Adam Improves Muon: Adaptive Moment Estimation with Orthogonalized Momentum

Minxin Zhang minxinzhang@math.ucla.edu
Yuxuan Liu yxliu@math.ucla.edu
Hayden Schaeffer hayden@math.ucla.edu
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095, USA

Abstract

Efficient stochastic optimization typically integrates an update direction that performs well in the deterministic regime with a mechanism adapting to stochastic perturbations. While Adam uses adaptive moment estimates to promote stability, Muon utilizes the weight layers' matrix structure via orthogonalized momentum, showing superior performance in large language model training. We propose a new optimizer and a diagonal extension, NAMO and NAMO-D, providing the first principled integration of orthogonalized momentum with norm-based Adam-type noise adaptation. NAMO scales orthogonalized momentum using a single adaptive stepsize, preserving orthogonality while improving upon Muon at negligible additional cost. NAMO-D instead right-multiplies orthogonalized momentum by a diagonal matrix with clamped entries. This design enables neuron-wise noise adaptation and aligns with the common near block-diagonal Hessian structure. Under standard assumptions, we establish optimal convergence rates for both algorithms in the deterministic setting and show that, in the stochastic setting, their convergence guarantees adapt to the noise level of stochastic gradients. Experiments on pretraining GPT-2 models demonstrate improved performance of both NAMO and NAMO-D compared to the AdamW and Muon baselines, with NAMO-D achieving further gains over NAMO via an additional clamping hyperparameter that balances the competing goals of maintaining a well-conditioned update direction and leveraging fine-grained noise adaptation.

Keywords: Stochastic optimization, Muon optimizer, Adam optimizer, noise-adaptive convergence, large language model pretraining

1 Introduction

Stochastic optimization is central to modern large-scale learning, where algorithms must update iterates using only noisy estimates of first-order information. A key challenge is to balance two goals: selecting an update direction that is effective in the noise-free regime, and incorporating mechanisms that adapt to stochastic perturbations. From this viewpoint, an efficient stochastic optimizer can be understood as combining two ingredients: (i) a principled direction-selection rule that performs well when gradients are exact, and (ii) an adaptive stepsize mechanism that stabilizes the iterates by attenuating updates under gradient uncertainty. This perspective offers a useful lens for interpreting existing methods and for guiding the development of new algorithms with strong convergence behavior. In this paper, we design a new optimization algorithm that couple the structural advantages of an orthogonalized update direction with adaptive moment estimation to account for gradient noise. We provide theoretical convergence guarantees and demonstrate strong empirical performance in training large language models (LLMs).

The code is available at: `https://github.com/minxin-zhg/namo`.

Adaptive methods such as Adam (Kingma and Ba, 2014) and its decoupled weight-decay variant AdamW (Loshchilov and Hutter, 2017) have long been the de facto optimizers for large-scale training. Their coordinate-wise adaptive stepsizes promote training stability and reduce sensitivity to hyperparameter choices. Given momentum coefficients $\beta_1, \beta_2 \in [0, 1)$, Adam iteratively updates a biased first-moment estimate of the stochastic gradient:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t,$$

and a biased second raw-moment estimate:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2,$$

where g_t denotes the stochastic gradient at iteration t, and the square is applied elementwise. With the standard bias corrections $\hat{m}_t := m_t/(1 - \beta_1^t)$ and $\hat{v}_t := v_t/(1 - \beta_2^t)$, the parameters are updated via:

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon},$$

where $\epsilon > 0$ is a small fixed scalar that avoids division by zero, and $\eta > 0$ is the learning rate. Throughout the paper, we refer to the hyperparameter η as the *learning rate*, and to its combination with an adaptive scaling as the effective *stepsize*. The ratio of the first- and second-moment estimates, $\hat{m}_t/\sqrt{\hat{v}_t}$, is often interpreted as a signal-to-noise ratio (SNR). When the update direction is dominated by noise, or as the iterates approach a stationary point, the resulting effective stepsize decreases, which is desirable for stable convergence. When $\beta_1 = \beta_2$, (Orvieto and Gower, 2025) interprets Adam as performing online estimation of the mean and variance of the stochastic gradients. Moreover, Adam's update rule can be viewed as combining a sign descent direction with a variance adaptation component. Theoretical analysis and experiments in (Balles and Hennig, 2018) show that the sign-descent component can lead to Adam's adverse generalization behavior (Wilson et al., 2017), indicating that coupling Adam-type noise adaptation with a different descent direction may yield improved generalization relative to Adam.

While Adam and other standard variants of stochastic gradient descent (SGD) treat trainable parameters of neural networks as flattened vectors, the Muon optimizer (Jordan et al., 2024b) exploits their matrix structure by updating weight matrices with orthogonalized momentum. Given a matrix $M \in \mathbb{R}^{m \times n}$, if $M = U\Sigma V^T$ is its reduced singular value decomposition (SVD), then its orthogonalization is given by:

$$\mathrm{Orth}\,(M) := UV^T.$$

The orthogonalization of M, $\mathrm{Orth}\,(M)$, is also called the orthogonal factor in the polar decomposition and is the nearest orthogonal matrix to M in the Frobenius norm (Higham, 2008). Muon updates matrix-structured parameters $\Theta \in \mathbb{R}^{m \times n}$ by:

$$\Theta_t = \Theta_{t-1} - \eta O_t,$$

where $\eta > 0$ is the prescribed learning rate, and $O_t \approx \mathrm{Orth}\,(M_t)$ denotes an approximate orthogonalization of the momentum matrix M_t at iteration t, computed via Newton-Schultz iterations (Bernstein, 2025). Growing evidence demonstrates that Muon achieves superior empirical performance compared to Adam in LLM training (Jordan et al., 2024a; Team et al., 2025; Liu et al., 2025). Recent studies show that orthogonalized updates can accelerate convergence, facilitate more reliable hyperparameter transfer across model sizes (Boreiko et al., 2025; Shah et al., 2025), and learn more effectively from heavy-tailed data (Wang et al., 2025). In the deterministic setting

without momentum, Muon's orthogonalized update direction can be interpreted as the steepest descent direction under the spectral norm, which is shown to be effective for training in deep learning (Davis and Drusvyatskiy, 2025). In stochastic settings, however, matrix orthogonalization is an unbounded operation (Higham, 1986) that may amplify the impact of noise in the original momentum matrix, leading to training instability (He et al., 2025) and increased sensitivity to hyperparameter choices (Crawshaw et al., 2025). This suggests that pairing Muon's orthogonalized updates with an explicit noise-adaptation mechanism could improve training robustness and further boost performance.

In this work, we develop a theoretically principled integration of Adam-type variance adaptation with an orthogonalized update direction. Prior work shows that orthogonalization decouples the direction and magnitude of a matrix update through a norm-duality characterization (Bernstein and Newhouse, 2024, Proposition 5), suggesting that a norm-based rescaling of the orthogonalized update is a natural design choice. We pair norm-based moment estimation with orthogonalized momentum and propose a new optimizer, NAMO (**N**orm-Based **A**daptive **M**oment Estimation with **O**rthogonalized Momentum), together with a diagonal extension, NAMO-D. NAMO scales the orthogonalized momentum with a norm-based adaptive scalar, preserving the orthogonality of the update direction while adapting the effective stepsize to the noise level. NAMO-D scales the orthogonalized momentum with a right-multiplied diagonal matrix, allowing an individual adaptive stepsize for each neuron. We establish theoretical convergence rates for both algorithms, and demonstrate that they outperform the AdamW and Muon baselines in GPT-2 pretraining.

1.1 Related work

Unlike Adam's coordinate-wise adaptive stepsizes, our methods introduce structured stepsize adaptation for orthogonalized updates: NAMO scales the orthogonalized momentum using a single adaptive stepsize, whereas NAMO-D employs a column-wise adaptive stepsize for the orthogonalized momentum. Related simplifications of Adam have also been studied. In (Chezhegov et al., 2024), a clipped norm-based Adam-type stepsize is applied to the momentum, and convergence is analyzed under the assumptions of bounded gradients and heavy-tailed noise. Motivated by the near block-diagonal Hessian structure of neural networks, Adam-mini (Zhang et al., 2024) partitions parameters into blocks and assigns a single adaptive learning rate per block, matching Adam's performance while reducing memory cost.

Several adaptive variants of Muon have been proposed as well. AdaMuon (Si et al., 2025) and NorMuon (Li et al., 2025) combine Muon's orthogonalized momentum with Adam-type scaling variants, though without theoretical convergence guarantees. AdaGO (Zhang et al., 2025) scales the orthogonalized momentum with an adaptively decaying stepsize and achieves optimal theoretical convergence rates, although its performance in LLM training has yet to be investigated. Layerwise adaptive learning rates for Muon are proposed in (Hao et al., 2025) based on gradient-variance estimation that requires evaluating a nuclear norm, thereby increasing the per-iteration cost. Tuning-robust variants of Muon are explored in (Crawshaw et al., 2025). PRISM (Yang, 2026) augments Muon with a moment-based adaptive preconditioner, but incurs higher additional computational cost than our proposed algorithms, and does not provide theoretical convergence guarantees. DeVA (Song et al., 2026) decouples variance adaptation from scale-invariant sign descent, yielding an Adam-style scaling for Muon's orthogonalized momentum that require high computational and memory overhead for maintaining Kronecker preconditioners and periodic eigendecompositions.

1.2 Contributions and Organization

We propose a new optimization algorithm and a diagonal extension for problems with matrix-structured parameters, NAMO and NAMO-D, which provide the first theoretically principled integration of an orthogonalized update direction with noise adaptation based on Adam-type moment estimation. NAMO scales the orthogonalized momentum using a single adaptive stepsize, incurring only a negligible $\mathcal{O}(mn)$ additional computational cost and no additional memory overhead, thus providing a useful improvement for Muon's performance. On the other hand, NAMO-D scales the orthogonalized momentum by right-multiplying with a diagonal matrix D_t, thereby assigning a neuron-wise adaptive stepsize determined by the column norms of the stochastic gradient and momentum matrices. This column-wise scaling enables finer-grained noise adaptation and aligns with the near block-diagonal Hessian structure commonly observed in neural networks (Dong et al., 2025; An et al., 2025), while no longer strictly preserving the orthogonalized update direction. The diagonal entries of D_t admit an upper bound and are clamped toward the average, ensuring that NAMO-D's update direction remains well-conditioned. Under standard smoothness and unbiased bounded-variance noise assumptions, we establish optimal convergence rates for both algorithms in the deterministic setting and show that, in the stochastic setting, their convergence guarantees adapt to the noise level of the stochastic gradients and attain the optimal rate when the batch size is sufficiently large. Experiments on pretraining GPT-2 models demonstrate that both NAMO and NAMO-D outperform AdamW and Muon baselines. NAMO-D achieves further gains over NAMO through an additional clamping hyperparameter c, which balances the competing goals of maintaining a well-conditioned update direction and leveraging finer-grained noise adaptation.

The rest of the paper is organized as follows. Section 2 describes the proposed algorithms, and Section 3 provides the convergence analysis, with detailed proofs deferred to Appendices A–E. Section 4 presents experiments on GPT-2 pretraining, and Section 5 concludes the paper.

2 New Optimization Algorithms: NAMO and NAMO-D

In this section, we describe the proposed algorithm NAMO and its diagonal extension NAMO-D. We consider optimization problems with matrix-structured parameters of the form:

$$\min_{\Theta \in \mathbb{R}^{m\times n}} \mathcal{L}(\Theta),$$

where $\mathcal{L} : \mathbb{R}^{m\times n} \to \mathbb{R}$ is the (nonconvex) loss function. We impose the following standard smoothness of loss function and bounded-variance noise assumptions.

Assumption 1 *The gradient of $\mathcal{L}(\Theta)$ is Lipschitz continuous, i.e., for arbitrary $\Theta, \Theta' \in \mathbb{R}^{m\times n}$:*

$$\left\|\nabla\mathcal{L}(\Theta) - \nabla\mathcal{L}(\Theta')\right\|_* \le L \left\|\Theta - \Theta'\right\|_2, \tag{1}$$

for some constant $L > 0$, where $\|\cdot\|_*$ and $\|\cdot\|_2$ denote the nuclear norm and the spectral norm respectively.

Note that gradient of the loss, $\nabla\mathcal{L}(\Theta)$, is in $\mathbb{R}^{m\times n}$ and recall that the nuclear norm is the sum of the singular values while the spectral norm is the maximum singular value.

Assumption 2 *At each iteration t, the stochastic gradient G_t is an unbiased estimate of the true gradient, i.e., $\mathbb{E}[G_t] = \nabla\mathcal{L}(\Theta_{t-1})$, with a uniformly bounded variance:*

$$\mathbb{E}\left[\|G_t - \nabla\mathcal{L}(\Theta_{t-1})\|_F^2\right] \le \frac{\sigma^2}{b},$$

where $b \geq 1$ is the batch size and $\|\cdot\|_F$ denotes the Frobenius norm.

Assumption 1 is equivalent to a more commonly seen smooth assumption:

$$\left\|\nabla\mathcal{L}(\Theta) - \nabla\mathcal{L}(\Theta')\right\|_F \leq L' \left\|\Theta - \Theta'\right\|_F \tag{2}$$

for a different Lipschitz constant $L' > 0$. Since orthogonalized gradient descent can be interpreted as steepest descent under the spectral norm (Shen et al., 2025), we state the smoothness assumption in the equivalent form (1).

Details of NAMO and NAMO-D are summarized in Algorithms 1 and 2, respectively. For the theoretical analysis in Section 3, we assume exact orthogonalization in both algorithms, as is standard in existing works (Shen et al., 2025; Sato et al., 2025; Li and Hong, 2025; Chen et al., 2025; Kovalev, 2025; Pethick et al., 2025). In practice, exact orthogonalization can be expensive to compute, and for the experiments in Section 4 we use Newton–Schulz iterations to obtain an approximate orthogonalization, as in the Muon optimizer.

NAMO maintains a biased second raw-moment estimate of the squared Frobenius norm:

$$v_t = \mu_2 v_{t-1} + (1-\mu_2)\left\|G_t\right\|_F^2,$$

where G_t is the stochastic gradient matrix at iteration t, and the momentum matrix M_t is a biased first moment estimate of the stochastic gradient. Applying bias correction yields $\hat{M}_t := M_t/(1-\mu_1^t)$ and $\hat{v}_t := v_t/(1-\mu_2^t)$. The parameters $\Theta \in \mathbb{R}^{m\times n}$ are updated by:

$$\Theta_t = \Theta_{t-1} - \eta\alpha_t O_t, \tag{3}$$

where $\eta > 0$ is a prescribed learning rate, and the orthogonalized momentum O_t is adaptively scaled by a scalar:

$$\alpha_t := \frac{(1-\mu_2^t)^{\frac{1}{2}}}{1-\mu_1^t}\frac{\|M_t\|}{\sqrt{v_t}+\epsilon} = \frac{\left\|\hat{M}_t\right\|}{\sqrt{\hat{v}_t}+\epsilon_t},$$

with $\epsilon_t := \epsilon/\sqrt{1-\mu_2^t}$ for a small $\epsilon > 0$. When stochastic gradients are noisy, or when the iterates approach a stationary point, α_t is small, promoting stable convergence.

Algorithm 1 NAMO: **N**orm-Based **A**daptive **M**oment Estimation with **O**rthogonalized Momentum

Require: Learning rate η, momentum $\mu_1, \mu_2 \in [0,1)$, batch size b, $\epsilon > 0$

Initialize $\Theta_0 \in \mathbb{R}^{m\times n}$, $M_0 = 0$, $v_0 = 0$
for $t = 1, 2, \ldots, T$ **do**
  Sample a minibatch of size b and compute stochastic gradient $G_t = \nabla\mathcal{L}_t(\Theta_{t-1})$
  $M_t \leftarrow \mu_1 M_{t-1} + (1-\mu_1)G_t$
  $v_t \leftarrow \mu_2 v_{t-1} + (1-\mu_2)\left\|G_t\right\|_F^2$
  $O_t \leftarrow \mathrm{Orth}\left(M_t\right)$
  $\alpha_t \leftarrow \frac{\sqrt{1-\mu_2^t}}{1-\mu_1^t}\frac{\|M_t\|_F}{\sqrt{v_t}+\epsilon}$
  Update parameters $\Theta_t \leftarrow \Theta_{t-1} - \eta\alpha_t O_t$
end for
return Θ_T

While NAMO scales the orthogonalized momentum using a single adaptive stepsize, NAMO-D applies a column-wise scaling based on the column norms of the momentum matrix. This design

assigns an individual adaptive stepsize to each neuron and is consistent with the near block-diagonal Hessian structure commonly observed in neural networks (Dong et al., 2025; An et al., 2025). For each $j = 1, 2, \cdots, n$, NAMO-D maintains a biased second raw-moment estimate of the squared norm of the j-th colum of the stochastic gradient:

$$[\mathbf{v}_t]_j = \mu_2 [\mathbf{v}_t]_{j-1} + (1-\mu_2) \left\| [G_t]_{:j} \right\|^2.$$

To write this in an equivalent vector form, define the operator $\mathcal{N}_c : \mathbb{R}^{m\times n} \to \mathbb{R}^n$ by:

$$[\mathcal{N}_c(M)]_j := \|M_{:j}\|, \quad j = 1, 2, \cdots, n,$$

where $\|\cdot\|$ denotes the Euclidean norm. Then the iterates can be written as:

$$\mathbf{v}_t = \mu_2 \mathbf{v}_{t-1} + (1-\mu_2)\mathcal{N}_c(G_t) \odot \mathcal{N}_c(G_t).$$

Applying bias correction yields $\hat{M}_t := M_t/(1-\mu_1^t)$ and $\hat{\mathbf{v}}_t := \mathbf{v}_t/(1-\mu_2^t)$. Now let:

$$\mathbf{d}_t = \mathcal{N}_c(\hat{M}_t) \oslash \left(\sqrt{\hat{\mathbf{v}}_t} + \epsilon_t \right),$$

where $\epsilon_t := \epsilon/\sqrt{1-\mu_2^t}$ for a small $\epsilon > 0$, and $\oslash$ denotes entrywise division. By Lemma 7, the entries of $\mathbf{d}_t$ are bounded above by:

$$\|\mathbf{d}_t\|_\infty = \max_j [\mathbf{d}_t]_j \le \sqrt{\frac{1-\mu_1}{1-\mu_2}}.$$

Compared with NAMO, NAMO-D's column-wise scaling enables finer-grained noise adaptation, but it no longer strictly preserves orthogonality of the update direction. Let $\bar{d}_t := \|\mathbf{d}_t\|_1 / n$ be the average of the entries of $\mathbf{d}_t$. To ensure that the scaled direction remains well-conditioned, we clamp the adaptive stepsizes toward this average via:

$$\tilde{\mathbf{d}}_t := \min\left\{ \max\left\{ \mathbf{d}_t, c\bar{d}_t \mathbf{1} \right\}, \bar{d}_t/c \right\},$$

for a prescribed constant $c \in (0, 1]$, where the maximum and minimum operations are applied entrywise, and $\mathbf{1} \in \mathbb{R}^n$ denotes the vector of all-ones. Let $D_t := \operatorname{diag}\left(\tilde{\mathbf{d}}_t\right)$. NAMO-D updates the parameters by:

$$\Theta_t = \Theta_{t-1} - \eta O_t D_t.$$

This update rule combines column-wise, noise-adaptive scaling with a simple clamping safeguard, tempering updates when gradients are noisy or the iterates are near a stationary point, while keeping the scaled update direction well-conditioned.

3 Convergence Analysis

In this section, we establish convergence rates for NAMO (Algorithm 1) and NAMO-D (Algorithm 2) in both the deterministic and stochastic settings under the standard Assumptions 1–2. The analysis of NAMO is presented in Section 3.1, and the analysis of NAMO-D is presented in Section 3.2. We provide a proof sketch for each of the theorems below and include detailed proofs in Appendices B–E. Proofs of useful lemmata are in Appendix A. The proofs repeatedly use the bias-corrected moment estimates and their convex-combination representations with coefficients:

$$w_{1,t,\tau} := \frac{1-\mu_1}{1-\mu_1^t}\mu_1^{t-\tau}, \qquad w_{2,t,\tau} := \frac{1-\mu_2}{1-\mu_2^t}\mu_2^{t-\tau}, \quad \text{for } \ t \ge 1 \text{ and } \tau \in \{1, \ldots, t\},$$

Algorithm 2 NAMO-D: Diagonal exension of NAMO

Require: Learning rate η, momentum $\mu_1, \mu_2 \in [0, 1)$, batch size b, $\epsilon > 0$, $c \in (0, 1]$

Initialize $\Theta_0 \in \mathbb{R}^{m\times n}$, $M_0 = 0$, $\mathbf{v}_0 = 0$
for $t = 1, 2, \ldots, T$ **do**
  Sample a minibatch of size b and compute stochastic gradient $G_t = \nabla \mathcal{L}_t(\Theta_{t-1})$
  $M_t \leftarrow \mu_1 M_{t-1} + (1 - \mu_1)G_t$
  $\mathbf{v}_t \leftarrow \mu_2 \mathbf{v}_{t-1} + (1 - \mu_2)\mathcal{N}_c(G_t) \odot \mathcal{N}_c(G_t)$
  $\mathbf{d}_t \leftarrow \frac{\sqrt{1-\mu_2^t}}{1-\mu_1^t}\mathcal{N}_c(M_t) \oslash \left(\sqrt{\mathbf{v}_t} + \epsilon\right)$
  $\bar{d}_t \leftarrow \|\mathbf{d}_t\|_1 / n$
  $O_t \leftarrow \mathrm{Orth}\left(M_t\right)$
  $D_t \leftarrow \mathrm{diag}\left(\min\left\{\max\left\{\mathbf{d}_t, c\bar{d}_t\mathbf{1}\right\}, \frac{1}{c}\bar{d}_t\mathbf{1}\right\}\right)$
  Update parameters $\Theta_t \leftarrow \Theta_{t-1} - \eta O_t D_t$
end for
return Θ_T

which satisfy $\sum_{\tau=1}^{t} w_{1,t,\tau} = \sum_{\tau=1}^{t} w_{2,t,\tau} = 1$. This representation allows one to control the deviation between the bias-corrected momentum and the current gradient via Assumption 1, as well as the magnitude of the bias-corrected second-moment estimate by elementary geometric-series bounds.

3.1 Analysis of NAMO

We first analyze the convergence of NAMO in the deterministic case where gradients are evaluated exactly. The $\mathcal{O}(T^{-1/2})$ rate established in the theorem below is the best possible for deterministic first-order methods under Assumption 1; see (Carmon et al., 2020, Theorem 2).

Theorem 3 (NAMO in the deterministic case) *Suppose Assumptions 1 holds. Let $\{\Theta_t\} \subset \mathbb{R}^{m\times n}$ be the sequence of iterates generated by Algorithm 1 with full-batch gradients and $0 \le \mu_1 \le \mu_2 < 1$. If choosing $\eta = \mathcal{O}\left(T^{-\frac{1}{2}}\right)$, $\epsilon = \mathcal{O}(T^{-\frac{1}{2}})$, $\mu_1 = \Theta(1)$ and $\mu_2 = \Theta(1)$, then for large $T > 0$:*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\Theta_{t-1})\|_F \le \mathcal{O}\left(T^{-\frac{1}{2}}\right).$$

Proof sketch. Let $\hat{M}_t := M_t/(1-\mu_1^t)$ and $\hat{v}_t := v_t/(1-\mu_2^t)$ denote the bias-corrected moments, so that $\hat{M}_t = \sum_{\tau=1}^{t} w_{1,t,\tau}\nabla L(\Theta_{\tau-1})$ and $\hat{v}_t = \sum_{\tau=1}^{t} w_{2,t,\tau}\|\nabla L(\Theta_{\tau-1})\|_F^2$. Define the adaptive scaling α_t as in (3). Applying Lemma 7 to the vectorization of the gradient matrix sequence $\{\nabla L(\Theta_{\tau-1})\}$ yields a uniform upper bound on α_t (Appendix B, Step 1). Next, using the orthogonalized descent inequality from (Zhang et al., 2025, Lemma B.1) and Assumption 1, one obtains an average inequality involving:

$$\frac{\|\hat{M}_t\|_F \|\nabla L(\Theta_{t-1})\|_*}{\sqrt{\hat{v}_t} + \epsilon/\sqrt{1-\mu_2^t}} \quad \text{and} \quad \|\nabla L(\Theta_{t-1}) - \hat{M}_t\|_*$$

(Appendix B, Step 2). The deviation of the bias-corrected momentum from the gradient is then controlled using the convex-combination representation of $\hat{M}_t$ together with Assumption 1 and a telescoping bound on $\|\Theta_{\tau-1} - \Theta_{t-1}\|_2$ (Appendix B, Step 3). Similarly, $\sqrt{\hat{v}_t}$ is upper bounded by $\|\nabla L(\Theta_{t-1})\|_*$ plus a constant obtained from geometric-series estimates (Appendix B, Step 4). Combining these bounds and applying Lemma 8 converts the bound on an averaged surrogate ϕ_ϵ

into the stated bound on the average gradient norm (Appendix B, Steps 5–6). The full proof is provided in Appendix B. ■

Next we analyze NAMO in the stochastic case where stochastic gradients have bounded variance. The best possible convergence rate for stochastic first-order methods under Assumptions 1–2 is $\mathcal{O}(T^{-1/4})$; see (Arjevani et al., 2023, Theorem 3). The following result shows that NAMO adapts to the noise level of the stochastic gradients and achieves the optimal convergence rate when the batch size is sufficiently large.

Theorem 4 (NAMO in the stochastic case) *Suppose Assumptions 1–2 hold. Let $\{\Theta_t\} \subset \mathbb{R}^{m\times n}$ be the sequence of iterates generated by Algorithm 1. For large $T > 0$, if we set $\eta = \mathcal{O}(T^{-\frac{3}{4}})$, $1-\mu_1 = \Theta(T^{-\frac{1}{2}})$, $1-\mu_2 = \Theta(T^{-\frac{1}{2}})$, $0 \le \mu_1 \le \mu_2 < 1$, and $\epsilon = \mathcal{O}(T^{-\frac{1}{2}})$, then:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla \mathcal{L}(\Theta_{t-1})\|_F\right] \le \mathcal{O}\left(T^{-\frac{1}{4}} + \sqrt{\sigma} b^{-\frac{1}{4}} T^{-\frac{1}{8}}\right),$$

where b is the batch size.

Proof sketch. The proof again works with the convex-combination representations of the bias-corrected moment estimates $\hat{M}_t$ and $\hat{v}_t$ with weight coefficients $w_{1,t,\tau}$ and $w_{2,t,\tau}$ respectively. Let $\mathbb{E}_t[\cdot]$ be the expectation conditioned on the iterate Θ_{t-1}, and define the deviation term $E_t := \hat{M}_t - \nabla L(\Theta_{t-1})$. Using (Zhang et al., 2025, Lemma B.1) and taking conditional expectations yields an average bound on $\mathbb{E}[\alpha_t \|\hat{M}_t\|_*]$ in terms of $\mathbb{E}\left[\|E_t\|_*\right]$ and $\mathbb{E}\left[\alpha_t^2\right]$ (Appendix C, Step 2). The quantity $\mathbb{E}\left[\|E_t\|_*\right]$ is controlled by decomposing E_t into a weighted sum of noise terms $G_\tau - \nabla L(\Theta_{\tau-1})$ and a drift term $\nabla L(\Theta_{\tau-1}) - \nabla L(\Theta_{t-1})$. The noise term is bounded above via Assumption 2, the drift term is bounded above using Assumption 1 along with a telescoping argument, and then Lemma 9 is applied to bound the averaging over t (Appendix C, Step 3). In parallel, Minkowski and Jensen's inequalities yield an upper bound on the expected denominator term $\mathbb{E}\left[\sqrt{\hat{v}_t} + \epsilon/\sqrt{1-\mu_2^t}\right]$ by $\mathbb{E}\left[\|\nabla L(\Theta_{t-1})\|\right]$ plus an explicit additive term containing $\sigma/\sqrt{b}$, then Lemma 10 is applied to bound the averaging over t (Appendix C, Step 4). Finally, a Cauchy–Schwarz argument relates $\mathbb{E}[\alpha_t \|\hat{M}_t\|_F]$ to $\mathbb{E}[\|\nabla L(\Theta_{t-1})\|_F]$, leading to a quadratic inequality in $\mathbb{E}\left[\|\nabla L(\Theta_{t-1})\|_F\right]$ that can be solved explicitly (Appendix C, Step 5). Substituting the averaged bounds and the parameter choices yields the stated rate. The full proof is provided in Appendix C. ■

This result indicates that the convergence rate of NAMO is adaptive to the noise level of the stochastic gradients, and that choosing the batch size as $b = \Omega(\sigma^2\sqrt{T})$ recovers the optimal $\mathcal{O}(T^{-1/4})$ rate.

3.2 Analysis of NAMO-D

NAMO-D introduces a column-wise diagonal scaling D_t, defined in Algorithm 2, instead of the scalar α_t. We first establish its optimal $\mathcal{O}\left(T^{-1/2}\right)$ rate for the deterministic case where gradients are evaluated exactly in the theorem below.

Theorem 5 (NAMO-D in the deterministic case) *Suppose Assumption 1 holds. Let $\{\Theta_t\} \subset \mathbb{R}^{m\times n}$ be the sequence of iterates generated by Algorithm 2 using full batch with $0 \le \mu_1 \le \mu_2 < 1$. If we choose $\eta = \mathcal{O}(T^{-\frac{1}{2}})$, $1-\mu_1 = \Theta(1)$, $1-\mu_2 = \Theta(1)$, $0 \le \mu_1 \le \mu_2 < 1$, $\epsilon = \mathcal{O}(T^{-\frac{1}{2}} n^{-1})$, and*

$c = \Theta(1)$, *then:*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\Theta_{t-1})\|_F \leq \mathcal{O}\left(T^{-\frac{1}{2}}\right),$$

for large $T > 0$.

Proof sketch. The proof works with the diagonal scaling matrix D_t defined as in Section 2 and its extremal diagonal entries $d_{t,\max}$ and $d_{t,\min}$. Applying Lemma 7 column-wise yields a uniform upper bound on the entries of D_t (Appendix D, Step 1). The clamping rule implies $d_{t,\min} \geq c^2\, d_{t,\max}$ and hence bounds the condition number of D_t. Applying the orthogonalized descent inequality (Zhang et al., 2025, Lemma B.1) gives an average bound on $d_{t,\max}\|\nabla L(\Theta_{t-1})\|_*$ in terms of the deviation $\left\|\nabla L(\Theta_{t-1}) - \hat{M}_t\right\|_*$ (Appendix D, Step 2). The deviation term and the term $\sqrt{\sum_{j=1}^{n} \hat{v}_t^j}$ are controlled using similar arguments as in the proof of Theorem 3, together with geometric-series bounds (Appendix D, Steps 3–4). The result again follows by combining these bounds and applying Lemma 8 (Appendix D, Steps 5–6). The full proof is provided in Appendix D. ■

In the stochastic case, we show that the convergence of NAMO-D adapts to the noise level of stochastic gradients, and achieves the optimal $\mathcal{O}\left(T^{-1/4}\right)$ rate when the batch size is sufficiently large.

Theorem 6 (NAMO-D in the stochastic case) *Suppose Assumptions 1–2 holds. Let* $\{\Theta_t\} \subset \mathbb{R}^{m\times n}$ *be the sequence of iterates generated by Algorithm 2. For large* $T > 0$, *if we set* $\eta = \mathcal{O}(T^{-\frac{3}{4}})$, $1-\mu_1 = \Theta(T^{-\frac{1}{2}})$, $1-\mu_2 = \Theta(T^{-\frac{1}{2}})$, $0 \leq \mu_1 \leq \mu_2 < 1$, $\epsilon = \mathcal{O}(T^{-\frac{1}{2}})$, *and* $c = \Theta(1)$, *then:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla \mathcal{L}(\Theta_{t-1})\|_F\right] \leq \mathcal{O}\left(T^{-\frac{1}{4}} + \sqrt{\sigma} b^{-\frac{1}{4}} T^{-\frac{1}{8}}\right),$$

where b *is the batch size.*

Proof sketch. Define $\hat{M}_t$, $\hat{v}_t$, and D_t as before, and let $\mathbb{E}_t[\cdot]$ denotes the expectation conditioned on iterate Θ_{t-1}, and $E_t := \hat{M}_t - \nabla L(\Theta_{t-1})$. The expected descent step uses Lemma 11 to lower bound $\langle \hat{M}_t, O_t D_t\rangle$ by $d_{t,\min}\left\|\hat{M}_t\right\|_*$, and the clamping rule controls the ratio of $d_{t,\max}$ and $d_{t,\min}$ (Appendix E, Step 2). The averaged deviation term $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]$ is bounded by the same variance-plus-drift decomposition as in the proof of Theorem 4, using Assumption 2 to control the noise term and Assumption 1 to control the drift term via a telescoping argument; then averaging uses Lemma 9 (Appendix E, Step 3). For the denominator, Minkowski and Jensen's inequalities yield an upper bound on $\mathbb{E}[\sqrt{\sum_{j=1}^{n}\hat{v}_t^j}]$ by $\mathbb{E}\left[\|\nabla L(\Theta_{t-1})\|\right]$ plus an explicit additive term containing $\sigma/\sqrt{b}$; then the averaging over t is bounded using Lemma 10 (Appendix E, Step 4). Finally, a Cauchy–Schwarz argument relate $\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]$ to the expected gradient norm, yielding a quadratic inequality in $\mathbb{E}\left[\|\nabla L(\Theta_{t-1})\|\right]$ that can be solved explicitly (Appendix E, Step 5). Combining these bounds with the specified hyperparameter choices gives the stated convergence rate (Appendix E, Step 6). The full proof is provided in Appendix E. ■

This result indicates that the convergence rate of NAMO-D, like that of NAMO, is adaptive to the noise level of the stochastic gradients, and that the optimal $\mathcal{O}(T^{-1/4})$ rate is recovered when choosing the batch size as $b = \Omega(\sigma^2\sqrt{T})$,

The results of this section demonstrate that both NAMO and NAMO-D can achieve optimal convergence rates in terms of the order on T. This occurs in both the deterministic and stochastic settings under Assumptions 1-2. In the deterministic setting, the $\mathcal{O}(T^{-1/2})$ bound matches the known lower complexity bound for smooth nonconvex optimization with first-order methods, thus demonstrating that utilizing orthogonalized update and adaptive scaling does not incur any deterioration in the convergence or complexity. In the stochastic regime, the bounds decompose naturally into an optimization term of order $\mathcal{O}(T^{-1/4})$ and an explicit variance-dependent term scaling as $\mathcal{O}\left(\sqrt{\sigma}\, b^{-1/4} T^{-1/8}\right)$, and thus quantifying the precise affect of gradient noise and batch size. Specifically, when $b = \Omega(\sigma^2 T^{1/2})$, the variance term becomes asymptotically dominated and the optimal $\mathcal{O}(T^{-1/4})$ rate is recovered. The analysis further shows that the orthogonalized descent inequality, combined with bias-corrected moment estimates and controlled adaptive scaling (scalar or diagonal), yields a unified proof strategy that is robust to both drift and stochastic perturbations. Altogether, these guarantees provide a rigorous complexity characterization of the proposed methods and confirm that their matrix-aware adaptive scaling preserves optimal first-order performance while maintaining stability under noise.

4 Experiments

4.1 Experimental setup

Baselines. We compare our proposed algorithms, NAMO and NAMO-D, against two popular baselines: AdamW (Loshchilov and Hutter, 2017) and Muon (Jordan et al., 2024b). Since Muon, NAMO, and NAMO-D are designed specifically for matrix parameters, we use AdamW to optimize all remaining scalar and vector parameters across all models. For brevity, we refer to these hybrid optimizers simply as Muon, NAMO, and NAMO-D.

Model Architectures. We evaluate all optimizers on GPT-2 pretraining experiments, which are based on the nanoGPT (Karpathy, 2022) implementation of the GPT-2 architecture (Radford et al., 2019). Two model sizes are considered: small (124M parameters), and medium (355M parameters). All experiments are conducted on 4× NVIDIA H100 GPUs.

Dataset. All experiments are conducted on the OpenWebText dataset (Gokaslan et al., 2019), which contains approximately 9 billion training tokens and 4.4 million validation tokens.

Training details. We use standard hyperparameter settings for the baselines. For AdamW, we set the momentum coefficients to $\beta_1 = 0.9$ and $\beta_2 = 0.95$. For Muon, the momentum coefficient is set to $\beta = 0.95$. For NAMO and NAMO-D, the momentum coefficients are set to $\mu_1 = 0.95$ and $\mu_2 = 0.99$. The weight decay coefficient is set to $\lambda = 0.01$ for all optimizers. The Muon optimizer applies decoupled weight decay as AdamW:

$$\Theta_t = \Theta_{t-1} - \eta \left(O_t + \lambda \Theta_{t-1}\right),$$

where η is the prescribed learning rate. For NAMO and NAMO-D, as part of the effective stepsize, the adaptive scaling is applied to the decoupled weight decay as well. Specifically, NAMO updates by:

$$\Theta_t = \Theta_{t-1} - \eta \alpha_t \left(O_t + \lambda \Theta_{t-1}\right),$$

where α_t is given in Algorithm 1, and NAMO-D updates by:

$$\Theta_t = \Theta_{t-1} - \eta \left(O_t + \lambda \Theta_{t-1}\right) D_t,$$

where D_t is given in Algorithm 2. For GPT-2 (124M) model pretraining, we train with context length $L = 1024$ and an effective batch size of 480 sequences (491,520 tokens) per optimizer update for all optimizers (micro-batch size $B = 60$ and gradient accumulation $K = 8$). For GPT-2 (355M) model, we also train with context length $L = 1024$ and an effective batch size of 480 sequences (491,520 tokens) per optimizer update for all optimizers (micro-batch size $B = 40$ and gradient accumulation $K = 12$). For both models and all optimziers, we use the following learning rate scheduler: 2000-step linear warm-up followed by constant learning rate. Also, for both models, we sweep for optimal learning rate η for each optimizer through a grid search. For NAMO-D, we fix the clamping hyperparameter $c = 0.1$ for GPT-2 (124M) experiments, and sweep for optimal c for GPT-2 (355M) experiments.

4.2 Empirical Results on GPT-2

For pretraining the GPT-2 (124M) model, we perform a grid search for each optimizer to find the optimal learning rate (LR) that achieves the lowest validation loss after 10K steps. The sweeping results are presented in Figure 1, which shows that NAMO and NAMO-D achieve lower training and validation losses across a wider range of learning-rate choices, demonstrating both accelerated convergence and improved tuning robustness comparing to Muon and AdamW.

The optimal learning rate selected from the 10K-step sweep for each optimizer is listed in Table 1. While NAMO and NAMO-D exhibit similar training and validation losses in the 10K-step learning-rate sweep, Figure 2 shows that when training is extended to 50K steps using the selected optimal learning rate, NAMO-D attains lower training and validation losses than NAMO, demonstrating advantages of its finer-grained neuron-wise adaptive scaling. The training and validation losses at termination for each optimizer are reported in Table 2.

We further conduct experiments on pretraining the GPT-2 (355M) model. For each optimizer, we sweep from five learning rates that are around the chosen optimal learning rate for GPT-2 (124M) model listed in Table 1. Specifically, for Muon and AdamW, we sweep for the optimal LR from the following set:

$$\{0.0006, 0.0009, 0.0013, 0.0018, 0.0025\};$$

for NAMO and for NAMO-D, we sweep for the optimal LR from the following set:

$$\{0.005, 0.007, 0.009, 0.012, 0.015\}.$$

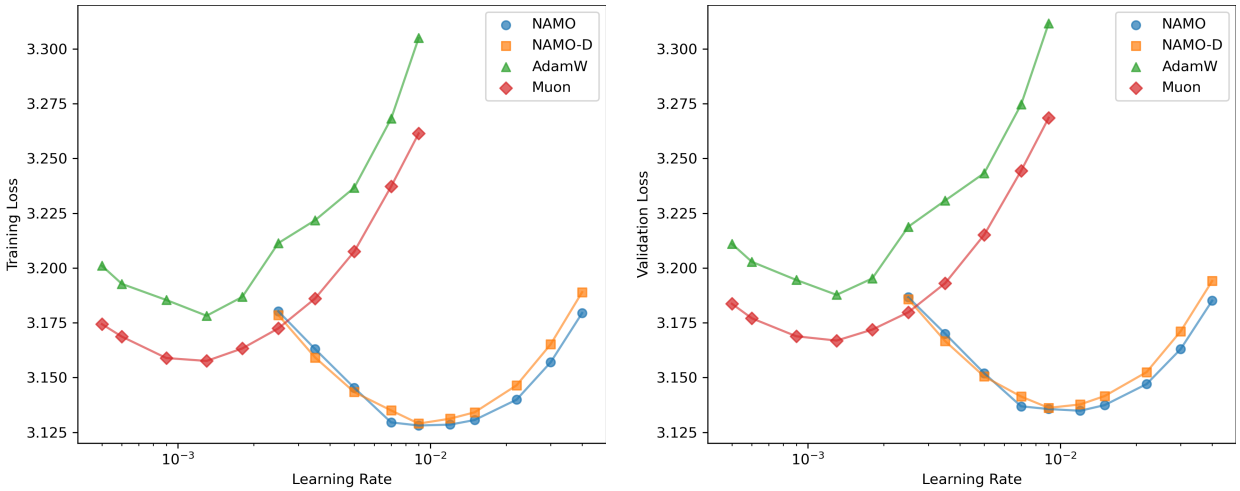
The optimal LR is the one that achieves the lowest validation loss after 10K steps. Additionally, for NAMO-D, we observe that its performance on the GPT-2 (355M) model can vary for different choices of the clamping hyperparameter c. Therefore, we also sweep for an optimal c from the following set:

$$\{0.12, 0.40, 0.75, 0.90\}.$$

The optimal hyperparameters for each optimizer for the GPT-2 (355M) model are listed in Table 1, and the training and validation losses at termination for each optimizer are reported in Table 2, together with those for the GPT-2 (124M) model. Figure 3 show that NAMO and NAMO-D outperforms both Muon and AdamW, with NAMO-D providing further gains over NAMO through the additional hyperparameter c. This observation is consistent with the algorithmic design of the two proposed algorithms. NAMO augments Muon with adaptivity while preserving the orthogonality of the update direction. On the other hand, NAMO-D enables neuron-wise adaptivity while no longer strictly preserving the orthogonalized direction, and a suitable choice of the clamping parameter c balances the two competing goals of maintaining the structural advantages of a well-conditioned update direction and benefiting from finer-grained noise adaptation.

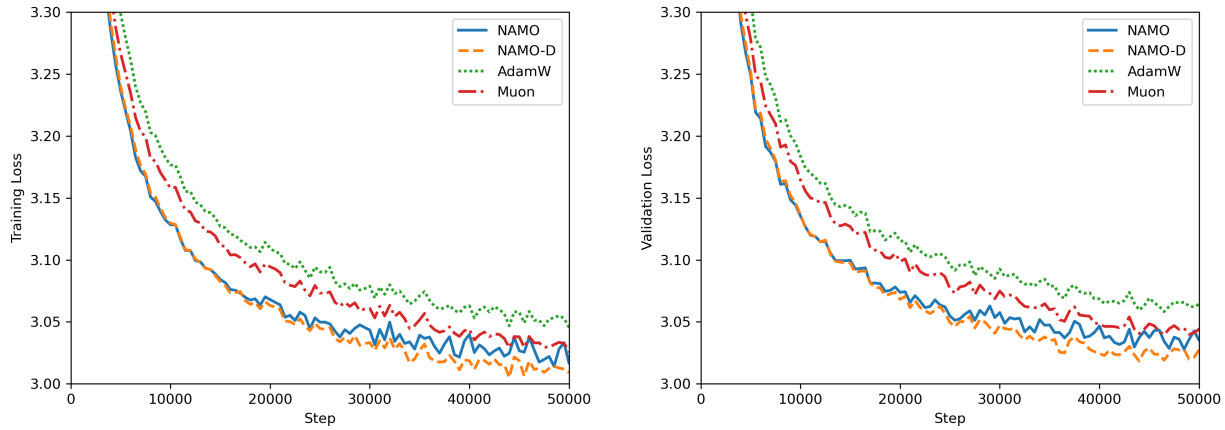
Table 1: **Optimal hyperparameters for all optimizers and pretraining of two model sizes.** Grid search is used to determine the optimal choices.

Optimizer	GPT-2 (124M) Hyperparameters	GPT-2 (355M) Hyperparameters
AdamW	$\eta = 0.0013$	$\eta = 0.0009$
Muon	$\eta = 0.0013$	$\eta = 0.0009$
NAMO	$\eta = 0.012$	$\eta = 0.007$
NAMO-D	$\eta = 0.009$, $c = 0.1$	$\eta = 0.009$, $c = 0.9$



(a) Training loss (b) Validation Loss

Figure 1: **Hyperparameter sweeping results for GPT-2 (124M).** The training and validation losses at step 10K are reported, where the x-axis is the learning rate.

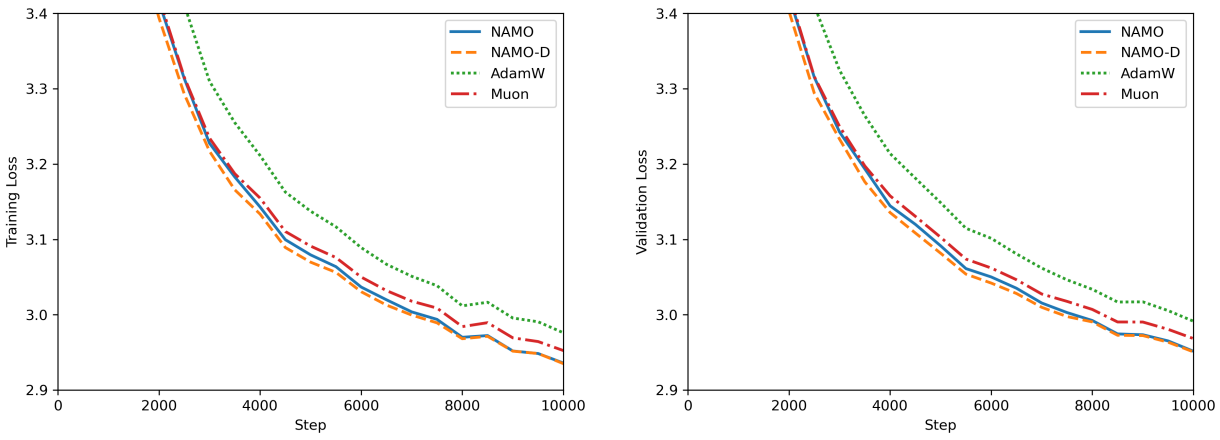


(a) Training loss (b) Validation Loss

Figure 2: **Pretraining GPT-2 (124M) for 50K steps.** The optimal LR from sweeping for 10K steps is used.

Table 2: **Final training and validation losses for GPT-2 (124M) and GPT-2 (355M).** GPT-2 (124M) is trained for 50K steps, and GPT-2 (355M) is trained for 10K steps.

Optimizer	GPT-2 (124M)		GPT-2 (355M)	
	Training Loss	Validation Loss	Training Loss	Validation Loss
AdamW	3.0456	3.0643	2.9760	2.9914
Muon	3.0265	3.0435	2.9524	2.9684
NAMO	2.9272	3.0351	2.9359	2.9516
NAMO-D	2.9167	3.0246	2.9351	2.9507



(a) Training loss

(b) Validation Loss

Figure 3: **Pretraining GPT-2 (355M) for 10K steps.** The optimal LR (and optimal c for NAMO-D) from sweeping for 10K steps are used.

5 Conclusions and Future Work

In this work, we propose a new optimizer and a diagonal extension, NAMO and NAMO-D, which provide the first theoretically principled integration of an orthogonalized update direction with norm-based adaptive moment estimation for noise adaptation. NAMO rescales Muon's orthogonalized momentum by a single adaptive scalar, thereby preserving the orthogonality of the update direction while yielding performance improvements over Muon at negligible additional computational cost. NAMO-D instead right-multiplies the orthogonalized momentum by a diagonal matrix, enabling finer-grained, neuron-wise noise adaptation for further improving performance, albeit without strictly preserving orthogonality of the update direction. Under standard smoothness and unbiased bounded-variance noise assumptions, we establish optimal convergence rates for both algorithms in the deterministic setting and show that, in the stochastic setting, their convergence guarantees adapt to the noise level of the stochastic gradients and attain the optimal rate when the batch size is sufficiently large. Experiments on pretraining GPT-2 (124M) and GPT-2 (355M) models demonstrate the improved performance of both NAMO and NAMO-D compared to the AdamW and Muon baselines. NAMO-D provides further gains over NAMO through an additional clamping hyperparameter, which can be tuned to balance two competing goals: maintaining the structural advantages of a well-conditioned update direction and leveraging finer-grained noise adaptation. Future work includes evaluating NAMO and NAMO-D on larger LLMs, developing tuning-light variants of NAMO-D, and further investigating theoretical and practical advantages of noise-adaptive scaling for orthogonalized updates.

Acknowledgement

Funding. This work was supported in part by NSF DMS 2502561.

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.

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

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413. PMLR, 2018.

Jeremy Bernstein. The modula docs, 2025. URL `https://docs.modula.systems/`.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Valentyn Boreiko, Zhiqi Bu, and Sheng Zha. Towards understanding of orthogonalization in muon. In *High-dimensional Learning Dynamics 2025*, 2025.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*, 2024.

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M Gower. An exploration of non-euclidean gradient descent: Muon and its many variants. *arXiv preprint arXiv:2510.09827*, 2025.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

Zhaorui Dong, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus, 2019.

Jie Hao, Xiaochuan Gong, Jie Xu, Zhengdao Wang, and Mingrui Liu. Noise-adaptive layerwise learning rates: Accelerating geometry-aware optimization for deep neural network training. *arXiv preprint arXiv:2510.14009*, 2025.

Wei He, Kai Han, Hang Zhou, Hanting Chen, Zhicheng Liu, Xinghao Chen, and Yunhe Wang. Root: Robust orthogonalized optimizer for neural network training. *arXiv preprint arXiv:2511.20626*, 2025.

Nicholas J Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Keller Jordan, Jeremy Bernstein, Ben Rappazzo, B Vlado, Y Jiacheng, F Cesista, and B Koszarsky. Modded-nanogpt: Speedrunning the nanogpt baseline. *GitHub repository*, 2024a.

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

Andrej Karpathy. Nanogpt. `https://github.com/karpathy/nanoGPT`, 2022.

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

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon. *arXiv preprint arXiv:2502.02900*, 2025.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 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*, 2025.

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

Antonio Orvieto and Robert Gower. In search of adam's secret sauce. *arXiv preprint arXiv:2505.21829*, 2025.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Analysis of muon's convergence and critical batch size. *arXiv preprint arXiv:2507.01598*, 2025.

Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J. Shah, et al. Practical efficiency of muon for pretraining, 2025.

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

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

Zitao Song, Cedar Site Bai, Zhe Zhang, Brian Bullins, and David F Gleich. Decoupling variance and scale-invariant updates in adaptive gradient descent for unified vector and matrix optimization. *arXiv preprint arXiv:2602.06880*, 2026.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Yujie Yang. Prism: Structured optimization via anisotropic spectral shaping. *arXiv preprint arXiv:2602.03096*, 2026.

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

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024.

Appendix A. Useful Lemmata

This section contains proofs of lemmata used in the analysis of NAMO and NAMO-D. The following lemma is used in the proofs of Theorems 3–6 for the derivation of a uniform upper bound of the adaptive stepsizes.

Lemma 7 *Assume* $\mu_1, \mu_2 \in (0,1]$ *with* $\mu_1 \le \mu_2$, $t > 0$, *and* $g_1, g_2, \cdots, g_t \in \mathbb{R}^d$. *Define* m_t *by* $m_0 = 0$ *and:*

$$m_\tau = \mu_1 m_{\tau-1} + (1-\mu_1) g_\tau,\ \tau = 1, 2, \cdots, t-1.$$

Define v_t *by* $v_0 = 0$ *and:*

$$v_\tau = \mu_2 m_{\tau-1} + (1-\mu_2) \|g_\tau\|^2,\ \tau = 1, 2, \cdots, t-1.$$

Let $\hat{m}_t := m_t/(1-\mu_1^t)$ *and* $\hat{v}_t := v_t/(1-\mu_2^t)$. *Then:*

$$\frac{\|\hat{m}_t\|}{\sqrt{\hat{v}_t}} \le \sqrt{\frac{1-\mu_1}{1-\mu_2}}.$$

Proof Write $w_{1,t,\tau} := \frac{1-\mu_1}{1-\mu_1^t}\mu_1^{t-\tau}$ and $w_{2,t,\tau} := \frac{1-\mu_2}{1-\mu_2^t}\mu_2^{t-\tau}$. Then:

$$\hat{m}_t = \frac{1-\mu_1}{1-\mu_1^t}\sum_{\tau=1}^t \mu_1^{t-\tau} g_\tau = \sum_{\tau=1}^t w_{1,t,\tau} g_\tau,$$

and

$$\hat{v}_t = \frac{1-\mu_2}{1-\mu_2^t}\sum_{\tau=1}^t \mu_2^{t-\tau} \|g_\tau\|^2 = \sum_{\tau=1}^t w_{2,t,\tau} \|g_\tau\|^2.$$

Since $\sum_{\tau=1}^t w_{1,t,\tau} = \sum_{\tau=1}^t w_{2,t,\tau} = 1$. Under the assumption that $\mu_1 \le \mu_2$,

$$w_{1,t,\tau} = \frac{1-\mu_1}{1-\mu_2}\frac{1-\mu_2^t}{1-\mu_1^t}\left(\frac{\mu_1}{\mu_2}\right)^{t-\tau} w_{2,t,\tau} \le \frac{1-\mu_1}{1-\mu_2} w_{2,t,\tau}.$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\hat{m}_t\|^2 =& \left\|\sum_{\tau=1}^t w_{1,t,\tau} g_\tau\right\|^2 \\
\le& \left(\sum_{\tau=1}^t w_{1,t,\tau}\right)\left(\sum_{\tau=1}^t w_{1,t,\tau}\|g_\tau\|^2\right) \\
\le& \frac{1-\mu_1}{1-\mu_2}\sum_{\tau=1}^t w_{2,t,\tau}\|g_\tau\|^2 \\
=& \frac{1-\mu_1}{1-\mu_2}\hat{v}_t.
\end{aligned}$$

Hence,

$$\alpha_t \le \frac{\|\hat{m}_t\|}{\sqrt{\hat{v}_t}} \le \sqrt{\frac{1-\mu_1}{1-\mu_2}}.$$

■

The lemma below is used in the analysis of NAMO and NAMO-D in the deterministic setting, i.e. in the proofs of Theorem 3 and Theorem 5.

Lemma 8 *For $\epsilon > 0$ and $x \geq 0$, define $\phi_\epsilon(x) := \frac{x^2}{x+\epsilon}$. Then:*

$$x \leq \phi_\epsilon(x) + \sqrt{\epsilon \phi_\epsilon(x)}.$$

Proof By the definition of $\phi_\epsilon(x)$,

$$x^2 - \phi_\epsilon(x)x - \epsilon\phi_\epsilon(x) \leq 0.$$

Solving for x gives:

$$x \leq \frac{\phi_\epsilon(x) + \sqrt{\phi_\epsilon(x)^2 + 4\epsilon\phi_\epsilon(x)}}{2} \leq \phi_\epsilon(x) + \sqrt{\epsilon\phi_\epsilon(x)}.$$

■

The following two lemmas are used in the proofs of Theorem 4 and Theorem 6.

Lemma 9 *For $\mu \in (0,1)$ and $T > 0$,*

$$\sum_{t=1}^{T} \frac{1}{1-\mu^t} \leq T + \frac{\mu}{1-\mu} - \frac{1}{\ln \mu} \ln\left(\frac{1-\mu^T}{1-\mu}\right).$$

Proof For $x \geq 1$, define:

$$f(x) := \frac{\mu^x}{1-\mu^x}.$$

Since:

$$f'(x) = \frac{\mu^x \ln \mu}{(1-\mu^x)^2} < 0,$$

the integral test gives:

$$\sum_{t=1}^{T} \frac{\mu^t}{1-\mu^t} = \sum_{t=1}^{T} f(t) \leq f(1) + \int_1^T f(x)dx.$$

To compute the integral, write $y := \mu^x$, then $dy = \mu^x \ln \mu dx = y \ln \mu dx$, and

$$\int_1^T \frac{\mu^x}{1-\mu^x} dx = \int_\mu^{\mu^T} \frac{y}{1-y} \frac{dy}{y \ln \mu} = -\frac{1}{\ln \mu} \int_\mu^{\mu^T} d\ln(1-y) = -\frac{1}{\ln \mu} \ln\left(\frac{1-\mu^T}{1-\mu}\right).$$

Hence,

$$\sum_{t=1}^{T} \frac{\mu^t}{1-\mu^t} \leq \frac{\mu}{1-\mu} - \frac{1}{\ln \mu} \ln\left(\frac{1-\mu^T}{1-\mu}\right).$$

It then follows that:

$$\sum_{t=1}^{T} \frac{1}{1-\mu^t} = \sum_{t=1}^{T} \left(1 + \frac{\mu^t}{1-\mu^t}\right) \leq T + \frac{\mu}{1-\mu} - \frac{1}{\ln \mu} \ln\left(\frac{1-\mu^T}{1-\mu}\right).$$

■

Lemma 10 *For $\mu \in (0,1)$ and $T > 0$,*

$$\sum_{t=1}^{T} \frac{1}{\sqrt{1-\mu^t}} \leq T - \frac{2\ln(1+\sqrt{1-\mu^T})}{\ln \mu}.$$

Proof For $x \geq 1$, define:

$$f(x) := \frac{1}{\sqrt{1-\mu^x}}.$$

Since:

$$f'(x) = \frac{\mu^x \ln \mu}{2(1-\mu^x)^{3/2}} < 0,$$

f is strictly decreasing on $[1, \infty)$. Hence,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{1-\mu^t}} \leq \int_0^T f(x)dx.$$

Let $y := \sqrt{1-\mu^x}$, then $dy = \frac{-\mu^x \ln \mu}{2\sqrt{1-\mu^x}}dx = \frac{(y^2-1)\ln \mu}{2y}dx$. It follows that:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{1-\mu^t}} \leq \int_0^T \frac{1}{\sqrt{1-\mu^x}}dx = \frac{-2}{\ln \mu}\int_0^{\sqrt{1-\mu^T}} \frac{1}{1-y^2}dy = \frac{1}{-\ln \mu}\ln\left(\frac{1+\sqrt{1-\mu^T}}{1-\sqrt{1-\mu^T}}\right).$$

Since:

$$\frac{1}{-\ln \mu}\ln\left(\frac{1+\sqrt{1-\mu^T}}{1-\sqrt{1-\mu^T}}\right) = \frac{1}{-\ln \mu}\ln\left(\frac{\left(1+\sqrt{1-\mu^T}\right)^2}{1-(1-\mu^T)}\right) = T - \frac{2\ln(1+\sqrt{1-\mu^T})}{\ln \mu},$$

Hence,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{1-\mu^t}} \leq T - \frac{2\ln(1+\sqrt{1-\mu^T})}{\ln \mu}.$$

■

The lemma below is used for the analysis for NAMO-D, i.e., in the proofs of Theorem 5 and Theorem 6.

Lemma 11 *Let $M \in \mathbb{R}^{m\times n}$ have reduced singular value decomposition (SVD) $M = U\Sigma V^T$, and let $O = UV^T$. Let $D = \operatorname{diag}(d_1, \cdots, dn)$ with $d_i \geq 0$ for all i, and write $d_{\min} := \min_i d_i$. Then*

$$\langle M, OD\rangle \geq d_{\min} \|M\|_* .$$

Proof Note that:

$$\langle M, OD\rangle = \operatorname{Tr}\left(M^T OD\right) = \operatorname{Tr}\left((U\Sigma V^T)^T UV^T D\right) = \operatorname{Tr}\left(V\Sigma V^T D\right) = \operatorname{Tr}\left(\Sigma V^T DV\right).$$

Since $D \succeq d_{\min} I_n$,

$$V^T(D - d_{\min} I_n)V \succeq 0.$$

It follows that:

$$\langle M, OD\rangle = d_{\min}\operatorname{Tr}(\Sigma) + \operatorname{Tr}\left(\Sigma V^T(D - d_{\min}I_n)V\right) \geq d_{\min}\|M\|_* .$$

■

Appendix B. Proof of Theorem 3

This section contains the detailed proof of Theorem 3 for the convergence of NAMO in the deterministic setting.

Proof

Step 1: A uniform upper bound on stepsize. For each $t \geq 0$ and $\tau \leq t$, define:

$$w_{1,t,\tau} := \frac{1-\mu_1}{1-\mu_1^t}\mu_1^{t-\tau} \quad \text{and} \quad w_{2,t,\tau} := \frac{1-\mu_2}{1-\mu_2^t}\mu_2^{t-\tau}.$$

It satisfies that $\sum_{\tau=1}^t w_{1,t,\tau} = \sum_{\tau=1}^t w_{2,t,\tau} = 1$. Let $\hat{M}_t := M_t/(1-\mu_1^t)$ and $\hat{v}_t := v_t/(1-\mu_2^t)$, then:

$$\hat{M}_t = \frac{1-\mu_1}{1-\mu_1^t}\sum_{\tau=1}^t \mu_1^{t-\tau}\nabla\mathcal{L}(\Theta_{\tau-1}) = \sum_{\tau=1}^t w_{1,t,\tau}\nabla\mathcal{L}(\Theta_{\tau-1}),$$

and

$$\hat{v}_t = \frac{1-\mu_2}{1-\mu_2^t}\sum_{\tau=1}^t \mu_2^{t-\tau}\|\nabla\mathcal{L}(\Theta_{\tau-1})\|^2 = \sum_{\tau=1}^t w_{2,t,\tau}\|\nabla\mathcal{L}(\Theta_{\tau-1})\|^2,$$

Define:

$$\alpha_t := \frac{\sqrt{1-\mu_2^t}}{1-\mu_1^t}\frac{\|M_t\|_F}{\sqrt{v_t}+\epsilon} = \frac{\left\|\hat{M}_t\right\|_F}{\sqrt{\hat{v}_t}+\epsilon/\sqrt{1-\mu_2^t}}.$$

Since the Frobenius norm of a matrix coincides with the Euclidean norm of its vectorization, by Lemma 7,

$$\alpha_t \leq \frac{\left\|\hat{M}_t\right\|_F}{\sqrt{\hat{v}_t}} = \sqrt{\frac{1-\mu_1}{1-\mu_2}}. \tag{4}$$

Step 2: A descent inequality and averaging. By (Zhang et al., 2025, Lemma B.1),

$$\begin{aligned}
&\mathcal{L}(\Theta_t) - \mathcal{L}(\Theta_{t-1}) \\
\leq& -\langle\nabla\mathcal{L}(\Theta_{t-1}), \eta\alpha_t O_t\rangle + \frac{L}{2}\eta^2\alpha_t^2 \\
\leq& -\eta\alpha_t\|\nabla\mathcal{L}(\Theta_{t-1})\|_* + 2\eta\alpha_t\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{L}{2}\eta^2\alpha_t^2 \\
=& -\eta\frac{\left\|\hat{M}_t\right\|\|\nabla\mathcal{L}(\Theta_{t-1})\|_*}{\sqrt{\hat{v}_t}+\epsilon/\sqrt{1-\mu_2}} + 2\eta\alpha_t\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{L}{2}\eta^2\alpha_t^2.
\end{aligned}$$

Rearranging the terms gives:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\frac{\left\|\hat{M}_t\right\|\|\nabla\mathcal{L}(\Theta_{t-1})\|_*}{\sqrt{\hat{v}_t}+\epsilon/\sqrt{1-\mu_2}} \leq & \frac{\Delta}{\eta T}+\frac{2}{T}\sum_{t=1}^{T}\alpha_t\left\|\nabla\mathcal{L}(\Theta_{t-1})-\hat{M}_t\right\|_*+\frac{\eta L}{2T}\sum_{t=1}^{T}\alpha_t^2 \\ \leq & \frac{\Delta}{\eta T}+\frac{2}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\left\|\nabla\mathcal{L}(\Theta_{t-1})-\hat{M}_t\right\|_*+\frac{\eta L}{2}\left(\frac{1-\mu_1}{1-\mu_2}\right), \end{aligned} \tag{5}$$

where $\Delta := \mathcal{L}(\Theta_0) - \min_\Theta \mathcal{L}(\Theta)$.

Step 3: Bounding the distance between bias-corrected momentum and true gradient. The difference between the scaled momentum $\hat{M}_t$ and the gradient $\nabla\mathcal{L}(\Theta_{t-1})$ can be bounded by:

$$\begin{aligned}\left\|\hat{M}_t-\nabla\mathcal{L}(\Theta_{t-1})\right\|_* \leq & \sum_{\tau=1}^{t} w_{1,t,\tau}\left\|\nabla\mathcal{L}(\Theta_{\tau-1})-\nabla\mathcal{L}(\Theta_{t-1})\right\|_* \\ \leq & \sum_{\tau=1}^{t} w_{1,t,\tau}L\left\|\Theta_{\tau-1}-\Theta_{t-1}\right\|_2 \\ \leq & \sum_{\tau=1}^{t} w_{1,t,\tau}L\left\|\sum_{s=\tau}^{t-1}\eta\alpha_s O_s\right\|_2 \\ \leq & \sum_{\tau=1}^{t} w_{1,t,\tau}\eta L\left(\sum_{s=\tau}^{t-1}\alpha_s\right). \end{aligned}$$

Then by (4) and the definition of $w_{1,t,\tau}$,

$$\begin{aligned}\left\|\hat{M}_t-\nabla\mathcal{L}(\Theta_{t-1})\right\|_* \leq & \sum_{\tau=1}^{t} w_{1,t,\tau}\eta L\,(t-\tau)\sqrt{\frac{1-\mu_1}{1-\mu_2}} \\ = & \sum_{\tau=1}^{t}\eta L\,(t-\tau)\,\mu_1^{t-\tau}\frac{1-\mu_1}{1-\mu_1^t}\sqrt{\frac{1-\mu_1}{1-\mu_2}} \\ \leq & \eta L\frac{1-\mu_1}{1-\mu_1^t}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{j=1}^{\infty} j\mu_1^j \\ \leq & \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2}. \end{aligned} \tag{6}$$

Step 4: Bounding $\sqrt{\hat{v}_t}$. For $\hat{v}_t$, it satisfies that:

$$\begin{aligned}\hat{v}_t = & \sum_{\tau=1}^{t} w_{2,t,\tau}\left\|\nabla\mathcal{L}(\Theta_{\tau-1})\right\|_F^2 \\ \leq & \sum_{\tau=1}^{t} w_{2,t,\tau}\left(\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + L\left\|\Theta_{\tau-1}-\Theta_{t-1}\right\|_2\right)^2 \\ \leq & \sum_{\tau=1}^{t} w_{2,t,\tau}\left(\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \eta L\left(\sum_{s=\tau}^{t-1}\alpha_s\right)\right)^2 \end{aligned}$$

Then by (4) and the definition of $w_{2,t,\tau}$,

$$\begin{aligned}\hat{v}_t \leq & \sum_{\tau=1}^{t} w_{2,t,\tau} \left(\left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_* + \eta L \left(t-\tau\right) \sqrt{\frac{1-\mu_1}{1-\mu_2}} \right)^2 \\ = & \left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_*^2 + 2\eta L \left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_* \sqrt{\frac{1-\mu_1}{1-\mu_2}} \sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau) \\ & + \eta^2 L^2 \left(\frac{1-\mu_1}{1-\mu_2} \right) \sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau)^2.\end{aligned}$$

By Cauchy-Schwarz inequality, it follows that:

$$\begin{aligned}\hat{v}_t \leq & \left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_*^2 + 2\eta L \left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_* \sqrt{\frac{1-\mu_1}{1-\mu_2}} \sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau)^2} \\ & + \eta^2 L^2 \left(\frac{1-\mu_1}{1-\mu_2} \right) \left(\sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau)^2 \right) \\ \leq & \left(\left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_* + \eta L \sqrt{\frac{1-\mu_1}{1-\mu_2}} \sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau)^2} \right)^2 \\ = & \left(\left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_* + a_t \right)^2,\end{aligned}$$

where

$$a_t := \eta L \sqrt{\frac{1-\mu_1}{1-\mu_2}} \sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau)^2}.$$

Since

$$\begin{aligned}a_t^2 \leq & \eta^2 L^2 \left(\frac{1-\mu_1}{1-\mu_2} \right) \left(\frac{1-\mu_2}{1-\mu_2^t} \right) \left(\sum_{\tau=1}^{t} \mu_2^{t-\tau} (t-\tau)^2 \right) \\ \leq & \eta^2 L^2 \left(\frac{1-\mu_1}{1-\mu_2} \right) \left(\sum_{\tau=1}^{\infty} \mu_2^{\tau} \tau^2 \right) \\ \leq & \eta^2 L^2 \left(\frac{1-\mu_1}{1-\mu_2} \right) \left(\frac{\mu_2(1+\mu_2)}{(1-\mu_2)^3} \right) := a^2. \end{aligned} \tag{7}$$

Hence,

$$\sqrt{\hat{v}_t} \leq \left\| \nabla \mathcal{L}(\Theta_{t-1}) \right\|_* + a. \tag{8}$$

Step 5: Relating $\alpha_t\|\nabla\mathcal{L}(\Theta_{t-1})\|_*$ to $\|\nabla\mathcal{L}(\Theta_{t-1})\|_*$. Combining (6) and (8) gives:

$$
\begin{aligned}
\alpha_t \left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* =& \frac{\left\|\hat{M}_t\right\|_F \left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\sqrt{\hat{v}_t}+\epsilon/\sqrt{1-\mu_2^t}} \\
\geq& \frac{\left(\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F - \left\|\hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right)\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \epsilon/\sqrt{1-\mu_2^t}+a} \\
\geq& \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F \left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \epsilon/\sqrt{1-\mu_2^t}+a} - \left\|\hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})\right\|_F \\
\geq& \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F \left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \epsilon/(1-\mu_2^t)}\left(1-\frac{a}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \epsilon/(1-\mu_2^t)+a}\right) - \left\|\hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})\right\|_F \\
\geq& \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F \left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \epsilon/(1-\mu_2^t)} - a - \left\|\hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})\right\|_F \\
\geq& \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F \left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \epsilon/(1-\mu_2^t)} - a - \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2} \\
\geq& \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*^2/\sqrt{r}}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* + \tilde{\epsilon}} - a - \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2}
\end{aligned}
$$

where $\tilde{\epsilon} := \epsilon/(1-\mu_2)$, $r := \min\{m,n\}$ and the constant a is given in (7). Define $\phi_{\tilde{\epsilon}}(x) := \frac{x^2}{x+\tilde{\epsilon}}$ for $x \geq 0$, it then follows from (5) and (6) that:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\phi_{\tilde{\epsilon}}(\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*) \\
\leq&\sqrt{r}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\left\|\hat{M}_t\right\|\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\sqrt{\hat{v}_t}+\epsilon/(1-\mu_2^t)} + a + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2}\right) \\
\leq&\sqrt{r}\left(\frac{\Delta}{\eta T} + \frac{2}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{\eta L}{2}\left(\frac{1-\mu_1}{1-\mu_2}\right) + a + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2}\right) \\
\leq&\sqrt{r}\left(\frac{\Delta}{\eta T} + \frac{2}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{\eta L}{2}\left(\frac{1-\mu_1}{1-\mu_2}\right) + a + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2}\right) \\
\leq&\sqrt{r}\left(\frac{\Delta}{\eta T} + \left(\frac{1-\mu_1}{1-\mu_2}\right)\frac{2\eta L\mu_1}{(1-\mu_1)^2} + \frac{\eta L}{2}\left(\frac{1-\mu_1}{1-\mu_2}\right) + \eta L\frac{\sqrt{\mu_2(1-\mu_1)(1+\mu_2)}}{(1-\mu_2)^2} + \sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\eta L\mu_1}{(1-\mu_1)^2}\right) \\
=&\frac{\Delta\sqrt{r}}{\eta T} + \eta L C_\mu\sqrt{r},
\end{aligned}
$$

where

$$
C_\mu := \left(\frac{1-\mu_1}{1-\mu_2}\right)\frac{2\mu_1}{(1-\mu_1)^2} + \frac{1}{2}\left(\frac{1-\mu_1}{1-\mu_2}\right) + \frac{\sqrt{\mu_2(1-\mu_1)(1+\mu_2)}}{(1-\mu_2)^2} + \sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2}
$$

is a constant that depends on μ_1 and μ_2.

Step 6: Deriving the convergence rate. By Lemma 8,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\Theta_{t-1}\|_* \leq & \frac{1}{T}\sum_{t=1}^{T}\phi_{\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_*\right)+\frac{1}{T}\sum_{t=1}^{T}\sqrt{\tilde{\epsilon}\phi_{\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_*\right)} \\ \leq & \frac{1}{T}\sum_{t=1}^{T}\phi_{\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_*\right)+\frac{\tilde{\epsilon}^{\frac{1}{2}}}{\sqrt{T}}\sqrt{\sum_{t=1}^{T}\phi_{\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_*\right)} \\ \leq & \frac{\Delta\sqrt{r}}{\eta T}+\eta LC_\mu\sqrt{r}+\tilde{\epsilon}^{\frac{1}{2}}\sqrt{\frac{\Delta\sqrt{r}}{\eta T}+\eta LC_\mu\sqrt{r}}.\end{aligned}$$

In particular, if choosing $\eta=\mathcal{O}\left(T^{-\frac{1}{2}}\right)$, $\mu_1=\Theta(1)$, $\mu_2=\Theta(1)$, and $\epsilon=\mathcal{O}(T^{-\frac{1}{2}})$, then:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\Theta_{t-1})\|_F \leq \frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\Theta_{t-1})\|_* \leq \mathcal{O}\left(T^{-\frac{1}{2}}\right)$$

for large $T>0$. The proof is thus completed. ■

Appendix C. Proof of Theorem 4

This section contains the detailed proof of Theorem 4 for the convergence of NAMO in the stochastic setting.

Proof

Step 1: A uniform upper bound on stepsize For each $t\geq 0$ and $\tau\leq t$, define:

$$w_{1,t,\tau}:=\frac{1-\mu_1}{1-\mu_1^t}\mu_1^{t-\tau} \quad \text{and} \quad w_{2,t,\tau}:=\frac{1-\mu_2}{1-\mu_2^t}\mu_2^{t-\tau}.$$

It satisfies that $\sum_{\tau=1}^{t}w_{1,t,\tau}=\sum_{\tau=1}^{t}w_{2,t,\tau}=1$. Let $\hat{M}_t:=M_t/(1-\mu_1^t)$ and $\hat{v}_t:=v_t/(1-\mu_2^t)$, then:

$$\hat{M}_t=\frac{1-\mu_1}{1-\mu_1^t}\sum_{\tau=1}^{t}\mu_1^{t-\tau}G_\tau=\sum_{\tau=1}^{t}w_{1,t,\tau}G_\tau,$$

and

$$\hat{v}_t=\frac{1-\mu_2}{1-\mu_2^t}\sum_{\tau=1}^{t}\mu_2^{t-\tau}\|G_\tau\|_F^2=\sum_{\tau=1}^{t}w_{2,t,\tau}\|G_\tau\|_F^2,$$

Define:

$$\alpha_t:=\frac{\sqrt{1-\mu_2^t}}{1-\mu_1^t}\frac{\|M_t\|_F}{\sqrt{v_t}+\epsilon}=\frac{\left\|\hat{M}_t\right\|_F}{\sqrt{\hat{v}_t}+\epsilon/\sqrt{1-\mu_2^t}}.$$

Since the Frobenius norm of a matrix coincides with the Euclidean norm of its vectorization, by Lemma 7,

$$\alpha_t\leq\frac{\left\|\hat{M}_t\right\|_F}{\sqrt{\hat{v}_t}}=\sqrt{\frac{1-\mu_1}{1-\mu_2}}. \tag{9}$$

Step 2: Expected descent and an average bound on $\alpha_t\|\hat{M}_t\|_*$. Let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|\Theta_{t-1}]$ denote the conditional expectation given the previous iterates $\Theta_0, \cdots, \Theta_{t-1}$, and write $E_t := \hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})$. Then by (Zhang et al., 2025, Lemma B.1),

$$\begin{aligned}
&\mathbb{E}_t\left[\mathcal{L}(\Theta_t) - \mathcal{L}(\Theta_{t-1})\right] \\
\leq&\mathbb{E}_t\left[-\left\langle\nabla\mathcal{L}(\Theta_{t-1}), \eta\alpha_t O_t\right\rangle\right] + \frac{\eta^2 L}{2}\mathbb{E}_t\left[\alpha_t^2\right] \\
=&\mathbb{E}_t\left[-\left\langle\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t, \eta\alpha_t O_t\right\rangle\right] - \mathbb{E}_t\left[\eta\alpha_t\left\|\hat{M}_t\right\|_*\right] + \frac{\eta^2 L}{2}\mathbb{E}_t\left[\alpha_t^2\right] \\
\leq&\left(\mathbb{E}_t\left[\eta\alpha_t\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_*\right] - \mathbb{E}_t\left[\eta\alpha_t\left\|\hat{M}_t\right\|_*\right]\right) + \frac{\eta^2 L}{2}\mathbb{E}_t\left[\alpha_t^2\right] \\
\leq&-\mathbb{E}_t\left[\eta\alpha_t\left\|\hat{M}_t\right\|_*\right] + \mathbb{E}_t\left[\eta\alpha_t\left\|E_t\right\|_*\right] + \frac{\eta^2 L}{2}\mathbb{E}_t\left[\alpha_t^2\right].
\end{aligned}$$

Rearranging the terms gives:

$$\mathbb{E}_t\left[\alpha_t\left\|\hat{M}_t\right\|_*\right] \leq \mathbb{E}_t\left[\mathcal{L}(\Theta_{t-1}) - \mathcal{L}(\Theta_t)\right] + \mathbb{E}_t\left[\alpha_t\left\|E_t\right\|_*\right] + \frac{\eta L}{2}\mathbb{E}_t\left[\alpha_t^2\right].$$

Then by the law of total expectation and (9),

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_*\right] \leq&\frac{\Delta}{\eta T} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t\left\|E_t\right\|_*\right] + \frac{\eta L}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2\right] \\
\leq&\frac{\Delta}{\eta T} + \frac{1}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\left\|E_t\right\|_*\right] + \frac{\eta L(1-\mu_1)}{2(1-\mu_2)}. \qquad (10)
\end{aligned}$$

Step 3: Bounding the distance between bias-corrected momentum and true gradient. For each t, it satisfies:

$$\begin{aligned}
E_t =&\hat{M}_t - \mathbb{E}\left[\hat{M}_t\right] + \mathbb{E}\left[\hat{M}_t\right] - \nabla\mathcal{L}(\Theta_{t-1}) \\
=&\sum_{\tau=1}^{t} w_{1,t,\tau}\left(G_\tau - \nabla\mathcal{L}(\Theta_{\tau-1})\right) + \sum_{\tau=1}^{t} w_{1,t,\tau}\left(\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right)
\end{aligned}$$

Hence, by (9),

$$\begin{aligned}
\mathbb{E}\left[\|E_t\|_F^2\right] \leq & \mathbb{E}\left[\left\|\sum_{\tau=1}^{t} w_{1,t,\tau}\left(G_\tau - \nabla\mathcal{L}(\Theta_{\tau-1})\right)\right\|_F^2\right] + \left\|\sum_{\tau=1}^{t} w_{1,t,\tau}\left(\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right)\right\|_F^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right)^2 \left(\sum_{\tau=1}^{t} \mu_1^{2(t-\tau)}\right) \frac{\sigma^2}{b} + \sum_{\tau=1}^{t} w_{1,t,\tau} \left\|\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right\|_*^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right)^2 \left(\frac{1-\mu_1^{2t}}{1-\mu_1^2}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \sum_{\tau=1}^{t} \mu_1^{t-\tau} L^2 \left\|\Theta_{\tau-1} - \Theta_{t-1}\right\|_2^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_1^t}\right) L^2\eta^2 \sum_{\tau=1}^{t} \mu_1^{t-\tau} \left(\sum_{s=\tau}^{t-1} \alpha_s\right)^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left[\left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_2}\right) L^2\eta^2 \sum_{\tau=1}^{t} \mu_1^{t-\tau} (t-\tau)^2\right] \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left[\left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_2}\right) L^2\eta^2 \sum_{\tau=1}^{\infty} \mu_1^{\tau} \tau^2\right] \\
= & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left[\left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{\mu_1(1+\mu_1)}{(1-\mu_2)(1-\mu_1)^2}\right) L^2\eta^2\right] \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \frac{\sigma^2}{b} + \frac{\mu_1(1+\mu_1)L^2\eta^2}{(1-\mu_2)(1-\mu_1)(1-\mu_1^t)}.
\end{aligned}$$

Then by Lemma 9, it follows that:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|E_t\|_F^2\right] \leq & \left(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\mu_1^t}\right) \left(\frac{\sigma^2(1-\mu_1)}{b} + \frac{\mu_1(1+\mu_1)L^2\eta^2}{(1-\mu_1)(1-\mu_2)}\right) \\
\leq & \left(1 + \frac{\mu_1}{(1-\mu_1)T} - \frac{1}{T\ln\mu_1} \ln\left(\frac{1-\mu_1^T}{1-\mu_1}\right)\right) \left(\frac{\sigma^2(1-\mu_1)}{b} + \frac{\mu_1(1+\mu_1)L^2\eta^2}{(1-\mu_1)(1-\mu_2)}\right).
\end{aligned}$$

By Cauchy-Schwarz inequality and Jensen's inequality,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|E_t\|_*\right] \leq & \sqrt{\frac{r}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|E_t\|_F^2\right]} \\
\leq & \sqrt{1 + \frac{\mu_1}{(1-\mu_1)T} - \frac{1}{T\ln\mu_1} \ln\left(\frac{1-\mu_1^T}{1-\mu_1}\right)} \left(\frac{\sigma\sqrt{r(1-\mu_1)}}{\sqrt{b}} + L\eta\sqrt{\frac{r\mu_1(1+\mu_1)}{(1-\mu_2)(1-\mu_1)}}\right).
\end{aligned} \tag{11}$$

Step 4: Bounding $\mathbb{E}[\sqrt{\hat{v}_t}]$. For $\hat{v}_t$, by Minkowski inequality and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\hat{v}_t}\right] =& \mathbb{E}\left[\sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau} \left\|G_\tau\right\|_F^2}\right] \\
\leq& \mathbb{E}\left[\sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau} \left\|\nabla\mathcal{L}(\Theta_{\tau-1}) - G_\tau\right\|_F^2}\right] + \mathbb{E}\left[\sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau} \left\|\nabla\mathcal{L}(\Theta_{\tau-1})\right\|_F^2}\right] \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \mathbb{E}\left[\sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau} \left\|\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right\|_F^2}\right] \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\mathbb{E}\left[\sqrt{\sum_{\tau=1}^{t} w_{2,t,\tau} \left(\sum_{\tau=1}^{t} \alpha_s\right)^2}\right] \\
\overset{(9)}{\leq}& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2^t}}\sqrt{\sum_{\tau=1}^{t} \mu_2^{t-\tau}\left(t-\tau\right)^2} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2^t}}\sqrt{\sum_{\tau=1}^{\infty} \mu_2^{t-\tau}\left(t-\tau\right)^2} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2^t}}\sqrt{\sum_{\tau=1}^{\infty} \mu_2^{\tau}\tau^2} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{(1-\mu_1)\mu_2(1+\mu_2)}{(1-\mu_2^t)(1-\mu_2)^3}} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2^t)(1-\mu_2)^3}}.
\end{aligned}$$

Hence,

$$\mathbb{E}\left[\sqrt{\hat{v}_t} + \frac{\epsilon}{\sqrt{1-\mu_2^t}}\right] \leq \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|\right] + a_t, \tag{12}$$

where

$$a_t := \frac{\sigma}{\sqrt{b}} + \sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2^t)(1-\mu_2)^3}} + \frac{\epsilon}{\sqrt{1-\mu_2^t}}.$$

Then by Lemma 10,

$$\frac{1}{T}\sum_{t=1}^{T} a_t \leq \frac{\sigma}{T}\sum_{t=1}^{T}\frac{1}{\sqrt{b}} + \left(\sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2)^3}} + \epsilon\right)\left(1 - \frac{2\ln(1+\sqrt{1-\mu_2^T})}{T\ln\mu_2}\right). \tag{13}$$

Step 5: Relating $\mathbb{E}\left[\alpha_t \left\|\hat{M}_t\right\|_F\right]$ to $\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]$. By Cauchy-Schwarz inequality,

$$\begin{aligned}\left(\mathbb{E}\left[\left\|\hat{M}_t\right\|_F\right]\right)^2 =& \left(\mathbb{E}\left[\frac{\left\|\hat{M}_t\right\|_F}{\left(\sqrt{\hat{v}_t}+\epsilon_t\right)^{\frac{1}{2}}}\cdot\left(\sqrt{\hat{v}_t}+\epsilon_t\right)^{\frac{1}{2}}\right]\right)^2 \\ \leq& \mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]\mathbb{E}\left[\sqrt{\hat{v}_t}+\epsilon_t\right],\end{aligned}$$

where $\epsilon_t := \epsilon/\sqrt{1-\mu_2^t}$. Combining the above with (12) gives:

$$\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right] \geq \frac{\left(\mathbb{E}\left[\left\|\hat{M}_t\right\|_F\right]\right)^2}{\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]+a_t} \geq \frac{\left(\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]-\mathbb{E}\left[\|E_t\|_F\right]\right)^2}{\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]+a_t}.$$

Rearranging the terms gives:

$$\left(\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]\right)^2 - \left(2\mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]\right)\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right] - a_t\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right] + \mathbb{E}\left[\|E_t\|_F\right]^2 \leq 0.$$

Then solving for $\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]$ gives:

$$\begin{aligned}&\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right] \\ \leq& \frac{2\mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]+\sqrt{\left(2\mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]\right)^2+4a_t\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]-4\mathbb{E}\left[\|E_t\|_F\right]^2}}{2} \\ \leq& \mathbb{E}\left[\|E_t\|_F\right]+\frac{1}{2}\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]+\sqrt{\frac{1}{4}\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]^2+\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]} \\ \leq& \mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]+\sqrt{\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]}\end{aligned}$$

Step 6: Deriving the convergence rate. By Cauchy-Schwarz inequality,

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right] \\ \leq& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_F\right]+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]+\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t\left\|\hat{M}_t\right\|_F\right]}.\end{aligned}$$

Combining the above with (10) and (13) gives:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]\\
\leq&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\Delta}{\eta T}+\frac{\eta L(1-\mu_1)}{2(1-\mu_2)}+\frac{1}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]\\
&+\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)}\cdot\sqrt{\frac{\Delta}{\eta T}+\frac{1}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\eta L(1-\mu_1)}{2(1-\mu_2)}}\\
\leq&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\Delta}{\eta T}+\frac{\eta L(1-\mu_1)}{2(1-\mu_2)}+\frac{1}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]\\
&+\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_F\right]+\frac{\sigma}{\sqrt{b}}+\left(\sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2)^3}}+\epsilon\right)\left(1-\frac{2\ln(1+\sqrt{1-\mu_2^T})}{T\ln\mu_2}\right)}\\
&\cdot\sqrt{\frac{\Delta}{\eta T}+\frac{1}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\eta L(1-\mu_1)}{2(1-\mu_2)}}.
\end{aligned}$$

In particular, for large $T>0$, if choosing $\eta=\mathcal{O}(T^{-\frac{3}{4}})$, $1-\mu_1=\Theta(T^{-\frac{1}{2}})$, $1-\mu_2=\Theta(T^{-\frac{1}{2}})$, $0\leq\mu_1\leq\mu_2<1$, and $\epsilon=\mathcal{O}(T^{-\frac{1}{2}})$, then, by (11),

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]\leq\mathcal{O}\left(T^{-\frac{1}{4}}\right),$$

and it follows that:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]\leq\mathcal{O}\left(T^{-\frac{1}{4}}+\sqrt{\sigma}b^{-\frac{1}{4}}T^{-\frac{1}{8}}\right),$$

where b is the batch size. The proof is thus completed. ■

Appendix D. Proof of Theorem 5

This section contains the detailed proof of Theorem 5 for the convergence of NAMO-D in the deterministic setting.

Proof

Step 1: A uniform upper bound on stepsize. For each $t\geq 0$ and $\tau\leq t$, define:

$$w_{1,t,\tau}:=\frac{1-\mu_1}{1-\mu_1^t}\mu_1^{t-\tau}\quad\text{and}\quad w_{2,t,\tau}:=\frac{1-\mu_2}{1-\mu_2^t}\mu_2^{t-\tau}.$$

It satisfies that $\sum_{\tau=1}^{t}w_{1,t,\tau}=\sum_{\tau=1}^{t}w_{2,t,\tau}=1$. For each $j=1,\cdots,n$ and $t>0$, let M_t^j denote the j-th column of M_t, and $\mathbf{v}_t^j$ denote the j-th element of $\mathbf{v}_t$. Write $\hat{M}_t:=\frac{1}{1-\mu_1^t}M_t$, $\hat{\mathbf{v}}_t:=\frac{1}{1-\mu_2^t}\mathbf{v}_t$, and

$$D_t:=\operatorname{diag}\left(\min\left\{\max\left\{\mathbf{d}_t,c\bar{d}_t\mathbf{1}\right\},\frac{1}{c}\bar{d}_t\mathbf{1}\right\}\right),$$

where $\mathbf{d}_t$ and $\bar{d}_t$ are given in Algorithm 2. Then:

$$\hat{M}_t = \frac{1-\mu_1}{1-\mu_1^t}\sum_{\tau=1}^{t}\mu_1^{t-\tau}\nabla\mathcal{L}(\Theta_{\tau-1}) = \sum_{\tau=1}^{t} w_{1,t,\tau}\nabla\mathcal{L}(\Theta_{\tau-1}),$$

and

$$\hat{\mathbf{v}}_t = \frac{1-\mu_2}{1-\mu_2^t}\sum_{\tau=1}^{t}\mu_2^{t-\tau}\mathcal{N}_c(\nabla\mathcal{L}(\Theta_{\tau-1}))\odot\mathcal{N}_c(\nabla\mathcal{L}(\Theta_{\tau-1})) = \sum_{\tau=1}^{t} w_{2,t,\tau}\mathcal{N}_c(\nabla\mathcal{L}(\Theta_{\tau-1}))\odot\mathcal{N}_c(\nabla\mathcal{L}(\Theta_{\tau-1})).$$

By Lemma 7,

$$[D_t]_{jj} \le \frac{\left\|\hat{M}_t^j\right\|}{\sqrt{\hat{\mathbf{v}}_t^j}} \le \sqrt{\frac{1-\mu_1}{1-\mu_2}}, \quad \forall j. \tag{14}$$

Write:

$$d_{t,\max} := \max_j [D_t]_{jj}, \quad \text{and } d_{t,\min} := \min_j [D_t]_{jj}. \tag{15}$$

Then the condition number of D_t is given by:

$$\kappa(D_t) = \frac{d_{t,\max}}{d_{t,\min}} \le \min\left\{\kappa_t, \frac{1}{c^2}\right\}, \tag{16}$$

where $c \in (0,1]$ is the fixed clamping hyperparameter, $\kappa_t := \kappa(\operatorname{diag}(\mathbf{d}_t))$ denotes the condition number of $\operatorname{diag}(\mathbf{d}_t)$.

Step 2: Descent inequality and averaging. By (Zhang et al., 2025, Lemma B.1),

$$\begin{aligned}
&\mathcal{L}(\Theta_t) - \mathcal{L}(\Theta_{t-1}) \\
\le& -\langle\nabla\mathcal{L}(\Theta_{t-1}), \eta O_t D_t\rangle + \frac{L}{2}\eta^2\|D_t\|_2^2 \\
\le& -\eta d_{t,\min}\|\nabla\mathcal{L}(\Theta_{t-1})\|_* + 2\eta d_{t,\max}\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{L}{2}\eta^2\|D_t\|_2^2 \\
=& -\eta\max\{\kappa_t^{-1}, c^2\} d_{t,\max}\|\nabla\mathcal{L}(\Theta_{t-1})\|_* + 2\eta d_{t,\max}\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{L}{2}\eta^2\|D_t\|_2^2.
\end{aligned}$$

Rearranging the terms gives:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} d_{t,\max}\|\nabla\mathcal{L}(\Theta_{t-1})\|_* \le& \frac{\Delta}{\eta T c^2} + \frac{2}{Tc^2}\sum_{t=1}^{T} d_{t,\max}\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{\eta L}{2Tc^2}\sum_{t=1}^{T} d_{t,\max}^2 \\
\le& \frac{\Delta}{\eta T c^2} + \frac{2}{Tc^2}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_* + \frac{\eta L}{2c^2}\left(\frac{1-\mu_1}{1-\mu_2}\right).
\end{aligned} \tag{17}$$

Step 3: Bounding $\|\hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})\|_*$. The difference between the scaled momentum $\hat{M}_t$ and the gradient $\nabla\mathcal{L}(\Theta_{t-1})$ can be bounded by:

$$\begin{aligned}\left\|\hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})\right\|_* \leq& \sum_{\tau=1}^{t} w_{1,t,\tau} \left\|\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right\|_* \\ \leq& \sum_{\tau=1}^{t} w_{1,t,\tau} L \left\|\Theta_{\tau-1} - \Theta_{t-1}\right\|_2 \\ \leq& \sum_{\tau=1}^{t} w_{1,t,\tau} L \left\|\sum_{s=\tau}^{t-1} \eta O_s D_s\right\|_2 \\ \leq& \sum_{\tau=1}^{t} w_{1,t,\tau} \eta L \left(\sum_{s=\tau}^{t-1} d_{s,\max}\right).\end{aligned}$$

Then by (4) and the definition of $w_{1,t,\tau}$,

$$\begin{aligned}\left\|\hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})\right\|_* \leq& \sum_{\tau=1}^{t} w_{1,t,\tau} \eta L (t-\tau) \sqrt{\frac{1-\mu_1}{1-\mu_2}} \\ =& \sum_{\tau=1}^{t} \eta L (t-\tau) \mu_1^{t-\tau} \frac{1-\mu_1}{1-\mu_1^t} \sqrt{\frac{1-\mu_1}{1-\mu_2}} \\ \leq& \eta L \frac{1-\mu_1}{1-\mu_1^t} \sqrt{\frac{1-\mu_1}{1-\mu_2}} \sum_{j=1}^{\infty} j\mu_1^j \\ \leq& \eta L \sqrt{\frac{1-\mu_1}{1-\mu_2}} \frac{\mu_1}{(1-\mu_1)^2}.\end{aligned} \tag{18}$$

Step 4: Bounding $\sqrt{\sum_{j=1}^{n} \hat{\mathbf{v}}_t^j}$. For $\hat{\mathbf{v}}_t$, it satisfies that:

$$\begin{aligned}\sum_{j=1}^{n} \hat{\mathbf{v}}_t^j =& \sum_{\tau=1}^{t} w_{2,t,\tau} \left\|\nabla\mathcal{L}(\Theta_{\tau-1})\right\|_F^2 \\ \leq& \sum_{\tau=1}^{t} w_{2,t,\tau} \left(\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F + L \left\|\Theta_{\tau-1} - \Theta_{t-1}\right\|_2\right)^2 \\ \leq& \sum_{\tau=1}^{t} w_{2,t,\tau} \left(\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F + \eta L \left(\sum_{s=\tau}^{t-1} d_{s,\max}\right)\right)^2\end{aligned}$$

Then by (4) and the definition of $w_{2,t,\tau}$,

$$\begin{aligned}\sum_{j=1}^{n} \hat{\mathbf{v}}_t^j \leq& \sum_{\tau=1}^{t} w_{2,t,\tau} \left(\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F + \eta L (t-\tau) \sqrt{\frac{1-\mu_1}{1-\mu_2}}\right)^2 \\ =& \left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F^2 + 2\eta L \left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F \sqrt{\frac{1-\mu_1}{1-\mu_2}} \sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau) \\ &+ \eta^2 L^2 \left(\frac{1-\mu_1}{1-\mu_2}\right) \sum_{\tau=1}^{t} w_{2,t,\tau}(t-\tau)^2.\end{aligned}$$

By Cauchy-Schwarz inequality, it follows that:

$$\begin{aligned}\sum_{j=1}^{n}\hat{\mathbf{v}}_t^j \leq & \left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F^2 + 2\eta L\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sqrt{\sum_{\tau=1}^{t}w_{2,t,\tau}(t-\tau)^2} \\ & + \eta^2L^2\left(\frac{1-\mu_1}{1-\mu_2}\right)\left(\sum_{\tau=1}^{t}w_{2,t,\tau}(t-\tau)^2\right) \\ \leq & \left(\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sqrt{\sum_{\tau=1}^{t}w_{2,t,\tau}(t-\tau)^2}\right)^2 \\ = & \left(\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F + a_t\right)^2,\end{aligned}$$

where

$$a_t := \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sqrt{\sum_{\tau=1}^{t}w_{2,t,\tau}(t-\tau)^2}.$$

Since:

$$\begin{aligned}a_t^2 \leq & \eta^2L^2\left(\frac{1-\mu_1}{1-\mu_2}\right)\left(\frac{1-\mu_2}{1-\mu_2^t}\right)\left(\sum_{\tau=1}^{t}\mu_2^{t-\tau}(t-\tau)^2\right) \\ \leq & \eta^2L^2\left(\frac{1-\mu_1}{1-\mu_2}\right)\left(\sum_{\tau=1}^{\infty}\mu_2^{\tau}\tau^2\right) \\ \leq & \eta^2L^2\left(\frac{1-\mu_1}{1-\mu_2}\right)\left(\frac{\mu_2(1+\mu_2)}{(1-\mu_2)^3}\right) := a^2. \end{aligned} \tag{19}$$

Hence,

$$\sqrt{\sum_{j=1}^{n}\hat{\mathbf{v}}_t^j} \leq \left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F + a_t \leq \left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F + a. \tag{20}$$

Step 5: Lower-bounding $d_{t,\max}\|\nabla\mathcal{L}(\Theta_{t-1})\|_*$. Combining (18) and (20) gives:

$$
\begin{aligned}
& d_{t,\max}\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_* \\
\geq & \frac{\left\|\hat{M}_t\right\|\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}+n\epsilon/\sqrt{1-\mu_2^t}} \\
\geq & \frac{\left(\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F-\left\|\hat{M}_t-\nabla\mathcal{L}(\Theta_{t-1})\right\|\right)\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F+n\epsilon/\sqrt{1-\mu_2^t}+a} \\
\geq & \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F^2}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F+n\epsilon/\sqrt{1-\mu_2^t}+a}-\left\|\hat{M}_t-\nabla\mathcal{L}(\Theta_{t-1})\right\| \\
\geq & \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F^2}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F+n\epsilon/\sqrt{1-\mu_2^t}}\left(1-\frac{a}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F+n\epsilon/\sqrt{1-\mu_2^t}+a}\right)-\left\|\hat{M}_t-\nabla\mathcal{L}(\Theta_{t-1})\right\| \\
\geq & \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F^2}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F+n\epsilon/\sqrt{1-\mu_2^t}}-a-\left\|\hat{M}_t-\nabla\mathcal{L}(\Theta_{t-1})\right\| \\
\geq & \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F^2}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F+n\epsilon/\sqrt{1-\mu_2^t}}-a-\eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2} \\
\geq & \frac{\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F^2}{\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F+n\tilde{\epsilon}}-a-\eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2}
\end{aligned}
$$

where $\tilde{\epsilon}:=\epsilon/\sqrt{1-\mu_2}$ and a is given in (19).

Step 6: Upper-bounding the gradient norm and deriving rate. Now define $\phi_{n\tilde{\epsilon}}(x):=\frac{x^2}{x+n\tilde{\epsilon}}$, it then follows from (17) and (18) that:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\phi_{n\tilde{\epsilon}}(\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_F) \leq & \frac{1}{T}\sum_{t=1}^T d_{t,\max}\left\|\nabla\mathcal{L}(\Theta_{t-1}\right\|_*+a+\eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2} \\
\leq & \frac{\Delta}{\eta Tc^2}+\frac{2}{Tc^2}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^T\left\|\nabla\mathcal{L}(\Theta_{t-1})-\hat{M}_t\right\|_*+\frac{\eta L}{2c^2}\left(\frac{1-\mu_1}{1-\mu_2}\right) \\
& +a+\eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2} \\
\leq & \frac{\Delta}{\eta Tc^2}+\frac{2}{Tc^2}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^T\left\|\nabla\mathcal{L}(\Theta_{t-1})-\hat{M}_t\right\|_*+\frac{\eta L}{2c^2}\left(\frac{1-\mu_1}{1-\mu_2}\right) \\
& +a+\eta L\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1}{(1-\mu_1)^2} \\
\leq & \frac{\Delta}{\eta Tc^2}+\left(\frac{1-\mu_1}{1-\mu_2}\right)\frac{2\eta L\mu_1}{(1-\mu_1)^2c^2}+\frac{\eta L}{2c^2}\left(\frac{1-\mu_1}{1-\mu_2}\right) \\
& +\eta L\frac{\sqrt{\mu_2(1-\mu_1)(1+\mu_2)}}{(1-\mu_2)^2}+\sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\eta L\mu_1}{(1-\mu_1)^2} \\
= & \frac{\Delta}{\eta Tc^2}+\frac{\eta LC_\mu}{c^2},
\end{aligned}
$$

where

$$C_\mu := \left(\frac{1-\mu_1}{1-\mu_2}\right)\frac{2\mu_1}{(1-\mu_1)^2} + \frac{1}{2}\left(\frac{1-\mu_1}{1-\mu_2}\right) + \frac{c^2\sqrt{\mu_2(1-\mu_1)(1+\mu_2)}}{(1-\mu_2)^2} + \sqrt{\frac{1-\mu_1}{1-\mu_2}}\frac{\mu_1 c^2}{(1-\mu_1)^2}$$

is a constant that depends on μ_1 and μ_2. Then by Lemma 8,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\Theta_{t-1}\|_F \leq & \frac{1}{T}\sum_{t=1}^{T}\phi_{n\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_F\right) + (n\tilde{\epsilon})^{\frac{1}{2}}\frac{1}{T}\sum_{t=1}^{T}\sqrt{\phi_{n\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_F\right)} \\ \leq & \frac{1}{T}\sum_{t=1}^{T}\phi_{n\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_F\right) + \frac{(n\tilde{\epsilon})^{\frac{1}{2}}}{\sqrt{T}}\sqrt{\sum_{t=1}^{T}\phi_{n\tilde{\epsilon}}\left(\|\nabla\mathcal{L}(\Theta_{t-1}\|_F\right)} \\ \leq & \frac{\Delta}{\eta T c^2} + \frac{\eta L C_\mu}{c^2} + (n\tilde{\epsilon})^{\frac{1}{2}}\sqrt{\frac{\Delta}{\eta T c^2} + \frac{\eta L C_\mu}{c^2}}.\end{aligned}$$

In particular, if choosing $\eta = \mathcal{O}(T^{-\frac{1}{2}})$, $1-\mu_1 = \Theta(1)$, $1-\mu_2 = \Theta(1)$, $0 \leq \mu_1 \leq \mu_2 < 1$, $\epsilon = \mathcal{O}(T^{-\frac{1}{2}}n^{-1})$, and $c = \Theta(1)$, then:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\Theta_{t-1})\| \leq \mathcal{O}\left(T^{-\frac{1}{2}}\right)$$

for large $T > 0$. The proof is thus completed. ■

Appendix E. Proof of Theorem 6

This section contains the detailed proof of Theorem 6 for the convergence of NAMO-D in the stochastic setting.

Proof

Step 1: A uniform upper bound on stepsize. For each $t \geq 0$ and $\tau \leq t$, define:

$$w_{1,t,\tau} := \frac{1-\mu_1}{1-\mu_1^t}\mu_1^{t-\tau} \quad \text{and} \quad w_{2,t,\tau} := \frac{1-\mu_2}{1-\mu_2^t}\mu_2^{t-\tau}.$$

It satisfies that $\sum_{\tau=1}^{t} w_{1,t,\tau} = \sum_{\tau=1}^{t} w_{2,t,\tau} = 1$. For each $j = 1, \cdots, n$ and $t > 0$, let M_t^j denote the j-th column of M_t, and $\mathbf{v}_t^j$ denote the j-th element of $\mathbf{v}_t$. Write $\hat{M}_t := \frac{1}{1-\mu_1^t}M_t$, $\hat{\mathbf{v}}_t := \frac{1}{1-\mu_2^t}\mathbf{v}_t$, and

$$D_t := \operatorname{diag}\left(\min\left\{\max\left\{\mathbf{d}_t, c\bar{d}_t\mathbf{1}\right\}, \frac{1}{c}\bar{d}_t\mathbf{1}\right\}\right),$$

where $c \in (0,1]$ is the fixed clamping hyperparameter, $\mathbf{d}_t$ and $\bar{d}_t$ are given in Algorithm 2. Then:

$$\hat{M}_t = \frac{1-\mu_1}{1-\mu_1^t}\sum_{\tau=1}^{t}\mu_1^{t-\tau}G_\tau = \sum_{\tau=1}^{t} w_{1,t,\tau}G_\tau,$$

and

$$\hat{\mathbf{v}}_t = \frac{1-\mu_2}{1-\mu_2^t}\sum_{\tau=1}^{t}\mu_2^{t-\tau}\mathcal{N}_c(G_\tau)\odot\mathcal{N}_c(G_\tau) = \sum_{\tau=1}^{t} w_{2,t,\tau}\mathcal{N}_c(G_\tau)\odot\mathcal{N}_c(G_\tau).$$

By Lemma 7,

$$[D_t]_{jj} \leq \frac{\left\|\hat{M}_t^j\right\|}{\sqrt{\hat{\mathbf{v}}_t^j}} \leq \sqrt{\frac{1-\mu_1}{1-\mu_2}}, \quad \forall j. \tag{21}$$

Write:

$$d_{t,\max} := \max_j [D_t]_{jj}, \quad \text{and } d_{t,\min} := \min_j [D_t]_{jj}. \tag{22}$$

Then the condition number of D_t is given by:

$$\kappa(D_t) = \frac{d_{t,\max}}{d_{t,\min}} \leq \min\left\{\kappa_t, \frac{1}{c^2}\right\}, \tag{23}$$

where $\kappa_t := \kappa(\mathrm{diag}(\mathbf{d}_t))$ denotes the condition number of $\mathrm{diag}(\mathbf{d}_t)$.

Step 2: Expected descent inequality and averaging. Let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|\Theta_{t-1}]$ denote the conditional expectation given the previous iterates $\Theta_0, \cdots, \Theta_{t-1}$, and write $E_t := \hat{M}_t - \nabla\mathcal{L}(\Theta_{t-1})$. Then, by Lemma 11,

$$\begin{aligned}
&\mathbb{E}_t\left[\mathcal{L}(\Theta_t) - \mathcal{L}(\Theta_{t-1})\right] \\
\leq &\mathbb{E}_t\left[-\left\langle \nabla\mathcal{L}(\Theta_{t-1}), \eta O_t D_t\right\rangle\right] + \frac{\eta^2 L}{2}\mathbb{E}_t\left[\|D_t\|_2^2\right] \\
= &\mathbb{E}_t\left[-\left\langle \nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t, \eta O_t D_t\right\rangle\right] - \mathbb{E}_t\left[\eta\left\langle \hat{M}_t, O_t D_t\right\rangle\right] + \frac{\eta^2 L}{2}\mathbb{E}_t\left[\|D_t\|_2^2\right] \\
\leq &\left(\mathbb{E}_t\left[\eta d_{t,\max}\left\|\nabla\mathcal{L}(\Theta_{t-1}) - \hat{M}_t\right\|_*\right] - \mathbb{E}_t\left[\eta d_{t,\min}\left\|\hat{M}_t\right\|_*\right]\right) + \frac{\eta^2 L}{2}\mathbb{E}_t\left[d_{t,\max}^2\right] \\
\leq &-\mathbb{E}_t\left[\eta d_{t,\min}\left\|\hat{M}_t\right\|_*\right] + \mathbb{E}_t\left[\eta d_{t,\max}\|E_t\|_*\right] + \frac{\eta^2 L}{2}\mathbb{E}_t\left[d_{t,\max}^2\right],
\end{aligned}$$

where $d_{t,\max}$ and $d_{t,\min}$ are defined in (22). Rearranging the terms gives:

$$\mathbb{E}_t\left[d_{t,\min}\left\|\hat{M}_t\right\|_*\right] \leq \mathbb{E}_t\left[\mathcal{L}(\Theta_{t-1}) - \mathcal{L}(\Theta_t)\right] + \mathbb{E}_t\left[d_{t,\max}\|E_t\|_*\right] + \frac{\eta L}{2}\mathbb{E}_t\left[d_{t,\max}^2\right].$$

Then by the law of total expectation and (21),

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[d_{t,\min}\left\|\hat{M}_t\right\|_*\right] \leq &\frac{\Delta}{\eta T} + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[d_{t,\max}\|E_t\|_*\right] + \frac{\eta L}{2T}\sum_{t=1}^T \mathbb{E}\left[d_{t,\max}^2\right] \\
\leq &\frac{\Delta}{\eta T} + \frac{1}{T}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^T \mathbb{E}\left[\|E_t\|_*\right] + \frac{\eta L(1-\mu_1)}{2(1-\mu_2)}.
\end{aligned} \tag{24}$$

Step 3: Bounding the distance between bias-corrected momentum and true gradient. For each t, it satisfies:

$$\begin{aligned}
E_t = &\hat{M}_t - \mathbb{E}\left[\hat{M}_t\right] + \mathbb{E}\left[\hat{M}_t\right] - \nabla\mathcal{L}(\Theta_{t-1}) \\
= &\sum_{\tau=1}^t w_{1,t,\tau}\left(G_\tau - \nabla\mathcal{L}(\Theta_{\tau-1})\right) + \sum_{\tau=1}^t w_{1,t,\tau}\left(\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right)
\end{aligned}$$

Hence, by (21),

$$\begin{aligned}
\mathbb{E}\left[\|E_t\|_F^2\right] \leq & \mathbb{E}\left[\left\|\sum_{\tau=1}^{t} w_{1,t,\tau}\left(G_\tau - \nabla\mathcal{L}(\Theta_{\tau-1})\right)\right\|_F^2\right] + \left\|\sum_{\tau=1}^{t} w_{1,t,\tau}\left(\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right)\right\|_F^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right)^2 \left(\sum_{\tau=1}^{t} \mu_1^{2(t-\tau)}\right) \frac{\sigma^2}{b} + \sum_{\tau=1}^{t} w_{1,t,\tau} \left\|\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right\|_F^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right)^2 \left(\frac{1-\mu_1^{2t}}{1-\mu_1^2}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \sum_{\tau=1}^{t} \mu_1^{t-\tau} L^2 \left\|\Theta_{\tau-1} - \Theta_{t-1}\right\|_2^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_1^t}\right) L^2\eta^2 \sum_{\tau=1}^{t} \mu_1^{t-\tau} \left(\sum_{s=\tau}^{t-1} \|D_s\|_2\right)^2 \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left[\left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_2}\right) L^2\eta^2 \sum_{\tau=1}^{t} \mu_1^{t-\tau} (t-\tau)^2\right] \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left[\left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{1-\mu_1}{1-\mu_2}\right) L^2\eta^2 \sum_{\tau=1}^{\infty} \mu_1^{\tau} \tau^2\right] \\
= & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \left[\left(\frac{1+\mu_1^t}{1+\mu_1}\right) \frac{\sigma^2}{b} + \left(\frac{\mu_1(1+\mu_1)}{(1-\mu_2)(1-\mu_1)^2}\right) L^2\eta^2\right] \\
\leq & \left(\frac{1-\mu_1}{1-\mu_1^t}\right) \frac{\sigma^2}{b} + \frac{\mu_1(1+\mu_1)L^2\eta^2}{(1-\mu_2)(1-\mu_1)(1-\mu_1^t)}.
\end{aligned}$$

Then by Lemma 9, it follows that:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|E_t\|_F^2\right] \leq & \left(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{1-\mu_1^t}\right) \left(\frac{\sigma^2(1-\mu_1)}{b} + \frac{\mu_1(1+\mu_1)L^2\eta^2}{(1-\mu_1)(1-\mu_2)}\right) \\
\leq & \left(1 + \frac{\mu_1}{(1-\mu_1)T} - \frac{1}{T\ln\mu_1} \ln\left(\frac{1-\mu_1^T}{1-\mu_1}\right)\right) \left(\frac{\sigma^2(1-\mu_1)}{b} + \frac{\mu_1(1+\mu_1)L^2\eta^2}{(1-\mu_1)(1-\mu_2)}\right).
\end{aligned}$$

By Cauchy-Schwarz inequality and Jensen's inequality,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|E_t\|_*\right] \leq & \sqrt{\frac{r}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|E_t\|_F^2\right]} \\
\leq & \sqrt{1 + \frac{\mu_1}{(1-\mu_1)T} - \frac{1}{T\ln\mu_1} \ln\left(\frac{1-\mu_1^T}{1-\mu_1}\right)} \left(\frac{\sigma\sqrt{r(1-\mu_1)}}{\sqrt{b}} + L\eta\sqrt{\frac{r\mu_1(1+\mu_1)}{(1-\mu_2)(1-\mu_1)}}\right).
\end{aligned} \tag{25}$$

Step 4: Bounding $\mathbb{E}\left[\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}\right]$. For each j, let G_τ^j denote the j-th column of G_τ. By Minkowski inequality and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}\right] =& \mathbb{E}\left[\sqrt{\sum_{j=1}^n \sum_{\tau=1}^t w_{2,t,\tau}\left\|G_\tau^j\right\|^2}\right] = \mathbb{E}\left[\sqrt{\sum_{\tau=1}^t w_{2,t,\tau}\left\|G_\tau\right\|_F^2}\right] \\
\leq& \mathbb{E}\left[\sqrt{\sum_{\tau=1}^t w_{2,t,\tau}\left\|\nabla\mathcal{L}(\Theta_{\tau-1}) - G_\tau\right\|_F^2}\right] + \mathbb{E}\left[\sqrt{\sum_{\tau=1}^t w_{2,t,\tau}\left\|\nabla\mathcal{L}(\Theta_{\tau-1})\right\|_F^2}\right] \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \mathbb{E}\left[\sqrt{\sum_{\tau=1}^t w_{2,t,\tau}\left\|\nabla\mathcal{L}(\Theta_{\tau-1}) - \nabla\mathcal{L}(\Theta_{t-1})\right\|_F^2}\right] \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\mathbb{E}\left[\sqrt{\sum_{\tau=1}^t w_{2,t,\tau}\left(\sum_{s=\tau}^{t-1}\left\|D_s\right\|_2\right)^2}\right] \\
\overset{(21)}{\leq}& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2^t}}\sqrt{\sum_{\tau=1}^t \mu_2^{t-\tau}(t-\tau)^2} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2^t}}\sqrt{\sum_{\tau=1}^\infty \mu_2^{t-\tau}(t-\tau)^2} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{1-\mu_1}{1-\mu_2^t}}\sqrt{\sum_{\tau=1}^\infty \mu_2^{\tau}\tau^2} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \eta L\sqrt{\frac{(1-\mu_1)\mu_2(1+\mu_2)}{(1-\mu_2^t)(1-\mu_2)^3}} \\
\leq& \frac{\sigma}{\sqrt{b}} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|_F\right] + \sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2^t)(1-\mu_2)^3}}.
\end{aligned}$$

Hence,

$$\mathbb{E}\left[\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j} + \frac{n\epsilon}{\sqrt{1-\mu_2^t}}\right] \leq \mathbb{E}\left[\left\|\nabla\mathcal{L}(\Theta_{t-1})\right\|\right] + a_t, \tag{26}$$

where

$$a_t := \frac{\sigma}{\sqrt{b}} + \sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2^t)(1-\mu_2)^3}} + \frac{n\epsilon}{\sqrt{1-\mu_2^t}}.$$

Then by Lemma 10,

$$\frac{1}{T}\sum_{t=1}^T a_t \leq \frac{\sigma}{\sqrt{b}} + \left(\sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2)^3}} + n\epsilon\right)\left(1 - \frac{2\ln(1+\sqrt{1-\mu_2^T})}{T\ln\mu_2}\right). \tag{27}$$

Step 5: Lower-bounding $\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]$ and relating to $\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]$. By Cauchy-Schwarz inequality,

$$\begin{aligned}
\left(\mathbb{E}\left[\left\|\hat{M}_t\right\|_F\right]\right)^2 =& \left(\mathbb{E}\left[\frac{\left\|\hat{M}_t\right\|_F}{\left(\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}+n\epsilon_t\right)^{\frac{1}{2}}}\cdot\left(\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}+n\epsilon_t\right)^{\frac{1}{2}}\right]\right)^2 \\
\leq& \mathbb{E}\left[\frac{\left\|\hat{M}_t\right\|_F^2}{\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}+n\epsilon_t}\right]\mathbb{E}\left[\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}+n\epsilon_t\right] \\
\leq& \mathbb{E}\left[\sqrt{\frac{\sum_{j=1}^n\left\|\hat{M}_t^j\right\|^2}{\sum_{j=1}^n \hat{\mathbf{v}}_t^j+n\epsilon_t}}\cdot\left\|\hat{M}_t\right\|_F\right]\mathbb{E}\left[\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}+n\epsilon_t\right] \\
\leq& \mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]\mathbb{E}\left[\sqrt{\sum_{j=1}^n \hat{\mathbf{v}}_t^j}+n\epsilon_t\right],
\end{aligned}$$

where $\epsilon_t := \epsilon/\sqrt{1-\mu_2^t}$. Combining the above with (26) gives:

$$\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right] \geq \frac{\left(\mathbb{E}\left[\left\|\hat{M}_t\right\|_F\right]\right)^2}{\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]+a_t} \geq \frac{\left(\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]-\mathbb{E}\left[\|E_t\|_F\right]\right)^2}{\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]+a_t}.$$

Rearranging the terms gives:

$$(\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right])^2-\left(2\mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]\right)\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right]-a_t\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]+\mathbb{E}\left[\|E_t\|_F\right]^2\leq 0.$$

Then solving for $\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|\right]$ gives

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right] \\
\leq& \frac{2\mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]+\sqrt{\left(2\mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]\right)^2+4a_t\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]-4\mathbb{E}\left[\|E_t\|_F\right]^2}}{2} \\
\leq& \mathbb{E}\left[\|E_t\|_F\right]+\frac{1}{2}\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]+\sqrt{\frac{1}{4}\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]^2+\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]} \\
\leq& \mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]+\sqrt{\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]} \\
\leq& \mathbb{E}\left[\|E_t\|_F\right]+\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]+\sqrt{\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]}.
\end{aligned}$$

Step 6: Deriving the convergence rate By Cauchy-Schwarz inequality and (23), it then follows that:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right] \leq& \frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\|E_t\|_F\right]+\frac{1}{T}\min\left\{\kappa_t,\frac{1}{c^2}\right\}\sum_{t=1}^T\mathbb{E}\left[d_{t,\min}\left\|\hat{M}_t\right\|_F\right] \\
&+\sqrt{\frac{1}{T}\sum_{t=1}^T\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)}\sqrt{\frac{1}{T}\min\left\{\kappa_t,\frac{1}{c^2}\right\}\sum_{t=1}^T\mathbb{E}\left[d_{t,\max}\left\|\hat{M}_t\right\|_F\right]}.
\end{aligned}$$

Combining the above with (24) and (27) gives:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|_F\right] \\
\leq&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\Delta}{\eta T c^2}+\frac{\eta L(1-\mu_1)}{2c^2(1-\mu_2)}+\frac{1}{Tc^2}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right] \\
&+\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|E_t\|_F\right]+a_t\right)}\cdot\sqrt{\frac{\Delta}{\eta T c^2}+\frac{1}{Tc^2}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\eta L(1-\mu_1)}{2c^2(1-\mu_2)}} \\
\leq&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\Delta}{\eta T c^2}+\frac{\eta L(1-\mu_1)}{2c^2(1-\mu_2)}+\frac{1}{Tc^2}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right] \\
&+\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_F\right]+\frac{\sigma}{\sqrt{b}}+\left(\sqrt{2}\eta L\sqrt{\frac{(1-\mu_1)}{(1-\mu_2)^3}}+n\epsilon\right)\left(1-\frac{2\ln(1+\sqrt{1-\mu_2^T})}{T\ln\mu_2}\right)} \\
&\cdot\sqrt{\frac{\Delta}{\eta T c^2}+\frac{1}{Tc^2}\sqrt{\frac{1-\mu_1}{1-\mu_2}}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]+\frac{\eta L(1-\mu_1)}{2c^2(1-\mu_2)}}.
\end{aligned}
$$

In particular, for large $T>0$, if choosing $\eta=\mathcal{O}(T^{-\frac{3}{4}})$, $1-\mu_1=\Theta(T^{-\frac{1}{2}})$, $1-\mu_2=\Theta(T^{-\frac{1}{2}})$, $\epsilon=\mathcal{O}(T^{-\frac{1}{2}})$, and $c=\Theta(1)$, then, by (25),

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|E_t\|_*\right]\leq\mathcal{O}\left(T^{-\frac{1}{4}}\right),
$$

and it follows that:

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\Theta_{t-1})\|\right]\leq\mathcal{O}\left(T^{-\frac{1}{4}}+\sqrt{\sigma}b^{-\frac{1}{4}}T^{-\frac{1}{8}}\right),
$$

where b is the batch size. The proof is thus completed. ■