more stable and reliable optimization at scale. Notably, the performance gap between our methods and Muon widens at OPT-1.3B compared with OPT-125M and OPT-350M. Meanwhile, unlike Muon, our proposed methods remain parameter-free and thus eliminate extensive manual tuning, further supporting the effectiveness of our adaptive design for large-scale training. Finally, AdaSGDA-VR achieves the best performance across all settings, consistent with our theory that variance reduction provides tighter stochastic error control and faster convergence.

Table 5: Wall-clock training loss of OPT-125M.

	Wall-clock training loss						
Algorithm	**600s**	**1200s**	**1800s**	**2400s**	**3000s**	**3600s**	**4200s**
Muon	6.6598	6.4266	6.3381	6.1984	6.0494	5.9709	5.8419
AdaSGDA	6.5628	6.3523	6.2203	6.0404	5.9412	5.8038	5.6744
AdaSGDA-VR	6.4094	6.2918	6.0617	5.9695	5.8521	5.7178	5.4577

Table 6: Wall-clock training loss of OPT-350M.

	Wall-clock training loss						
Algorithm	**600s**	**1200s**	**1800s**	**2400s**	**3000s**	**3600s**	**4200s**
Muon	7.0506	6.7895	6.4186	6.3847	6.2401	6.1844	6.1175
AdaSGDA	6.9492	6.5542	6.3820	6.2238	6.1351	6.0878	5.9382
AdaSGDA-VR	6.7569	6.5519	6.3191	6.2176	6.0907	5.9691	5.8582

Table 7: Wall-clock training loss of OPT-1.3B.

	Wall-clock training loss						
Algorithm	**600s**	**1200s**	**1800s**	**2400s**	**3000s**	**3600s**	**4200s**
Muon	7.8490	7.1454	6.7588	6.6044	6.3795	6.2671	6.1316
AdaSGDA	7.5296	6.9236	6.4627	6.3201	6.2450	6.1573	6.0156
AdaSGDA-VR	7.3432	6.6373	6.4266	6.2602	6.1029	6.0209	5.8820

Table 8: Test accuracy with stepsizes smaller than η^* on FMNIST.

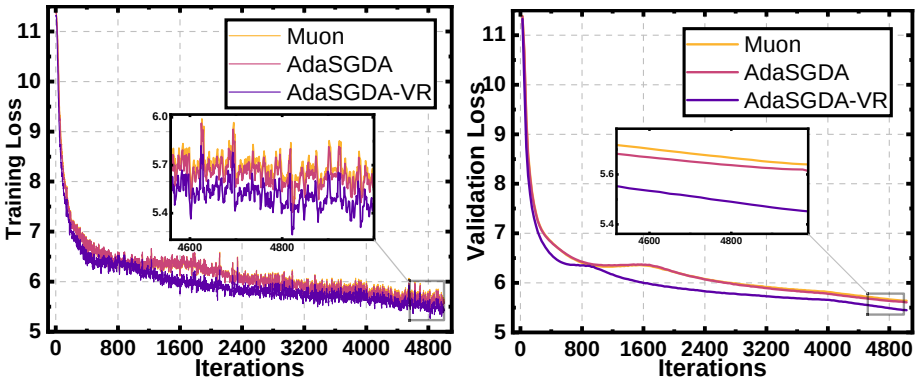
Algorithm	$\eta^* \times 0.05$	$\eta^* \times 0.1$	$\eta^* \times 0.5$	η^*
AdaSGDA	$88.85_{\pm 0.07}$	$88.83_{\pm 0.07}$	$88.16_{\pm 0.08}$	$89.53_{\pm 0.06}$
AdaSGDA-VR	$90.15_{\pm 0.05}$	$90.12_{\pm 0.07}$	$90.22_{\pm 0.06}$	$90.26_{\pm 0.04}$
Muon	$87.90_{\pm 0.08}$	$88.01_{\pm 0.06}$	$88.11_{\pm 0.08}$	$88.19_{\pm 0.09}$
D-AdaST	$81.82_{\pm 0.03}$	$83.65_{\pm 0.04}$	$86.76_{\pm 0.08}$	$87.79_{\pm 0.08}$
SGDA	$80.33_{\pm 0.07}$	$82.03_{\pm 0.08}$	$86.55_{\pm 0.06}$	$87.32_{\pm 0.06}$

Table 9: Test accuracy with stepsizes larger than η^* on FMNIST.

Algorithm	$\eta^* \times 5$	$\eta^* \times 10$	$\eta^* \times 50$	$\eta^* \times 100$
AdaSGDA	$88.88_{\pm 0.10}$	$88.94_{\pm 0.08}$	$88.93_{\pm 0.09}$	$88.96_{\pm 0.11}$
AdaSGDA-VR	$90.10_{\pm 0.09}$	$90.12_{\pm 0.09}$	$90.05_{\pm 0.12}$	$89.99_{\pm 0.07}$
Muon	$84.21_{\pm 0.11}$	$80.86_{\pm 0.12}$	$10.09_{\pm 0.03}$	$10.09_{\pm 0.03}$
D-AdaST	$87.52_{\pm 0.11}$	$87.30_{\pm 0.11}$	$86.51_{\pm 0.16}$	$9.59_{\pm 0.05}$
SGDA	$85.03_{\pm 0.13}$	$82.26_{\pm 0.15}$	$81.98_{\pm 0.12}$	$19.63_{\pm 0.06}$

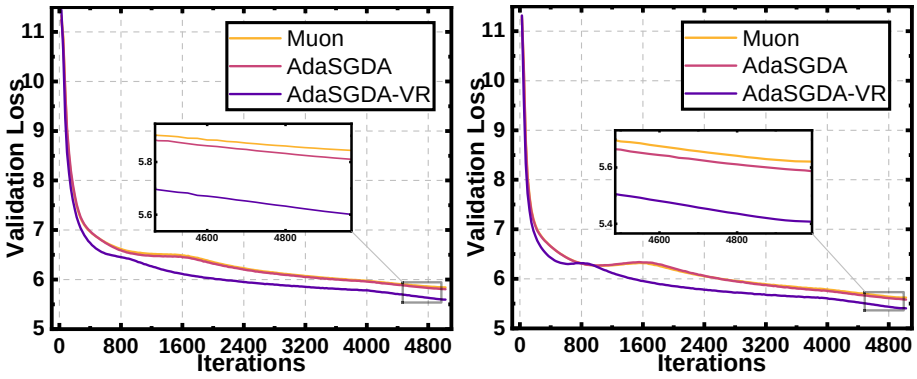
Table 10: Test accuracy versus momentum coefficient β on the FMNIST dataset.

Algorithm	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.9$
AdaSGDA	$89.20_{\pm 0.07}$	$89.32_{\pm 0.06}$	$89.48_{\pm 0.09}$	$89.46_{\pm 0.08}$	$89.49_{\pm 0.06}$
AdaSGDA-VR	$89.48_{\pm 0.05}$	$90.10_{\pm 0.07}$	$90.06_{\pm 0.10}$	$90.24_{\pm 0.09}$	$90.25_{\pm 0.06}$
Muon	$87.46_{\pm 0.09}$	$87.34_{\pm 0.10}$	$86.68_{\pm 0.12}$	$86.82_{\pm 0.10}$	$87.96_{\pm 0.10}$



(a) Training loss (b) Validation loss

Figure 11: Training and validation loss curves for OPT-350M on the FineWeb dataset.



(a) Validation loss (OPT-125M) (b) Validation loss (OPT-1.3B)

Figure 12: Validation loss curves for OPT-125M and OPT-1.3B on the FineWeb dataset.

We further report wall-clock training loss in Tables 5 to 7 for OPT-125M, OPT-350M, and OPT-1.3B. Across all three model scales, the same ranking is observed at every reported time budget: AdaSGDA-VR achieves the lowest training loss at every reported wall-clock checkpoint across all three model scales, while AdaSGDA is uniformly better than Muon. This demonstrates that the empirical advantage of our methods is not limited to loss-versus-iteration comparisons, but also yields consistently better time-to-quality tradeoffs in real training time. This consistent trend across model scales suggests that our adaptive orthogonal design continues to remain effective as the model size increases. Overall, these wall-clock results provide additional evidence that our proposed methods are practically more efficient for large-scale minimax-based language pre-training.

E.6 Sensitivity Analysis on Initial Stepsize and Momentum Coefficient

We further conduct sensitivity analyses on the FMNIST dataset for the initial stepsize and the momentum coefficient β, with results summarized in Tables 8 and 10.

- **Initial stepsize.** As shown in Tables 8 and 9, $\eta^\star$ denotes the best-tuned stepsize for each baseline, while for our methods $\eta^\star$ is selected following the theoretical guidelines in Theorems 4.5 and 4.6. We then scale $\eta^\star$ by $\{0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$ to test robustness. All algorithms are evaluated for 100 rounds for a fair comparison. The results show that our methods achieve the best accuracy across all stepsize scalings and exhibit remarkable resilience to stepsize selection, maintaining stable performance over a much broader range than the baselines. In sharp contrast, Muon fails to converge at the $50\times$ scale and all baselines break down at $100\times$, whereas our methods remain accurate and stable even with overly aggressive stepsizes, continuing to make steady progress without divergence or severe degradation. This highlights not only the strong

Table 11: Test accuracy for the ablation study on FMNIST, CIFAR-10, and CIFAR-100.

Algorithm	Fashion-MNIST	CIFAR-10	CIFAR-100
w/o Orthogonalization	$88.09_{\pm 0.10}$	$75.56_{\pm 0.11}$	$32.76_{\pm 0.11}$
w/o Adaptivity	$89.08_{\pm 0.08}$	$80.86_{\pm 0.08}$	$38.73_{\pm 0.09}$
w/o Hierarchical Scaling	$89.05_{\pm 0.08}$	$80.52_{\pm 0.11}$	$38.82_{\pm 0.10}$
w/o Clamping	$89.02_{\pm 0.06}$	$80.56_{\pm 0.06}$	$39.01_{\pm 0.08}$
w/o Momentum	$86.42_{\pm 0.06}$	$75.97_{\pm 0.09}$	$33.52_{\pm 0.11}$
AdaSGDA	$89.53_{\pm 0.06}$	$80.91_{\pm 0.08}$	$40.33_{\pm 0.08}$
AdaSGDA-VR	$90.26_{\pm 0.04}$	$81.56_{\pm 0.07}$	$44.85_{\pm 0.08}$

stepsize robustness of our approach, but also its practical value: the method can be deployed with minimal tuning effort while remaining reliable across a wide range of hyperparameter choices.

- **Momentum coefficient.** Table 10 varies the momentum coefficient over $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Across all choices, our methods consistently outperform Muon and show much smaller performance variation, indicating strong robustness to momentum tuning. This is expected: our adaptive stepsizes continuously calibrate the effective update magnitude using accumulated gradient information, which reduces the impact of momentum-induced overshooting and oscillations.

E.7 Ablation Study

We conduct ablation studies on FMNIST, CIFAR-10, and CIFAR-100 to examine the contributions of the individual components of AdaSGDA. Each ablated variant removes one component while keeping the remaining settings unchanged:

- **w/o orthogonalization:** replace the polar direction with a Frobenius-normalized direction while preserving the update magnitude;
- **w/o adaptivity:** use fixed stepsizes selected through the same validation protocol;
- **w/o hierarchical scaling:** scale the primal update using only its own accumulator;
- **w/o clamping:** accumulate the original, unclipped gradient norms;
- **w/o momentum:** set the momentum coefficient to zero.

Table 11 shows that removing any individual component consistently leads to a degradation in performance, indicating that each design choice contributes meaningfully to the overall performance of the algorithm. In particular, orthogonalization and momentum play significant roles, as their removal results in the largest accuracy drops across all datasets. The adaptive stepsize scheme, hierarchical scaling, and gradient clamping also provide consistent and non-negligible improvements, further stabilizing and enhancing optimization across tasks. Moreover, variance reduction yields additional gains, and AdaSGDA-VR achieves the best overall performance, which is consistent with its improved theoretical convergence rate. Overall, these results demonstrate that all components are indispensable and jointly contribute to the strong empirical performance of the proposed method.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper has discussed its limitations in the Section on Conclusions and Limitations.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The paper has provided the full set of assumptions and a complete and correct proof.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We have disclosed all the information needed to reproduce the main experimental results.

   Guidelines:

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

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

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

Answer: [No]

Justification: The experimental results shown in the submitted manuscript do not depend on private datasets and can be reproduced by following the provided instructions and settings.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We have provided detailed descriptions of the training and test setups, as well as all experimental configurations.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The paper has provided appropriate information about the statistical significance of the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We have provided the description of the type of compute workers and memory.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All the datasets used in our experiments are publicly available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.