NON-EUCLIDEAN SGD FOR STRUCTURED OPTIMIZATION: UNIFIED ANALYSIS AND IMPROVED RATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, several instances of non-Euclidean SGD, including SignSGD, Lion, and Muon, have attracted significant interest from the optimization community due to their practical success in training deep neural networks. Consequently, a number of works have attempted to explain this success by developing theoretical convergence analyses. Unfortunately, these results cannot properly justify the superior performance of these methods, as they could not beat the convergence rate of vanilla Euclidean SGD. We resolve this important open problem by developing a new unified convergence analysis under the structured smoothness and gradient noise assumption. In particular, our results indicate that non-Euclidean SGD **(i)** can exploit the sparsity or low-rank structure of the upper bounds on the Hessian and gradient noise, **(ii)** can provably benefit from popular algorithmic tools such as extrapolation or momentum variance reduction, and **(iii)** can match the state-of-the-art convergence rates of adaptive and more complex optimization algorithms such as AdaGrad and Shampoo.

1 INTRODUCTION

Optimization algorithms play a fundamental role in deep neural network training. Over the past couple of decades, a large number of optimization methods have been developed and successfully applied in practice. These range from vanilla stochastic gradient descent (SGD) with momentum (Robbins & Monro, 1951; Polyak, 1964) to more advanced and complex adaptive algorithms such as AdaGrad (Duchi et al., 2011), Shampoo (Gupta et al., 2018), and Adam/AdamW (Kingma, 2014; Loshchilov & Hutter, 2017). The latter have become the standard choice in most deep learning tasks.

However, recently, the algorithmic memory requirements have been tightening up due to the significant growth of model sizes, especially of the current large language models (LLMs) (Dubey et al., 2024; Liu et al., 2024a; Achiam et al., 2023; Team et al., 2023). Consequently, the optimization community has started shifting its attention to more memory-efficient, non-Euclidean variants of SGD. These include SignSGD (Bernstein et al., 2018), Lion (Chen et al., 2023b), and Muon (Jordan et al., 2024), the latter starting to replace AdamW as the standard choice in training LLMs (Team et al., 2025; Zeng et al., 2025; Liu et al., 2025a). Unfortunately, despite the practical effectiveness of these methods, their theoretical understanding remains highly limited. In particular, existing convergence theory remains insufficient to explain the superior practical performance of these methods even over vanilla SGD, as we will discuss later. In this paper, we aim to resolve this important open question and advance the theoretical understanding of non-Euclidean SGD methods.

1.1 NON-EUCLIDEAN SGD: OVERVIEW

Euclidean methods. The convergence properties of Euclidean methods are relatively well-studied. In particular, Ghadimi & Lan (2013) established the iteration complexity of vanilla SGD for non-convex Lipschitz-smooth functions with uniformly bounded gradient noise. Additionally, Cutkosky & Mehta (2020) analyzed the convergence of normalized SGD with momentum under the same assumptions. Both complexity results have optimal dependence on the target precision $\epsilon > 0$, i.e., they cannot be improved by any stochastic first-order method without extra assumptions (Arjevani et al., 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Sign-based methods. The most notable sign-based SGD methods include SignSGD with momentum (Bernstein et al., 2018) and Lion (Chen et al., 2023b). The convergence rates for these methods were established by Sun et al. (2023) and Dong et al. (2024), respectively. Note that we will not consider Lion further because its theoretical properties are very similar to those of SignSGD with momentum. In particular, generalizing the results of this paper to Lion is straightforward.

Orthogonolized methods. Orthogonalized SGD methods are specifically designed for functions defined on the space of $m \times n$ matrices. They apply the so-called orthogonalization procedure $G \mapsto (GG^\top)^{\frac{1}{2}}G$ to the current gradient estimator,[1] where $G \in \mathbb{R}^{m\times n}$. Orthogonalization was first proposed by Carlson et al. (2015) and was later rediscovered by Jordan et al. (2024), leading to the development of Muon – an orthogonalized SGD with momentum. This algorithm has remarkable practical performance across a wide range of deep learning tasks. Pethick et al. (2025); Kovalev (2025b) developed a convergence analysis of Muon under non-Euclidean Lipschitz smoothness with respect to the spectral norm and uniformly bounded variance with respect to the Euclidean norm. Furthermore, because of Muon's exceptional performance, many works have appeared studying various aspects of this method (Khaled et al., 2025; Zhang et al., 2025; Gruntkowska et al., 2025; Crawshaw et al., 2025; Wang et al., 2025; Shen et al., 2025; He et al., 2025a; Zhang & Gao, 2025; Chen et al., 2025; He et al., 2025b; Lau et al., 2025; Huang et al., 2025; Riabinin et al., 2025; Shulgin et al., 2025; Liu et al., 2025b; Sato et al., 2025).

General non-Euclidean methods. Motivated by the interest in sign-based and orthogonalized gradient methods, Pethick et al. (2025); Kovalev (2025b) developed general analysis of non-Euclidean SGD with momentum and arbitrary norms, which recovers the convergence results for Muon and SignSGD.

Connection with adaptive methods. Recently, Xie et al. (2025); Kovalev (2025a) developed a unified convergence analysis for the adaptive meta-algorithm of Gupta et al. (2017), which recovers many popular adaptive methods as special cases, including AdaGrad-Norm (Streeter & McMahan, 2010), AdaGrad, and ASGO/One-sided Shampoo (An et al., 2025; Xie et al., 2025), a more streamlined variant of Shampoo. More importantly, Xie et al. (2025); Kovalev (2025a) showed how various instances of the adaptive meta-algorithm can exploit non-Euclidean smoothness and bounded gradient noise properties or their variants with respect to a family of non-Euclidean norms.

1.2 Summary of Contributions

To the best of our knowledge, *existing theoretical explanations for the superior practical performance of non-Euclidean SGD remain insufficient*. In particular, existing convergence guarantees[2] do not improve upon the corresponding guarantees for Euclidean SGD methods. We resolve this issue by developing a unified convergence analysis of non-Euclidean SGD with momentum and weight decay (Algorithm 1) for a family of non-Euclidean norms, under structured smoothness and gradient noise assumptions. Algorithm 1 recovers normalized SGD with momentum (Cutkosky & Mehta, 2020), SignSGD with momentum (Bernstein et al., 2018), and Muon (Jordan et al., 2024). Our analysis yields *new state-of-the-art convergence guarantees for Muon and SignSGD, which beat Euclidean SGD and may finally explain the superior practical performance of these methods*. In addition, we establish the convergence guarantees for Algorithm 1 with extrapolation (option 12b) and momentum variance reduction (option 12c) tricks, which provably accelerate convergence. Finally, under the additional convexity assumption for the objective function, we show that *Muon and SignSGD with momentum variance reduction can match the existing state-of-the-art convergence rates for their respective adaptive counterparts: ASGO/One-sided Shampoo (An et al., 2025; Xie et al., 2025) and AdaGrad (Duchi et al., 2011; Liu et al., 2024b)*. Thus, our results support the recent trend of abandoning adaptive methods in favor of simpler and more memory-efficient instances of non-Euclidean SGD.

2 Preliminaries

Formally, in this paper, we are concerned with the problem of minimizing a continuously differentiable objective function $f(\cdot)\colon \mathcal{X} \to \mathbb{R}$ defined on a finite-dimensional Euclidean space $\mathcal{X}$.

[1] Practical Muon implementations approximate orthogonalization using iterative procedures, e.g., Newton-Schulz. Considering inexact orthogonalization is beyond the scope of this paper and is left for future research.

[2] Refer to Section A for an overview of existing convergence results.

Notation. $\mathbb{N}_0$ is the set of non-negative integers; $\mathbb{R}_{\geq}$ is the set of non-negative real numbers; $\mathbb{L}$ is the space of linear operators $\mathcal{X} \to \mathcal{X}$; $\mathbb{S}$ and $\mathbb{S}_{\succ}$ are the sets of self-adjoint and self-adjoint positive definite operators $\mathcal{X} \to \mathcal{X}$, respectively; $\|x\|_{\mathbf{B}} = \sqrt{\langle x, \mathbf{B}x\rangle}$, where $x \in \mathcal{X}$ and $\mathbf{B} \in \mathbb{S}_{\succ}$; $\|\mathbf{B}\|_{\mathrm{tr}}$ and $\|\mathbf{B}\|_{\mathrm{op}}$ are the nuclear and spectral norms of an operator $\mathbf{B} \in \mathbb{L}$; for all $z \in \mathcal{X}$, by $z\langle z, \cdot\rangle \in \mathbb{S}$ we denote the rank-1 (or rank-0 if $z = 0$) self-adjoint operator $x \mapsto z\langle z, x\rangle$; $\mathbf{I} \in \mathbb{S}$ denotes the identity operator. For a vector $x \in \mathbb{R}^d$, by $[x]_{[i]}$ we denote its i-th component; $\mathrm{D}_f(\cdot\,;\cdot)$ denotes the Bregman divergence associated with the function $f(\cdot)$.

2.1 Assumptions on the Objective Function

In this section, we describe the assumptions that we impose on the objective function $f(\cdot)$. Throughout this work, we will always use Assumptions 1 and 2 for the analysis of Algorithm 1, while Assumptions 3 and 4 will only be used for the analysis of momentum with extrapolation and momentum variance reduction options, respectively. Note that all Assumptions 1 to 4 are given in terms of weighted Euclidean norms of the form $\|\cdot\|_{\mathbf{H}}$, where $\mathbf{H} \in \mathbb{S}_{\succ}$. The main reason is that when combined with Assumption 5, they allow us to describe the non-Euclidean structure in the properties of the objective function $f(\cdot)$. More details are given in Section 2.2.

Smoothness and unbiased gradient estimator. We assume that the objective function $f(\cdot)$ is 1-smooth with respect to the weighted Euclidean norm $\|\cdot\|_{\mathbf{L}}$, where $\mathbf{L} \in \mathbb{S}_{\succ}$ is a self-adjoint positive definite operator. That is, the following Assumption 1 holds:

$$|\mathrm{D}_f(x; x')| \leq \tfrac{1}{2}\|x - x'\|_{\mathbf{L}}^2 \quad \text{for all } \; x, x' \in \mathcal{X}. \tag{A1}$$

We also assume that there exists a stochastic estimator $g(\cdot;\xi)\colon \mathcal{X} \to \mathcal{X}$ of the gradient $\nabla f(\cdot)$, where ξ is a random variable sampled from some probability distribution $\mathcal{D}$. We assume that $g(\cdot;\xi)$ is unbiased and has bounded variance, i.e., the following Assumption 2 holds:

$$\mathbb{E}_{\xi\sim\mathcal{D}}[g(x;\xi)] = \nabla f(x) \quad \text{and} \quad \mathbb{E}_{\xi\sim\mathcal{D}}[\|n(x;\xi)\|_{\mathbf{\Sigma}^{-1}}^2] \leq \|\mathbf{\Sigma}\|_{\mathrm{tr}} \quad \text{for all } \; x \in \mathcal{X}, \tag{A2}$$

where $\mathbf{\Sigma} \in \mathbb{S}_{\succ}$ is a self-adjoint positive definite operator, $n(x;\xi) = g(x;\xi) - \nabla f(x)$ is the noise of the gradient estimator.

Second-order smoothness for momentum with extrapolation. To achieve improved convergence rates for SGD using momentum with extrapolation (eq. (12b) in Algorithm 1), we assume that the objective function $f(x)$ is second-order Lipschitz-smooth with respect to the weighted Euclidean norm $\|\cdot\|_{\mathbf{T}}$, where $\mathbf{T} \in \mathbb{S}_{\succ}$. That is, the following Assumption 3 holds:

$$\|[\nabla^2 f(x) - \nabla^2 f(x')](x - x')\|_{\mathbf{T}^{-1}}^2 \leq \|\mathbf{T}\|_{\mathrm{tr}}^{-1} \cdot \|x - x'\|_{\mathbf{T}}^4 \quad \text{for all } \; x, x' \in \mathcal{X}. \tag{A3}$$

Note that the simplified version of this assumption with $\mathbf{T} \propto \mathbf{I}$ was previously used by Cutkosky & Mehta (2020) to obtain improved convergence rates for normalized SGD with momentum and extrapolation.

Mean-square Lipschitz gradients for momentum variance reduction. To achieve even better convergence rates for SGD with momentum variance reduction (eq. (12c) in Algorithm 1), we assume that the stochastic gradient estimator $g(\cdot;\xi)$ is mean-square 1-Lipschitz with respect to the weighted Euclidean norm $\|\cdot\|_{\mathbf{M}}$, where $\mathbf{M} \in \mathbb{S}_{\succ}$. That is, the following Assumption 4 holds:

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|g(x;\xi) - g(x';\xi)\|_{\mathbf{M}^{-1}}^2] \leq \|x - x'\|_{\mathbf{M}}^2 \quad \text{for all } \; x, x' \in \mathcal{X}. \tag{A4}$$

Note that the simplified version of this assumption with $\mathbf{M} \propto \mathbf{I}$ and its variants was previously used by Cutkosky & Orabona (2019); Fatkhullin et al. (2022) to analyze SGD with momentum variance reduction.

On the restrictiveness of the assumptions. In the case where $\mathbf{L}, \mathbf{\Sigma} \propto \mathbf{I}$, Assumptions 1 and 2 reduces to the standard Lipschitz continuity of the gradient $\nabla f(\cdot)$ and the uniformly bounded noise of the gradient estimator $g(\cdot;\xi)$, which may not be the best choice for modeling the properties of complex functions such as deep neural networks. For instance, Zhang et al. (2019); Chen et al. (2023c); Gorbunov et al. (2024); Vankov et al. (2024); Borodich & Kovalev (2025) analyzed the convergence of gradient methods under the (L_0, L_1)-smoothness, and Vaswani et al. (2019); Faw et al. (2022); Attia & Koren (2023); Sadchikov et al. (2024) analyzed SGD under the affine variance. Both are more realistic assumptions than Lipschitz smoothness and bounded variance. However, we do not consider these settings because our main goal is to show how non-Euclidean SGD can exploit the structure in the operators $\mathbf{L}, \mathbf{\Sigma}$ (and $\mathbf{T}, \mathbf{M}$ in Assumptions 3 and 4), and applying the techniques from our paper to generalized variants of Assumptions 1 and 2 should be straightforward.

Table 1: Instances of Algorithm 1 induced by the choice of the linear space $\mathcal{X}$, the space of operators $\mathcal{H}$ satisfying Assumption 5, and the non-Euclidean norm $\rho(\cdot)$ defined in eq. (1): normalized SGD (Cutkosky & Mehta, 2020), SignSGD (Bernstein et al., 2018), and Muon (Jordan et al., 2024), along with their adaptive counterparts: AdaGrad-Norm (Streeter & McMahan, 2010), AdaGrad (Duchi et al., 2011), and ASGO/One-sided Shampoo (An et al., 2025; Xie et al., 2025).

Algorithm 1	**Adaptive Counterpart**	$\mathcal{X}$	$\mathcal{H}$	$\rho(\cdot)$
Normalized SGD	AdaGrad-Norm	$\mathbb{R}^d$	$\{g \mapsto \beta g : \beta \in \mathbb{R}\}$	$\frac{1}{\sqrt{d}}\|\cdot\|_2$
SignSGD	AdaGrad	$\mathbb{R}^d$	$\{g \mapsto \boldsymbol{b} \odot g : \boldsymbol{b} \in \mathbb{R}^d\}$	$\|\cdot\|_\infty$
Muon	One-sided Shampoo	$\mathbb{R}^{m\times n}$	$\{G \mapsto BG : B \in \mathbb{S}^m\}$	$\frac{1}{\sqrt{n}}\|\cdot\|_{\mathrm{op}}$

2.2 Main Structural Assumption

We impose the additional structural assumptions on the operators $\mathbf{L}, \mathbf{M}, \mathbf{T}, \mathbf{\Sigma} \in \mathbb{S}_\succ$. In particular, we assume that $\mathbf{L}$, $\mathbf{M}$, $\mathbf{T}$, and $\mathbf{\Sigma}$ belong to a certain space of linear operators $\mathcal{H} \subset \mathbb{S}$ that satisfies the conditions in Assumption 5.

Assumption 5. *$\mathbf{L}, \mathbf{M}, \mathbf{T}, \mathbf{\Sigma} \in \mathcal{H} \cap \mathbb{S}_\succ$, where $\mathcal{H} \subset \mathbb{S}$ is a linear subspace of self-adjoint operators, satisfying the following conditions:*

- **(i)** *The space $\mathcal{H}$ contains identity operator: $\mathbf{I} \in \mathcal{H}$.*
- **(ii)** *The space $\mathcal{H}$ is closed under symmetric multiplications: $\mathbf{HFH} \in \mathcal{H}$ for all $\mathbf{H}, \mathbf{F} \in \mathcal{H}$*
- **(iii)** *The orthogonal projection onto $\mathcal{H}$ is order-preserving: $\mathrm{proj}_{\mathcal{H}}(\mathbf{H}) \in \mathbb{S}_\succ$ for all $\mathbf{H} \in \mathbb{S}_\succ$.*

Next, we define functions $\rho(\cdot)\colon \mathcal{X} \to \mathbb{R}_\geq$ and $\rho_*(\cdot)\colon \mathcal{X} \to \mathbb{R}_\geq$ as follows:

$$\rho(x) = \|\sqrt{\mathrm{proj}_{\mathcal{H}}(\mathbf{X})}\|_{\mathrm{op}} \quad \text{and} \quad \rho_*(x) = \|\sqrt{\mathrm{proj}_{\mathcal{H}}(\mathbf{X})}\|_{\mathrm{tr}}, \quad \text{where} \quad \mathbf{X} = x\langle x, \cdot\rangle. \tag{1}$$

One can verify that both functions $\rho(\cdot)$ and $\rho_*(\cdot)$ are norms, possibly non-Euclidean, as indicated by Lemma 1. Moreover, Lemma 1 implies that these norms are dual to each other.

Lemma 1 (↓). *$\rho(\cdot)$ and $\rho_*(\cdot)$ are norms, that is, they are subadditive, absolutely homogeneous, and positive-definite functions. Moreover, these norms are dual to each other:*

$$\rho(x) = \sup_{\rho_*(y)\leq 1}\langle x, y\rangle \quad \textit{and} \quad \rho_*(y) = \sup_{\rho(x)\leq 1}\langle x, y\rangle. \tag{2}$$

In Table 1, we describe the main instances of Algorithm 1 induced by the choice of the operator space $\mathcal{H}$ and the corresponding non-Euclidean norm $\rho(\cdot)$: normalized SGD and SignSGD with momentum and weight decay, and Muon with weight decay. In addition, we provide details on how these algorithms exploit the structure of the operators $\mathbf{L}, \mathbf{M}, \mathbf{T}, \mathbf{\Sigma}$ in Assumption 5 for improved convergence guarantees in Section 3.3.

Connection with non-Euclidean setting. In the following Lemma 2, we establish the connection between the non-Euclidean norm $\rho(\cdot)$ and an arbitrary weighted Euclidean norm of the form $\|\cdot\|_{\mathbf{H}}$, where $\mathbf{H} \in \mathcal{H} \cap \mathbb{S}_\succ$.

Lemma 2 (↓). *Let Assumption 5 hold and let $\mathbf{H} \in \mathcal{H} \cap \mathbb{S}_\succ$. Then the following inequalities hold:*

$$\|x\|_{\mathbf{H}}^2 \leq \|\mathbf{H}\|_{\mathrm{tr}} \cdot \rho^2(x) \quad \textit{and} \quad \|x\|_{\mathbf{H}^{-1}}^2 \geq \|\mathbf{H}\|_{\mathrm{tr}}^{-1} \cdot \rho_*^2(x) \quad \textit{for all} \;\; x \in \mathcal{X}. \tag{3}$$

Using Lemma 2, it is not hard to verify that the assumptions on the objective function from Section 2.1 imply their non-Euclidean counterparts, that is, for all $x, x' \in \mathcal{X}$, the following relations hold:

$$\text{Assumption 1} \;\Rightarrow\; |\mathrm{D}_f(x; x')| \leq \tfrac{1}{2}\|\mathbf{L}\|_{\mathrm{tr}} \cdot \rho^2(x - x'); \tag{4}$$

$$\text{Assumption 2} \;\Rightarrow\; \mathbb{E}_{\xi\sim\mathcal{D}}\big[\rho_*^2(n(x;\xi))\big] \leq \|\mathbf{\Sigma}\|_{\mathrm{tr}}^2; \tag{5}$$

$$\text{Assumption 3} \;\Rightarrow\; \rho_*\big([\nabla^2 f(x) - \nabla^2 f(x')](x - x')\big) \leq \|\mathbf{T}\|_{\mathrm{tr}} \cdot \rho^2(x - x'); \tag{6}$$

$$\text{Assumption 4} \;\Rightarrow\; \mathbb{E}_{\xi\sim\mathcal{D}}\big[\rho_*^2(g(x;\xi) - g(x';\xi))\big] \leq \|\mathbf{M}\|_{\mathrm{tr}}^2 \cdot \rho^2(x - x'). \tag{7}$$

It is worth highlighting that the reverse implications do not hold. Indeed, Assumptions 1 to 4 are not implied by the non-Euclidean counterparts in general. However, the additional structure implied by these assumptions is exactly what allows us to achieve the main goal: justifying the efficiency of non-Euclidean SGD. This also contrasts with the existing state-of-the-art results (Pethick et al., 2025; Kovalev, 2025b), which used the non-Euclidean assumptions above, but could not improve upon the baseline Euclidean SGD.

Connection with adaptive methods. Xie et al. (2025); Kovalev (2025a) developed a unified convergence analysis for the adaptive-meta algorithm (Gupta et al., 2017), which recovers AdaGrad-Norm, AdaGrad, and ASGO/One-sided Shampoo as special cases. As we will see further, these algorithms are closely related to normalized SGD, SignSGD, and Muon, respectively. In particular, our Assumption 5 is similar to Assumption 1 by Kovalev (2025a), and the definition of $\rho(\cdot)$ in eq. (1) is identical to that by Kovalev (2025a), used in the analysis of the adaptive meta-algorithm. More details are available in Section 3.1.

2.3 Constrained Optimization Setting

Further, in this paper, we focus on the following constrained variant of the optimization problem:

$$f^* = \min_{x \in Q} f(x), \tag{8}$$

where $Q = \{x \in \mathcal{X} : \rho(x) \leq \mathcal{R}\}$ is a ball of possibly infinite radius $\mathcal{R} \in (0, +\infty]$, defined in terms of the non-Euclidean norm $\rho(\cdot)$. As we will see further, this additional constraint does not introduce any overhead, as it is automatically handled by adding weight decay in Algorithm 1. On the other hand, we can solve the unconstrained version of the problem by simply choosing a large enough radius $\mathcal{R}$. Note that the presence of this constraint is often desirable: it can be seen as a regularization, which may encourage dense (Chen et al., 2023a) or high-rank (Chen et al., 2025) solution in the case of sign-based or orthogonalized methods, respectively.

Convergence criterion. We define the function $\mathcal{G}_{\mathcal{R}}(\cdot) \colon Q \to \mathbb{R}_{\geq}$, which we are going to use as a convergence criterion, as follows:

$$\mathcal{G}_{\mathcal{R}}(x) = \begin{cases} \frac{1}{\mathcal{R}} \max_{\rho(x') \leq \mathcal{R}} \langle \nabla f(x), x - x' \rangle & \mathcal{R} \in (0, +\infty) \\ \rho_*(\nabla f(x)) & \mathcal{R} = +\infty \end{cases}. \tag{9}$$

In the case $\mathcal{R} = +\infty$, the function $\mathcal{G}_{\mathcal{R}}(\cdot)$ is simply the dual gradient norm $\rho_*(\nabla f(\cdot))$, which is the standard convergence criterion for the unconstrained non-convex setting. In the constrained case, $\mathcal{R} < +\infty$, the function $\mathcal{G}_{\mathcal{R}}(\cdot)$ coincides with the Frank-Wolfe gap (Jaggi, 2013) up to the scaling factor $1/\mathcal{R}$. Note that the scaling factor $1/\mathcal{R}$ ensures that the function $\mathcal{R} \mapsto \mathcal{G}_{\mathcal{R}}(x)$ is continuous at $\mathcal{R} = +\infty$ for all $x \in \mathcal{X}$. In addition, $\mathcal{G}_{\mathcal{R}}(\cdot)$ coincides with the convergence criterion used by Dong et al. (2024); Chen et al. (2025). For completeness, in Lemma 3, we verify that the function $\mathcal{G}_{\mathcal{R}}(\cdot)$ is indeed a meaningful convergence criterion in the constrained case: null points of $\mathcal{G}_{\mathcal{R}}(\cdot)$ are the first-order stationary points for problem (8). Additionally, in the constrained convex case, $\mathcal{G}_{\mathcal{R}}(\cdot)$ can be used to upper-bound the objective function optimality gap, as stated in Lemma 4.

Lemma 3 (↓). *$\mathcal{G}_{\mathcal{R}}(x^*) = 0$ if and only if $x^* \in Q$ is a first-order stationary point in problem* (8)*, that is, $-\nabla f(x^*) \in \mathcal{N}_Q(x^*)$, where $\mathcal{N}_Q(x) = \{y \in \mathcal{X} : \langle y, x - x' \rangle \geq 0$ for all $x' \in Q\}$ is the normal cone to the set Q at the point $x \in Q$.*

Lemma 4 (↓). *Let $\mathcal{R} < +\infty$ and let the objective function $f(x)$ be convex. Then the following inequality holds for all $x \in Q$:*

$$\mathcal{R} \cdot \mathcal{G}_{\mathcal{R}}(x) \geq f(x) - f^*. \tag{10}$$

3 Non-Euclidean SGD with Momentum and Weight Decay

3.1 Non-Euclidean Trust-Region Gradient Step

In this section, we develop the convergence analysis for the non-Euclidean SGD with momentum and weight decay, which is formalized as Algorithm 1. We start with the non-Euclidean trust-region gradient step with the gradient estimator m_k in eq. (11). It was analyzed in details for general non-Euclidean norms Equation (1) by Kovalev (2025b). Further, we develop the analysis under the structural assumptions in Section 2.2. We start with the basic properties in Lemma 5.

Algorithm 1 Non-Euclidean SGD with Momentum and Weight Decay

1: **input:** $x_0 \in Q$
2: **parameters:** momentum $\alpha \in (0, 1]$, weight decay $\beta \in [0, 1]$, stepsize $\eta > 0$
3: sample $\xi_0 \sim \mathcal{D}$ and compute $m_0 = g(x_0; \xi_0)$
4: **for** $k = 0, 1, \ldots, K-1$ **do**
5: compute x_{k+1} as follows:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \langle m_k, x - x_k \rangle \quad \text{s.t.} \quad \rho(x - (1-\beta)x_k) \leq \eta \tag{11}$$

6: sample $\xi_{k+1} \sim \mathcal{D}$ and compute m_{k+1} using one of the following options:

$$m_{k+1} = (1-\alpha)m_k + \alpha g(x_{k+1}, \xi_{k+1}) \tag{12a}$$
$$m_{k+1} = (1-\alpha)m_k + \alpha g(\overline{x}_{k+1}, \xi_{k+1}) \tag{12b}$$
$$m_{k+1} = (1-\alpha)(m_k - g(x_k, \xi_{k+1})) + g(x_{k+1}; \xi_{k+1}) \tag{12c}$$

where $\overline{x}_{k+1} = x_k + (1/\alpha)(x_{k+1} - x_k)$.
7: **output:** $x_K \in \mathcal{X}$

Lemma 5 (↓). *Let Assumption 5 hold and let $\beta = \eta/\mathcal{R}$. Then, $\beta\rho(x_k) \leq \eta$ and $\rho(x_{k+1} - x_k) \leq 2\eta$.*

Next, we define the operator $\mathbf{B} \in \mathbb{S}_\succ$ as follows:

$$\mathbf{B} = \tfrac{1}{4}\|\mathbf{L}\|_{\mathrm{tr}}^{-1}\mathbf{L} + \tfrac{1}{4}\|\mathbf{M}\|_{\mathrm{tr}}^{-1}\mathbf{M} + \tfrac{1}{4}\|\mathbf{\Sigma}\|_{\mathrm{tr}}^{-1}\mathbf{\Sigma} + \tfrac{1}{4}\|\mathbf{T}\|_{\mathrm{tr}}^{-1}\mathbf{T}. \tag{13}$$

It is easy to verify that $\mathbf{B} \in \mathcal{H}$ due to Assumption 5, and that $\|\mathbf{B}\|_{\mathrm{tr}} = 1$. We also provide useful bounds on the operator $\mathbf{B}$ in Lemma 6 and the main result for the trust-region step in Lemma 7.

Lemma 6 (↓). *Let Assumption 5 hold. Then for $\mathbf{H} \in \{\mathbf{L}, \mathbf{M}, \mathbf{\Sigma}, \mathbf{T}\}$, the following inequalities hold:*

$$\mathbf{H} \preceq 4\|\mathbf{H}\|_{\mathrm{tr}} \cdot \mathbf{B}, \quad 4\|\mathbf{H}\|_{\mathrm{tr}} \cdot \mathbf{H}^{-1} \succeq \mathbf{B}^{-1} \tag{14}$$

Lemma 7 (↓). *Let Assumptions 1 and 5 hold and let $\beta = \eta/\mathcal{R}$. Then the following inequality holds, where we use the convention $1/0 = +\infty$:*

$$\eta\mathcal{G}_{\eta/\beta}(x_k) \leq f(x_k) - f(x_{k+1}) + 4\eta\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2. \tag{15}$$

Connection with adaptive methods. As previously mentioned, Xie et al. (2025); Kovalev (2025a) developed a unified analysis of the adaptive meta-algorithm. More concretely, they analyzed the preconditioned SGD of the form $x_{k+1} = x_k - \mathbf{H}_k g_k$, where $g_k \in \mathcal{X}$ is the current gradient estimator. Under Assumption 5, the preconditioning operator can be expressed as $\mathbf{H}_k = \eta[\delta\mathbf{I} + \mathrm{proj}_{\mathcal{H}}(\mathbf{S}_k)]^{\dagger/2}$, where $\eta, \delta \geq 0$ and $\mathbf{S}_k = \sum_{i=0}^k g_i\langle g_i, \cdot\rangle$ is the cumulative sum of gradient outer squares (Kovalev, 2025a; 2026). One can verify that when $\delta = 0$ and the gradient accumulation is turned off, i.e., $\mathbf{S}_k = g_k\langle g_k, \cdot\rangle$, the preconditioned SGD step above is equivalent to the trust-region step in eq. (11) with $m_k = g_k$ and $\beta = 0$. Hence, the adaptive meta-algorithm can be seen as an adaptive counterpart to Algorithm 1. Consequently, AdaGrad and ASGO/One-sided Shampoo are adaptive "twins" to SignSGD and Muon, as indicated in Table 1.

3.2 Main Convergence Results

In this section, we obtain the main convergence results for Algorithm 1, which are summarized in Table 2. First, in the following Lemmas 8, 9 and 10, we analyze three options for computing the gradient estimator m_{k+1} on line 6 using momentum, momentum with extrapolation, and momentum variance reduction in eqs. (12a), (12b) and (12c), respectively.

Lemma 8 (↓). *Let Assumptions 1, 2 and 5 hold and let $\beta = \eta/\mathcal{R}$. Let m_{k+1} be defined using eq.* (12a) *at each iteration of Algorithm 1. Then the following inequality holds:*

$$\mathbb{E}[\|m_{k+1} - \nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2] \leq (1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|_{\mathbf{B}^{-1}}^2] + 16\|\mathbf{L}\|_{\mathrm{tr}}^2\eta^2/\alpha + 4\alpha^2\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2. \tag{16}$$

Lemma 9 (↓). *Let Assumptions 2, 3 and 5 hold and let $\beta = \eta/\mathcal{R}$. Let m_{k+1} be defined using eq.* (12b) *at each iteration of Algorithm 1. Then the following inequality holds:*

$$\mathbb{E}[\|m_{k+1} - \nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2] \leq (1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|_{\mathbf{B}^{-1}}^2] + 48\|\mathbf{T}\|_{\mathrm{tr}}^2\eta^4/\alpha^3 + 4\alpha^2\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2. \tag{17}$$

Table 2: Summary of the results in Section 3. The number of iterations K of Algorithm 1 required to ensure $\mathbb{E}[\min_{k\in\{0,\dots,K-1\}} \mathcal{G}_{\mathcal{R}}(x_k)] \leq \epsilon$ in the non-convex case and $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$ in the convex case is given. Only the leading terms (with the worst dependence on ϵ) are reported. Universal and logarithmic factors are omitted. Extra assumptions beyond Assumptions 1, 2 and 5 are provided.

Reference	Non-Convex Case	Convex Case	Extra Assumption
Theorem 1	$\frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \|\mathbf{L}\|_{\mathrm{tr}} \Delta_0}{\epsilon^4}$	$\frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \|\mathbf{L}\|_{\mathrm{tr}} \mathcal{R}^4}{\epsilon^3}$	—
Theorem 2	$\frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \|\mathbf{T}\|_{\mathrm{tr}}^{0.5} \Delta_0}{\epsilon^{3.5}}$	$\frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \|\mathbf{T}\|_{\mathrm{tr}}^{0.5} \mathcal{R}^{3.5}}{\epsilon^{2.5}}$	Assumption 3
Theorem 3	$\frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^3 + \|\mathbf{\Sigma}\|_{\mathrm{tr}} \|\mathbf{M}\|_{\mathrm{tr}} \Delta_0}{\epsilon^3}$	$\frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \mathcal{R}^2 + \|\mathbf{\Sigma}\|_{\mathrm{tr}} \|\mathbf{M}\|_{\mathrm{tr}} \mathcal{R}^3}{\epsilon^2}$	Assumption 4

Lemma 10 (↓). *Let Assumptions 2, 4 and 5 hold and let $\beta = \eta/\mathcal{R}$. Let m_{k+1} be defined using eq.* (12c) *at each iteration of Algorithm 1. Then the following inequality holds:*

$$\mathbb{E}[\|m_{k+1} - \nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2] \leq (1-\alpha)^2 \mathbb{E}[\|m_k - \nabla f(x_k)\|_{\mathbf{B}^{-1}}^2] + 32\|\mathbf{M}\|_{\mathrm{tr}}^2 \eta^2 + 8\alpha^2 \|\mathbf{\Sigma}\|_{\mathrm{tr}}^2. \quad (18)$$

Using the above lemmas, we obtain the following Corollaries 1, 2 and 3

Corollary 1 (↓). *Under the conditions of Lemma 8, the following inequality holds:*

$$\mathbb{E}[\|m_k - \nabla f(x_k)\|_{\mathbf{B}^{-1}}] \leq 2\big[(1-\alpha/2)^k + \sqrt{\alpha}\big] \|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4(\eta/\alpha)\|\mathbf{L}\|_{\mathrm{tr}}. \quad (19)$$

Corollary 2 (↓). *Under the conditions of Lemma 9, the following inequality holds:*

$$\mathbb{E}[\|m_k - \nabla f(x_k)\|_{\mathbf{B}^{-1}}] \leq 2\big[(1-\alpha/2)^k + \sqrt{\alpha}\big] \|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4\sqrt{3}(\eta/\alpha)^2 \|\mathbf{T}\|_{\mathrm{tr}}. \quad (20)$$

Corollary 3 (↓). *Under the conditions of Lemma 10, the following inequality holds:*

$$\mathbb{E}[\|m_k - \nabla f(x_k)\|_{\mathbf{B}^{-1}}] \leq 2[(1-\alpha)^k + \sqrt{2\alpha}] \|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4\sqrt{2/\alpha} \|\mathbf{M}\|_{\mathrm{tr}} \eta. \quad (21)$$

Now, we are ready to establish the convergence rates for Algorithm 1 with momentum, momentum with extrapolation, and momentum variance reduction in the following Theorems 1, 2 and 3, respectively. In particular, we get the number of iterations K that are required to reach the precision $\mathbb{E}\big[\min_{k\in\{0,\dots,K-1\}} \mathcal{G}_{\mathcal{R}}(x_k)\big] \leq \epsilon$ in the non-convex case, and the precision $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$ in the convex case. Note that we also cover the non-convex unconstrained case, where $\mathcal{R} = +\infty$. Therefore, Theorems 1, 2 and 3 treat the constrained and unconstrained settings in a unified manner.

Theorem 1 (↓). *Let Assumptions 1, 2 and 5 hold. Let m_{k+1} be defined using eq.* (12a) *at each iteration of Algorithm 1. Then, choosing the parameters of Algorithm 1 as*

$$\alpha = \min\Big\{1, \frac{c_1 \cdot \epsilon^2}{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2}\Big\}, \quad \beta = \min\Big\{1, \frac{c_2 \cdot \epsilon\alpha}{\|\mathbf{L}\|_{\mathrm{tr}} \mathcal{R}}\Big\}, \quad \eta = \beta\mathcal{R}, \quad (22)$$

where $c_1, c_2 > 0$ are universal constants,[3] *and the number of iterations $K \in \mathbb{N}$ of Algorithm 1 as*

$$K = \mathcal{O}\Big(\max\Big\{1, \frac{\Delta_0}{\epsilon\mathcal{R}}, \frac{\|\mathbf{L}\|_{\mathrm{tr}} \Delta_0}{\epsilon^2}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^3}{\epsilon^3}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \|\mathbf{L}\|_{\mathrm{tr}} \Delta_0}{\epsilon^4}\Big\}\Big), \quad (23)$$

where $\Delta_0 = f(x_0) - f^*$, implies $\mathbb{E}\big[\min_{k\in\{0,\dots,K-1\}} \mathcal{G}_{\mathcal{R}}(x_k)\big] \leq \epsilon$. Furthermore, if the function $f(x)$ is assumed to be convex and $\mathcal{R} < +\infty$, then choosing the parameters of Algorithm 1 as

$$\alpha = \min\Big\{1, \frac{c_1 \cdot \epsilon^2}{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \mathcal{R}^2}\Big\}, \quad \beta = \min\Big\{\alpha, \frac{c_2 \cdot \epsilon\alpha}{\|\mathbf{L}\|_{\mathrm{tr}} \mathcal{R}^2}\Big\}, \quad \eta = \beta\mathcal{R}, \quad (24)$$

where $c_1, c_2 > 0$ are universal constants, and the number of iterations $K \in \mathbb{N}$ of Algorithm 1 as

$$K = \tilde{\mathcal{O}}\Big(\max\Big\{1, \frac{\|\mathbf{L}\|_{\mathrm{tr}} \mathcal{R}^2}{\epsilon}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \mathcal{R}^2}{\epsilon^2}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 \|\mathbf{L}\|_{\mathrm{tr}} \mathcal{R}^4}{\epsilon^3}\Big\}\Big). \quad (25)$$

implies $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$.

[3] In the case $\mathcal{R} = +\infty$, we define η by continuous extension.

Table 3: Comparison of the iterations complexities in the convex case for non-Euclidean SGD with momentum variance reduction (Algorithm 1, Theorem 3) with the results for adaptive meta-algorithm (Kovalev, 2025a; Xie et al., 2025). Universal constant and logarithmic factors are omitted.

Method	Algorithm 1 (Option 12c)	Adaptive Meta-Algorithm
Complexity	$\frac{\|\mathbf{M}\|_{\mathrm{tr}}\mathcal{R}^2}{\epsilon} + \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\mathcal{R}^2}{\epsilon^2} + \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}\|\mathbf{M}\|_{\mathrm{tr}}\mathcal{R}^3}{\epsilon^2}$	$\frac{\|\mathbf{L}\|_{\mathrm{tr}}\mathcal{R}^2}{\epsilon} + \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\mathcal{R}^2}{\epsilon^2}$
Assumptions	2, 4 and 5	1, 2 and 5

Theorem 2 (↓). *Let Assumptions 1, 2, 3 and 5 hold. Let m_{k+1} be defined using eq.* (12b) *at each iteration of Algorithm 1. Then, choosing the parameters of Algorithm 1 as*

$$\alpha = \min\Big\{1, \frac{c_1 \cdot \epsilon^2}{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2}\Big\}, \quad \beta = \min\Big\{1, \frac{c_2 \cdot \epsilon}{\|\mathbf{L}\|_{\mathrm{tr}}\mathcal{R}}, \frac{c_3 \cdot \sqrt{\epsilon}\alpha}{\sqrt{\|\mathbf{T}\|_{\mathrm{tr}}}\mathcal{R}}\Big\}, \quad \eta = \beta\mathcal{R}. \tag{26}$$

where $c_1, c_2, c_3 > 0$ are universal constants,[3] *and the number of iterations $K \in \mathbb{N}$ of Algorithm 1 as*

$$K = \mathcal{O}\Big(\max\Big\{1, \frac{\Delta_0}{\epsilon\mathcal{R}}, \frac{\|\mathbf{T}\|_{\mathrm{tr}}^{0.5}\Delta_0}{\epsilon^{1.5}}, \frac{\|\mathbf{L}\|_{\mathrm{tr}}\Delta_0}{\epsilon^2}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^3}{\epsilon^3}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\|\mathbf{T}\|_{\mathrm{tr}}^{0.5}\Delta_0}{\epsilon^{3.5}}\Big\}\Big), \tag{27}$$

where $\Delta_0 = f(x_0) - f^*$, implies $\mathbb{E}\big[\min_{k\in\{0,\ldots,K-1\}} \mathcal{G}_{\mathcal{R}}(x_k)\big] \leq \epsilon$. Furthermore, if the function $f(x)$ is assumed to be convex and $\mathcal{R} < +\infty$, then choosing the parameters of Algorithm 1 as

$$\alpha = \min\Big\{1, \frac{c_1 \cdot \epsilon^2}{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\mathcal{R}^2}\Big\}, \quad \beta = \min\Big\{\alpha, \frac{c_2 \cdot \epsilon}{\|\mathbf{L}\|_{\mathrm{tr}}\mathcal{R}^2}, \frac{c_3 \cdot \sqrt{\epsilon}\alpha}{\sqrt{\|\mathbf{T}\|_{\mathrm{tr}}\mathcal{R}^3}}\Big\}, \quad \eta = \beta\mathcal{R}. \tag{28}$$

where $c_1, c_2, c_3 > 0$ are universal constants, and the number of iterations $K \in \mathbb{N}$ of Algorithm 1 as

$$K = \tilde{\mathcal{O}}\Big(\max\Big\{1, \frac{\|\mathbf{T}\|_{\mathrm{tr}}^{0.5}\mathcal{R}^{1.5}}{\epsilon^{0.5}}, \frac{\|\mathbf{L}\|_{\mathrm{tr}}\mathcal{R}^2}{\epsilon}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\mathcal{R}^2}{\epsilon^2}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\|\mathbf{T}\|_{\mathrm{tr}}^{0.5}\mathcal{R}^{3.5}}{\epsilon^{2.5}}\Big\}\Big). \tag{29}$$

implies $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$.

Theorem 3 (↓). *Let Assumptions 2, 4 and 5 hold. Let m_{k+1} be defined using eq.* (12c) *at each iteration of Algorithm 1. Then, choosing the parameters of Algorithm 1 as*

$$\alpha = \min\Big\{1, \frac{c_1 \cdot \epsilon^2}{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2}\Big\}, \quad \beta = \min\Big\{1, \frac{c_2 \cdot \epsilon\sqrt{\alpha}}{\|\mathbf{M}\|_{\mathrm{tr}}\mathcal{R}}\Big\}, \quad \eta = \beta\mathcal{R}, \tag{30}$$

where $c_1, c_2 > 0$ are universal constants,[3] *and the number of iterations $K \in \mathbb{N}$ of Algorithm 1 as*

$$K = \mathcal{O}\Big(\max\Big\{1, \frac{\Delta_0}{\epsilon\mathcal{R}}, \frac{\|\mathbf{M}\|_{\mathrm{tr}}\Delta_0}{\epsilon^2}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^3}{\epsilon^3}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}\|\mathbf{M}\|_{\mathrm{tr}}\Delta_0}{\epsilon^3}\Big\}\Big), \tag{31}$$

where $\Delta_0 = f(x_0) - f^*$, implies $\mathbb{E}\big[\min_{k\in\{0,\ldots,K-1\}} \mathcal{G}_{\mathcal{R}}(x_k)\big] \leq \epsilon$. Furthermore, if the function $f(x)$ is assumed to be convex and $\mathcal{R} < +\infty$, then choosing the parameters of Algorithm 1 as

$$\alpha = \min\Big\{1, \frac{c_1 \cdot \epsilon^2}{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\mathcal{R}^2}\Big\}, \quad \beta = \min\Big\{\alpha, \frac{c_2 \cdot \epsilon\sqrt{\alpha}}{\|\mathbf{M}\|_{\mathrm{tr}}\mathcal{R}^2}\Big\}, \quad \eta = \beta\mathcal{R}, \tag{32}$$

where $c_1, c_2 > 0$ are universal constants, and the number of iterations $K \in \mathbb{N}$ of Algorithm 1 as

$$K = \tilde{\mathcal{O}}\Big(\max\Big\{1, \frac{\|\mathbf{M}\|_{\mathrm{tr}}\mathcal{R}^2}{\epsilon}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\mathcal{R}^2}{\epsilon^2}, \frac{\|\mathbf{\Sigma}\|_{\mathrm{tr}}\|\mathbf{M}\|_{\mathrm{tr}}\mathcal{R}^3}{\epsilon^2}\Big\}\Big). \tag{33}$$

implies $\mathbb{E}[f(x_K) - f^*] \leq \epsilon$.

3.3 Discussion

First, we compare the convergence results for non-Euclidean SGD with momentum variance reduction (Algorithm 1, Option 12c) in Theorem 3 with the convergence results for the adaptive meta-algorithm by Kovalev (2025a, Theorems 1 and 3) in the convex setting. This comparison is summarized in Table 3. Theorem 3 requires Assumption 4 in place of Assumption 1. Additionally, both rates

have optimal dependence on ϵ (Nemirovskij & Yudin, 1983; Lan, 2012) and match as long as $\|\mathbf{M}\|_{\rm tr}\mathcal{R} \lesssim \|\mathbf{\Sigma}\|_{\rm tr}$ and $\|\mathbf{L}\|_{\rm tr} \approx \|\mathbf{M}\|_{\rm tr}$.

Next, we show that Algorithm 1 can exploit the structural Assumption 5. In particular, we show that sign-based and orthogonalized variants of Algorithm 1 can provably outperform the Euclidean variant of Algorithm 1. For simplicity, we consider Option 12a, although a similar comparison clearly applies to Options 12b and 12c as well.

SignSGD vs Normalized SGD. In the case of sign-based methods, the properties of the objective function are measured in terms of the weighted Euclidean norms of the form $\|x\|_{\boldsymbol{h}}^2 = \langle x, \boldsymbol{h} \odot x\rangle$, where $\boldsymbol{h} \in \mathbb{R}^d_+$ and $x \in \mathcal{X} = \mathbb{R}^d$. Consequently, Assumptions 1 and 2 reduce to

$$|\mathrm{D}_f(x;x')| \leq \tfrac{1}{2}\|x-x'\|^2_{\mathrm{diag}(\boldsymbol{l})}, \qquad \mathbb{E}_{\xi\sim\mathcal{D}}[[n_\xi(x)]^2_{[i]}] \leq [\boldsymbol{\sigma}]^2_{[i]} \text{ for all } i, \tag{34}$$

where $\boldsymbol{l}, \boldsymbol{\sigma} \in \mathbb{R}^d_+$. Note that, for simplicity and for illustrative purposes, we use here a slightly more restrictive variant of Assumption 2, namely, the coordinate-wise bound on the variance. Under these assumptions, the convergence rates in Table 2 (Theorem 1, convex case) reduce to

$$\text{SignSGD:} \quad \frac{\|\boldsymbol{\sigma}\|_1^2\|\boldsymbol{l}\|_1\|x^*\|_\infty^4}{\epsilon^3}, \qquad \text{Normalized SGD:} \quad \frac{\|\boldsymbol{\sigma}\|_2^2\|\boldsymbol{l}\|_\infty\|x^*\|_2^4}{\epsilon^3}. \tag{35}$$

We can conclude that the complexity of SignSGD improves upon the complexity of Normalized SGD, when the solution $x^* \in \mathbb{R}^d$ is dense, i.e., $\|x^*\|_2 \sim \sqrt{d}\cdot\|x^*\|_\infty$, and when the coordinates in the objective function have very different scales, i.e., the vectors $\boldsymbol{l}, \boldsymbol{\sigma}$ have a few dominant components that are much greater than the rest, which implies $\|\boldsymbol{l}\|_1 \ll d\cdot\|\boldsymbol{l}\|_\infty$, $\|\boldsymbol{\sigma}\|_1 \ll \sqrt{d}\cdot\|\boldsymbol{\sigma}\|_2$. This observation aligns with and complements the existing theory for sign-based methods (Liu et al., 2024b; Jiang et al., 2024a).

Muon vs Normalized SGD. In the case of orthogonalized methods, the properties of the objective function are measured in terms of the weighted Euclidean norms of the form $\|X\|_H^2 = \langle X, HX\rangle$, where $H \in \mathbb{S}^m_\succ$ and $X \in \mathcal{X} = \mathbb{R}^{m\times n}$. Assumptions 1 and 2 reduce to the following:

$$|\mathrm{D}_f(X;X')| \leq \tfrac{1}{2}\|X-X'\|_L^2, \qquad \mathbb{E}_{\xi\sim\mathcal{D}}[N_\xi(X)N_\xi(X)^\top] \preceq \Sigma^2, \tag{36}$$

where $L, \Sigma \in \mathbb{S}^m_\succ$, and $N_\xi(X) \in \mathbb{R}^{m\times n}$ is the gradient noise reshaped into an $m \times n$ matrix. Note that, for simplicity and for illustrative purposes, we use here a slightly more restrictive variant of Assumption 2, namely, the matrix upper-bound on the variance. Under these assumptions, the convergence rates in Table 2 (Theorem 1, convex case) reduce to

$$\text{Muon:} \quad \frac{\|\Sigma\|_{\rm tr}^2\|L\|_{\rm tr}\|X^*\|_{\rm op}^4}{\epsilon^3}, \qquad \text{Normalized SGD:} \quad \frac{\|\Sigma\|_F^2\|L\|_{\rm op}\|X^*\|_F^4}{\epsilon^3}. \tag{37}$$

We can conclude that the complexity of Muon improves upon the complexity of Normalized SGD, when the solution $X^* \in \mathbb{R}^{m\times n}$ is high-rank, i.e., $\|X^*\|_F \sim \sqrt{r}\cdot\|X^*\|_{\rm op}$, where $r = \min\{m,n\}$, and when the matrices L, Σ are approximately low-rank, i.e., $\|L\|_{\rm tr} \ll r\cdot\|L\|_{\rm op}$, $\|\Sigma\|_{\rm tr} \ll \sqrt{r}\cdot\|\Sigma\|_F$. This observation aligns with and complements the existing theory for orthogonalized/matrix-based optimizers (An et al., 2025).

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including grammar correction, removal of repetition, and rewording to improve clarity and readability. AI-assisted editing was limited to text segments no longer than a single paragraph at a time. All suggested edits were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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.

Amit Attia and Tomer Koren. Sgd with adagrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, pp. 1147–1171. PMLR, 2023.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

Ekaterina Borodich and Dmitry Kovalev. Nesterov finds graal: Optimal and adaptive gradient method for convex optimization. *arXiv preprint arXiv:2507.09823*, 2025.

Eric Carlen. Trace inequalities and quantum entropy: an introductory course. *Entropy and the quantum*, 529(73-140):146, 2010.

David E Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. *Advances in neural information processing systems*, 28, 2015.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves constrained optimization: As lyapunov predicts. *arXiv preprint arXiv:2310.05898*, 2023a.

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

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023b.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pp. 5396–5427. PMLR, 2023c.

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.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pp. 2260–2268. PMLR, 2020.

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

Yiming Dong, Huan Li, and Zhouchen Lin. Convergence rate analysis of lion. *arXiv preprint arXiv:2411.07724*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

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

Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global kurdyka-lojasiewicz inequality. *Advances in Neural Information Processing Systems*, 35:15836–15848, 2022.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

Kaja Gruntkowska, Yassine Maziane, Zheng Qu, and Peter Richtárik. Drop-muon: Update less, converge faster. *arXiv preprint arXiv:2510.02239*, 2025.

Vineet Gupta, Tomer Koren, and Yoram Singer. A unified approach to adaptive regularization in online and stochastic optimization. *arXiv preprint arXiv:1706.06569*, 2017.

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

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

Chuan He, Shuyi Ren, Jingwei Mao, and Erik G Larsson. Demuon: A decentralized muon for matrix optimization over graphs. *arXiv preprint arXiv:2510.01377*, 2025b.

Feihu Huang, Yuning Luo, and Songcan Chen. Limuon: Light and fast muon optimizer for large models. *arXiv preprint arXiv:2509.14562*, 2025.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pp. 427–435. PMLR, 2013.

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

Wei Jiang, Sifan Yang, Wenhao Yang, and Lijun Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. *Advances in Neural Information Processing Systems*, 37: 33891–33932, 2024b.

Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, and Lijun Zhang. Improved analysis for sign-based methods with momentum updates. *arXiv preprint arXiv:2507.12091*, 2025.

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

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

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

Dmitry Kovalev. Sgd with adaptive preconditioning: Unified analysis and momentum acceleration. *arXiv preprint arXiv:2506.23803*, 2025a.

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

Dmitry Kovalev. Optimal projection-free adaptive sgd for matrix optimization. *arXiv preprint arXiv:2604.02505*, 2026.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

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

Yifeng Liu, Angela Yuan, and Quanquan Gu. Mars-m: When variance reduction meets matrices. *arXiv preprint arXiv:2510.21800*, 2025b.

Yuxing Liu, Rui Pan, and Tong Zhang. Adagrad under anisotropic smoothness. *arXiv preprint arXiv:2406.15244*, 2024b.

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

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

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.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making muon & scion great again!(bridging theory and practice of lmo-based optimizers for llms). *arXiv preprint arXiv:2505.13416*, 2025.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Andrey Sadchikov, Savelii Chezhegov, Aleksandr Beznosikov, and Alexander Gasnikov. Local sgd for near-quadratic problems: Improving convergence under unconstrained noise conditions. *arXiv preprint arXiv:2409.10478*, 2024.

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

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

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact muon update. *arXiv preprint arXiv:2510.19933*, 2025.

Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signsgd under weaker assumptions. In *International Conference on Machine Learning*, pp. 33077–33099. PMLR, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. *arXiv preprint arXiv:2410.10800*, 2024.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019.

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.

Shuo Xie, Tianhao Wang, Sashank Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. *arXiv preprint arXiv:2503.10537*, 2025.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

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

Xinwen Zhang and Hongchang Gao. On provable benefits of muon in federated learning. *arXiv preprint arXiv:2510.03866*, 2025.

Appendix

A Non-Euclidean vs Euclidean SGD: Existing Results

In this section, we summarize existing convergence results for non-Euclidean SGD and highlight that these results cannot surpass the best guarantees for Euclidean SGD.

Sign-based methods. Bernstein et al. (2018) analyzed SignSGD without momentum under assumptions similar to those described in Section 3.3. Their convergence guarantee requires the batch size to grow with the number of iterations, which is restrictive and rarely used in practice. Sun et al. (2023) and Jiang et al. (2024b) considered SignSGD with extrapolation and momentum variance reduction, respectively. While their algorithms can be seen as special instances of Algorithm 1 with options 12b and 12c, their convergence rates contain extra dimension factors and are worse than those for Euclidean SGD methods. Additionally, Jiang et al. (2025) established dimension-independent convergence rates for SignSGD with momentum. However, their result is based on ordinary momentum, whose state-of-the-art convergence rate in the convex setting has worse dependence on the precision ϵ than vanilla SGD. In contrast, we show that SignSGD with momentum variance reduction achieves optimal dependence on the precision ϵ and can show an overall improvement over the convergence rate of Euclidean SGD by exploiting the problem structure.

Orthogonalized methods. There are several works that show how convergence guarantees for orthogonalized gradient methods in the deterministic setting can outperform corresponding guarantees for Euclidean methods. However, to the best of our knowledge, all such results fall short in the stochastic setting, because they contain extra dimension factors in the complexities (e.g., Pethick et al. (2025, Appendix D), Kovalev (2025b, Corollary 2), Shen et al. (2025, Theorem 4.3)) and are worse than those for Euclidean SGD methods. In particular, to the best of our knowledge, Theorem 3 is the first result showing that Muon-type algorithms can achieve better convergence guarantees than vanilla SGD in the convex setting.

B PROOFS FROM SECTION 2.2

B.1 PROOF OF LEMMA 1

Part I. From Assumption 5 it follows that $\mathbf{H}^k \in \mathcal{H}$ for all $\mathbf{H} \in \mathcal{H}$, $k \in \mathbb{N}_0$. Based on this fact, using polynomial interpolation, it is easy to show that $\mathcal{H}$ is closed under arbitrary operator functions, that is, $\psi(\mathbf{H}) \in \mathcal{H}$ for all $\mathbf{H} \in \mathcal{H}$ and $\psi(\cdot)\colon \mathbb{R} \to \mathbb{R}$. Hence, we can use Lemma 4 of Kovalev (2025a) to prove that the function $\rho(\cdot)$ is a norm.

Part II. We start by proving the following Fenchel-Young inequality for all $x, y \in \mathcal{X}$:

$$\langle x, y\rangle \leq \tfrac{1}{2}\rho^2(x) + \tfrac{1}{2}\rho_*^2(y). \tag{38}$$

Indeed, we can lower-bound $\frac{1}{2}\rho^2(x)$ as follows:

$$\begin{aligned}
\tfrac{1}{2}\rho^2(x) &\overset{\text{(a)}}{=} \langle x, y\rangle + \sup_{\|\mathbf{H}\|_{\text{tr}}\leq 1, \mathbf{H}\in\mathbb{S}_\succ} \left[\tfrac{1}{2}\langle \text{proj}_{\mathcal{H}}(\mathbf{X}), \mathbf{H}\rangle - \langle x, y\rangle\right] \\
&\geq \langle x, y\rangle + \inf_{x\in\mathcal{X}} \sup_{\|\mathbf{H}\|_{\text{tr}}\leq 1, \mathbf{H}\in\mathbb{S}_\succ} \left[\tfrac{1}{2}\langle \text{proj}_{\mathcal{H}}(\mathbf{X}), \mathbf{H}\rangle - \langle x, y\rangle\right] \\
&\overset{\text{(b)}}{\geq} \langle x, y\rangle + \sup_{\|\mathbf{H}\|_{\text{tr}}\leq 1, \mathbf{H}\in\mathbb{S}_\succ} \inf_{x\in\mathcal{X}} \left[\tfrac{1}{2}\langle \text{proj}_{\mathcal{H}}(\mathbf{X}), \mathbf{H}\rangle - \langle x, y\rangle\right] \\
&\overset{\text{(c)}}{\geq} \langle x, y\rangle + \sup_{\|\mathbf{H}\|_{\text{tr}}\leq 1, \mathbf{H}\in\mathbb{S}_\succ} \inf_{x\in\mathcal{X}} \left[\tfrac{1}{2}\|x\|^2_{\text{proj}_{\mathcal{H}}(\mathbf{H})} - \langle x, y\rangle\right] \\
&\overset{\text{(d)}}{=} \langle x, y\rangle - \inf_{\|\mathbf{H}\|_{\text{tr}}\leq 1, \mathbf{H}\in\mathbb{S}_\succ} \tfrac{1}{2}\|y\|^2_{\text{proj}_{\mathcal{H}}(\mathbf{H})^{-1}},
\end{aligned}$$

where (a) uses the definition of $\rho(\cdot)$ in eq. (1) and the properties of $\|\cdot\|_{\text{op}}$; (b) uses the fact that $\inf_x \sup_{\mathbf{H}}[\ldots] \geq \sup_{\mathbf{H}} \inf_x[\ldots]$; (c) uses the properties of the projection and the definition of $\mathbf{X}$ in eq. (1); (d) uses the fact that $\text{proj}_{\mathcal{H}}(\mathbf{H}) \in \mathbb{S}_\succ$ due to Assumption 5. After rearranging, we obtain the following inequality for all $\mathbf{H} \in \mathbb{S}_\succ$ such that $\|\mathbf{H}\|_{\text{tr}} \leq 1$:

$$\langle x, y\rangle \leq \tfrac{1}{2}\rho^2(x) + \tfrac{1}{2}\|y\|^2_{\text{proj}_{\mathcal{H}}(\mathbf{H})^{-1}}. \tag{39}$$

Let $\delta > 0$ and let $\mathbf{H} \in \mathbb{S}$ be defined as follows:

$$\mathbf{H} = \|\sqrt{\delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y})}\|_{\text{tr}}^{-1} \sqrt{\delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y})}, \quad \text{where} \quad \mathbf{Y} = y\langle y, \cdot\rangle. \tag{40}$$

Using Assumption 5 and the fact that $\mathcal{H}$ is closed under operator functions from Part I, we can show that $\mathbf{H}, \mathbf{H}^{-1} \in \mathbb{S}_\succ \cap \mathcal{H}$. Hence, we can upper-bound $\langle x, y\rangle$ as follows:

$$\begin{aligned}
\langle x, y\rangle &\overset{\text{(a)}}{\leq} \tfrac{1}{2}\rho^2(x) + \tfrac{1}{2}\|y\|^2_{\text{proj}_{\mathcal{H}}(\mathbf{H})^{-1}} \\
&\overset{\text{(b)}}{=} \tfrac{1}{2}\rho^2(x) + \tfrac{1}{2}\|y\|^2_{\mathbf{H}^{-1}} \\
&\overset{\text{(c)}}{\leq} \tfrac{1}{2}\rho^2(x) + \tfrac{1}{2}\langle \delta\mathbf{I} + \mathbf{Y}, \mathbf{H}^{-1}\rangle \\
&\overset{\text{(d)}}{=} \tfrac{1}{2}\rho^2(x) + \tfrac{1}{2}\langle \delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y}), \mathbf{H}^{-1}\rangle \\
&\overset{\text{(e)}}{=} \tfrac{1}{2}\rho^2(x) + \tfrac{1}{2}\|\sqrt{\delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y})}\|^2_{\text{tr}},
\end{aligned}$$

where (a) uses the upper bound on $\langle x, y\rangle$ above; (b) uses the fact that $\mathbf{H} \in \mathcal{H}$; (c) uses the definition of $\mathbf{Y}$ above and the fact that $\mathbf{H} \in \mathbb{S}_\succ$; (d) uses the properties of the projection and the fact that $\mathbf{H}^{-1} \in \mathcal{H}$; (e) uses the definition of $\mathbf{H}$ above. After taking the limit $\delta \to +0$ and using the definition of $\rho_*(\cdot)$ in eq. (1), we obtain the desired inequality in eq. (38).

Next, it is easy to show that $\rho_*(\cdot)$ is absolutely homogeneous. Hence, using eq. (38), is easy to show that $\langle x, y\rangle \leq \rho_*(y)$ for all $x \in \mathcal{X}$ such that $\rho(x) \leq 1$. This immediately implies the inequality $\sup_{\rho(x)\leq 1}\langle x, y\rangle \leq \rho_*(y)$. Now, we will prove the reverse inequality. Let $\delta > 0$ and let $x \in \mathcal{X}$ be defined as follows:

$$x = \mathbf{F}y, \quad \text{where} \quad \mathbf{F} = [\delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y})]^{-1/2}. \tag{41}$$

Using Assumption 5 and the fact that $\mathcal{H}$ is closed under operator functions from Part I, we can show that $\mathbf{F} \in \mathbb{S}_{\succ} \cap \mathcal{H}$. Hence, we can verify that $\rho(x) \leq 1$ as follows:

$$\begin{aligned}
\rho^2(x) &\overset{\text{(a)}}{=} \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \text{proj}_{\mathcal{H}}(\mathbf{X}), \mathbf{H} \rangle \\
&\overset{\text{(b)}}{=} \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \mathbf{X}, \text{proj}_{\mathcal{H}}(\mathbf{H}) \rangle \\
&\overset{\text{(c)}}{=} \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \mathbf{F}\mathbf{Y}\mathbf{F}, \text{proj}_{\mathcal{H}}(\mathbf{H}) \rangle \\
&= \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \mathbf{Y}, \mathbf{F}\, \text{proj}_{\mathcal{H}}(\mathbf{H})\mathbf{F} \rangle \\
&\overset{\text{(d)}}{=} \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \text{proj}_{\mathcal{H}}(\mathbf{Y}), \mathbf{F}\, \text{proj}_{\mathcal{H}}(\mathbf{H})\mathbf{F} \rangle \\
&\overset{\text{(e)}}{\leq} \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y}), \mathbf{F}\, \text{proj}_{\mathcal{H}}(\mathbf{H})\mathbf{F} \rangle \\
&\overset{\text{(f)}}{=} \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \mathbf{F}^{-2}, \mathbf{F}\, \text{proj}_{\mathcal{H}}(\mathbf{H})\mathbf{F} \rangle \\
&\overset{\text{(g)}}{=} \sup_{\|\mathbf{H}\|_{\text{tr}} \leq 1, \mathbf{H} \in \mathbb{S}_{\succ}} \langle \mathbf{I}, \mathbf{H} \rangle = 1,
\end{aligned}$$

where (a) uses the definition of $\rho(\cdot)$ in eq. (1) and the properties of $\|\cdot\|_{\text{op}}$; (b) uses the properties of the projection; (c) uses the definitions of x and $\mathbf{Y}$ above and the definition of $\mathbf{X}$ in eq. (1); (d) uses the properties of the projection and the fact that $\mathbf{F}\, \text{proj}_{\mathcal{H}}(\mathbf{H})\mathbf{F} \in \mathcal{H}$ due to Assumption 5; (e) uses the fact that $\text{proj}_{\mathcal{H}}(\mathbf{H}) \in \mathbb{S}_{\succ}$ due to Assumption 5; (f) uses the definition of $\mathbf{F}$ above; (g) uses the properties of the projection and fact that $\mathbf{I} \in \mathcal{H}$ due to Assumption 5. Hence, we can lower-bound $\sup_{\rho(x) \leq 1} \langle x, y \rangle$ as follows:

$$\begin{aligned}
\sup_{\rho(x) \leq 1} \langle x, y \rangle &\overset{\text{(a)}}{\geq} \sup_{\delta > 0} \langle \mathbf{F}, \text{proj}_{\mathcal{H}}(\mathbf{Y}) \rangle \\
&\overset{\text{(b)}}{=} \sup_{\delta > 0} \langle [\delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y})]^{-1/2}, \text{proj}_{\mathcal{H}}(\mathbf{Y}) \rangle \\
&= \sup_{\delta > 0} [\|[\delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y})]^{1/2}\|_{\text{tr}} - \delta \|[\delta\mathbf{I} + \text{proj}_{\mathcal{H}}(\mathbf{Y})]^{-1/2}\|_{\text{tr}}] \\
&\overset{\text{(c)}}{\geq} \|[\text{proj}_{\mathcal{H}}(\mathbf{Y})]^{1/2}\|_{\text{tr}} \overset{\text{(d)}}{=} \rho_*(y),
\end{aligned}$$

where (a) uses the definitions of x and $\mathbf{Y}$ above, the properties of the projection, and the fact that $\mathbf{F} \in \mathcal{H}$; (b) uses the definition of $\mathbf{F}$ above; (c) is obtained by taking the limit $\delta \to +0$; (d) uses the definition of $\rho_*(\cdot)$ in eq. (1). Therefore, we can conclude that $\sup_{\rho(x) \leq 1} \langle x, y \rangle = \rho_*(y)$.

Part III. From Part II, it follows that $\rho_*(\cdot)$ is a dual norm to $\rho(\cdot)$. This implies that $\rho(\cdot)$ is a dual norm to $\rho_*(\cdot)$, that is, $\sup_{\rho_*(y) \leq 1} \langle x, y \rangle = \rho(x)$. □

B.2 Proof of Lemma 2

We can upper-bound $\|x\|_{\mathbf{H}}^2$ as follows:

$$\|x\|_{\mathbf{H}}^2 \overset{\text{(a)}}{=} \langle \mathbf{H}, \mathbf{X} \rangle \overset{\text{(b)}}{=} \langle \mathbf{H}, \text{proj}_{\mathcal{H}}(\mathbf{X}) \rangle \overset{\text{(c)}}{\leq} \|\mathbf{H}\|_{\text{tr}} \cdot \|\text{proj}_{\mathcal{H}}(\mathbf{X})\|_{\text{op}} \overset{\text{(d)}}{=} \|\mathbf{H}\|_{\text{tr}} \cdot \rho^2(x),$$

where (a) uses the definition of $\mathbf{X}$ in eq. (1); (b) uses the assumption $\mathbf{H} \in \mathcal{H}$ and the properties of the projection; (c) uses Hölder's inequality for Schatten norms; (d) uses the definition of $\rho(\cdot)$ in eq. (1). Furthermore, we can upper-bound $\rho_*^2(x)$ as follows:

$$\rho_*^2(x) \overset{\text{(a)}}{=} \sup_{\rho(x') \leq 1} \langle x, x' \rangle^2 \overset{\text{(b)}}{\leq} \sup_{\|x'\|_{\mathbf{H}}^2 \leq \|\mathbf{H}\|_{\text{tr}}} \langle x, x' \rangle^2 = \|\mathbf{H}\|_{\text{tr}} \cdot \|x\|_{\mathbf{H}^{-1}}^2,$$

where (a) uses Lemma 1; (b) uses the previously obtained inequality $\|x'\|_{\mathbf{H}}^2 \leq \|\mathbf{H}\|_{\text{tr}} \cdot \rho^2(x')$. □

C PROOFS FOR SECTION 2.3

C.1 PROOF OF LEMMA 3

Assume $\mathcal{R} < +\infty$, otherwise the proof is trivial. Suppose that $x^* \in Q$ is a first-order stationary point for the optimization problem $\min_{x \in Q} f(x)$, that is, $-\nabla f(x^*) \in \mathcal{N}_Q(x^*)$ By the definition of the normal cone, we have

$$\langle -\nabla f(x^*), x^* - x \rangle \geq 0 \quad \text{for all} \;\; x \in Q,$$

which implies $\mathcal{G}_{\mathcal{R}}(x^*) \leq 0$. Hence, using the fact that $\mathcal{G}_{\mathcal{R}}(x) \geq 0$ for all $x \in Q$, we get $\mathcal{G}_{\mathcal{R}}(x^*) = 0$.

Conversely, suppose that $\mathcal{G}_{\mathcal{R}}(x^*) = 0$. Then, by the definition of $\mathcal{G}_{\mathcal{R}}(\cdot)$ in eq. (9), we have

$$\langle \nabla f(x^*), x^* - x \rangle \leq 0 \quad \text{for all} \;\; x \in Q,$$

which by the definition of the normal cone implies $-\nabla f(x^*) \in \mathcal{N}_Q(x^*)$. □

C.2 PROOF OF LEMMA 4

We can upper-bound $f(x) - f^*$ for all $x \in Q$ as follows:

$$f(x) - f^* \overset{\text{(a)}}{\leq} \max_{x' \in Q} [f(x) - f(x')] \overset{\text{(b)}}{\leq} \max_{x' \in Q} \langle \nabla f(x), x - x' \rangle \overset{\text{(c)}}{\leq} \mathcal{R} \cdot \mathcal{G}_{\mathcal{R}}(x),$$

where (a) uses the definition of f^* in problem (8); (b) uses the convexity assumption; (c) uses the definition of $\mathcal{G}_{\mathcal{R}}(x)$ in eq. (9) and the definition of Q. □

D PROOFS FOR SECTION 3.1

D.1 PROOF OF LEMMA 5

We prove the inequality $\beta\rho(x_k) \leq \eta$ by induction. The base case $k = 0$ is implied by the inclusion $x_0 \in Q$ on line 1 and the definition of Q. Furthermore, we assume the inequality $\beta\rho(x_k) \leq \eta$ and prove the inequality $\beta\rho(x_{k+1}) \leq \eta$ as follows:

$$\begin{aligned}\beta\rho(x_{k+1}) &\overset{\text{(a)}}{\leq} \beta\rho(x_{k+1} - (1-\beta)x_k) + (1-\beta)\beta\rho(x_k) \\ &\overset{\text{(b)}}{\leq} \beta\eta + (1-\beta)\beta\rho(x_k) \overset{\text{(c)}}{\leq} \beta\eta + (1-\beta)\eta = \eta,\end{aligned}$$

where (a) uses the properties of the norm $\rho(\cdot)$; (b) uses the definition of x_{k+1} in eq. (11); (c) uses the induction hypothesis. Next, we prove the inequality $\rho(x_{k+1} - x_k) \leq 2\eta$ as follows:

$$\begin{aligned}\rho(x_{k+1} - x_k) &\overset{\text{(a)}}{\leq} \rho(x_{k+1} - (1-\beta)x_k) + \beta\rho(x_k) \\ &\overset{\text{(b)}}{\leq} \eta + \beta\rho(x_k) \overset{\text{(c)}}{\leq} 2\eta,\end{aligned}$$

where (a) uses the properties of the norm $\rho(\cdot)$; (b) uses the definition of x_{k+1} in eq. (11); (c) uses the previously obtained inequality above. □

D.2 PROOF OF LEMMA 6

The inequality $\mathbf{H} \preceq 4\|\mathbf{H}\|_{\text{tr}} \cdot \mathbf{B}$ is trivially implied by the definition of $\mathbf{B}$ in eq. (13) and Assumption 5. Hence, using the Löwner-Heinz theorem (Carlen, 2010, Theorem 2.6), we obtain the inequality $4\|\mathbf{H}\|_{\text{tr}} \cdot \mathbf{H}^{-1} \succeq \mathbf{B}^{-1}$. □

D.3 PROOF OF LEMMA 7

In the case $\beta > 0$, we can upper-bound $f(x_{k+1})$ as follows:

$$\begin{aligned}f(x_{k+1}) &\overset{\text{(a)}}{\leq} f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \tfrac{1}{2}\|\mathbf{L}\|_{\text{tr}}\rho^2(x_{k+1} - x_k) \\ &\overset{\text{(b)}}{\leq} f(x_k) + \langle \nabla f(x_k) - m_k, x_{k+1} - x_k \rangle + \langle m_k, x_{k+1} - x_k \rangle + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 \\ &\overset{\text{(c)}}{\leq} f(x_k) + \langle \nabla f(x_k) - m_k, x_{k+1} - x_k \rangle + \min_{\rho(x-(1-\beta)x_k)\leq\eta} \langle m_k, x - x_k \rangle + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 \\ &\overset{\text{(d)}}{=} f(x_k) + \langle \nabla f(x_k) - m_k, x_{k+1} - x_k \rangle + \min_{\beta\rho(z)\leq\eta} \beta\langle m_k, z - x_k \rangle + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 \\ &\overset{\text{(e)}}{\leq} f(x_k) + \langle \nabla f(x_k) - m_k, x_{k+1} - \beta z - (1-\beta)x_k \rangle + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 - \eta\mathcal{G}_{\eta/\beta}(x_k) \\ &\overset{\text{(f)}}{\leq} f(x_k) + \|x_{k+1} - \beta z - (1-\beta)x_k\|_{\mathbf{B}}\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 - \eta\mathcal{G}_{\eta/\beta}(x_k) \\ &\overset{\text{(g)}}{\leq} f(x_k) + \rho(x_{k+1} - \beta z - (1-\beta)x_k)\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 - \eta\mathcal{G}_{\eta/\beta}(x_k) \\ &\overset{\text{(h)}}{\leq} f(x_k) + 4\eta\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 - \eta\mathcal{G}_{\eta/\beta}(x_k),\end{aligned}$$

where (a) uses Assumptions 1 and 5 and Lemma 2; (b) uses Lemma 5; (c) uses the definition of x_{k+1} in eq. (11); (d) uses the variable exchange $x = \beta z + (1-\beta)x_k$; (e) uses the definition of $\mathcal{G}_{\eta/\beta}(\cdot)$ in eq. (9) and $z = \text{argmin}_{\beta\rho(z)\leq\eta}\langle \nabla f(x_k), z - x_k \rangle$; (f) uses the Cauchy-Schwarz inequality; (g) uses Lemma 2 and the definition of $\mathbf{B}$ in eq. (13); (h) uses the triangle inequality, Lemma 5 and the fact that $\beta\rho(z) \leq \eta$. In the case $\beta = 0$, we can upper-bound $f(x_{k+1})$ as follows:

$$\begin{aligned}f(x_{k+1}) &\overset{\text{(a)}}{\leq} f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \tfrac{1}{2}\|\mathbf{L}\|_{\text{tr}}\rho^2(x_{k+1} - x_k) \\ &\overset{\text{(b)}}{\leq} f(x_k) + \langle \nabla f(x_k) - m_k, x_{k+1} - x_k \rangle + \langle m_k, x_{k+1} - x_k \rangle + 2\|\mathbf{L}\|_{\text{tr}}\eta^2\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(c)}}{\leq} f(x_k) + \langle \nabla f(x_k) - m_k, x_{k+1} - x_k \rangle + \min_{\rho(x - x_k) \leq \eta} \langle m_k, x - x_k \rangle + 2\|\mathbf{L}\|_{\text{tr}} \eta^2 \\
&\overset{\text{(d)}}{=} f(x_k) + \langle \nabla f(x_k) - m_k, x_{k+1} - x_k \rangle - \eta \rho_*(m_k) + 2\|\mathbf{L}\|_{\text{tr}} \eta^2 \\
&\overset{\text{(e)}}{\leq} f(x_k) + \|x_{k+1} - x_k\|_{\mathbf{B}} \|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} - \eta \rho_*(m_k) + 2\|\mathbf{L}\|_{\text{tr}} \eta^2 \\
&\overset{\text{(f)}}{\leq} f(x_k) + \|x_{k+1} - x_k\|_{\mathbf{B}} \|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} + \eta \rho_*(\nabla f(x_k) - m_k) - \eta \rho_*(\nabla f(x_k)) \\
&\quad + 2\|\mathbf{L}\|_{\text{tr}} \eta^2 \\
&\overset{\text{(g)}}{\leq} f(x_k) + [\rho(x_{k+1} - x_k) + \eta] \|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} - \eta \rho_*(\nabla f(x_k)) + 2\|\mathbf{L}\|_{\text{tr}} \eta^2 \\
&\overset{\text{(h)}}{\leq} f(x_k) + 4\eta \|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} - \eta \rho_*(\nabla f(x_k)) + 2\|\mathbf{L}\|_{\text{tr}} \eta^2 \\
&\overset{\text{(i)}}{=} f(x_k) + 4\eta \|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} - \eta \mathcal{G}_{\eta/\beta}(x_k) + 2\|\mathbf{L}\|_{\text{tr}} \eta^2,
\end{aligned}
$$

where (a) uses Assumptions 1 and 5 and Lemma 2; (b) uses Lemma 5; (c) uses the definition of x_{k+1} in eq. (11) and $\beta = 0$; (d) uses Lemma 1; (e) uses the Cauchy-Schwarz inequality; (f) uses the triangle inequality; (g) uses Lemma 2 and the definition of $\mathbf{B}$ in eq. (13); (h) uses Lemma 5 and the inequality $3 \leq 4$; (i) uses the definition of $\mathcal{G}_{+\infty}(\cdot)$ in eq. (9) and $\eta/\beta = +\infty$. □

E PROOFS FOR SECTION 3.2 (OPTION 12A)

E.1 PROOF OF LEMMA 8

We can upper-bound $\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}]$ as follows:

$$\begin{aligned}
&\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}] = \\
&\quad= \mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[m_{k+1}]-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}] + \mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[m_{k+1}]\|^2_{\mathbf{B}^{-1}}] \\
&\quad\overset{(a)}{\leq} \|(1-\alpha)(m_k-\nabla f(x_{k+1}))\|^2_{\mathbf{B}^{-1}} + \alpha^2\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|n(x_{k+1};\xi_{k+1})\|^2_{\mathbf{B}^{-1}}] \\
&\quad\overset{(b)}{\leq} \|(1-\alpha)(m_k-\nabla f(x_k)) + (1-\alpha)(\nabla f(x_k)-\nabla f(x_{k+1}))\|^2_{\mathbf{B}^{-1}} + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{(c)}{\leq} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + ((1-\alpha)^2/\alpha)\|\nabla f(x_k)-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{(d)}{\leq} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + (4\|\mathbf{L}\|_{\mathrm{tr}}/\alpha)\|x_k-x_{k+1}\|^2_{\mathbf{L}} + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{(e)}{\leq} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + 16\|\mathbf{L}\|^2_{\mathrm{tr}}\eta^2/\alpha + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}},
\end{aligned}$$

where (a) uses eq. (12a) and the definition of $n(x;\xi)$ in Assumption 2; (b) uses Lemma 6 and Assumption 2; (c) uses the convexity of $\|\cdot\|^2_{\mathbf{B}^{-1}}$; (d) is implied by Assumption 1 and Lemma 6; (e) uses Lemma 2, the definition of $\mathbf{B}$ in eq. (13), and Lemma 5. It remains to take the full expectation. □

E.2 PROOF OF COROLLARY 1

We can upper-bound $\mathbb{E}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}]$ as follows:

$$\begin{aligned}
&\mathbb{E}\big[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}\big] \leq \\
&\qquad\overset{(a)}{\leq} (1-\alpha)^{k+1}\mathbb{E}\big[\|m_0-\nabla f(x_0)\|^2_{\mathbf{B}^{-1}}\big] + \big[16\|\mathbf{L}\|^2_{\mathrm{tr}}\eta^2/\alpha + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\big]\textstyle\sum_{i=0}^k(1-\alpha)^i \\
&\qquad\overset{(b)}{\leq} (1-\alpha)^{k+1}\mathbb{E}\big[\|n(x_0;\xi_0)\|^2_{\mathbf{B}^{-1}}\big] + \big[16\|\mathbf{L}\|^2_{\mathrm{tr}}\eta^2/\alpha + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\big]\textstyle\sum_{i=0}^k(1-\alpha)^i \\
&\qquad\overset{(c)}{\leq} 4(1-\alpha)^{k+1}\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + \big[16\|\mathbf{L}\|^2_{\mathrm{tr}}\eta^2/\alpha + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\big]\textstyle\sum_{i=0}^k(1-\alpha)^i \\
&\qquad\overset{(d)}{\leq} 4\big[(1-\alpha)^{k+1}+\alpha\big]\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + 16(\eta/\alpha)^2\|\mathbf{L}\|^2_{\mathrm{tr}},
\end{aligned}$$

where (a) uses Lemma 8; (b) uses the definition $m_0 = g(x_0;\xi_0)$ on line 3 and the definition of $n(x;\xi)$ in Assumption 2; (c) uses Lemma 6 and Assumption 2; (d) uses the inequality $\sum_{i=0}^k(1-\alpha)^i \leq 1/\alpha$. Now, we can upper-bound $\mathbb{E}[\|m_k-\nabla f(x_k)\|_{\mathbf{B}^{-1}}]$ as follows:

$$\begin{aligned}
\mathbb{E}[\|m_k-\nabla f(x_k)\|_{\mathbf{B}^{-1}}] &\overset{(a)}{\leq} \sqrt{\mathbb{E}\big[\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}}\big]} \\
&\overset{(b)}{\leq} \sqrt{4[(1-\alpha)^k+\alpha]\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + 16(\eta/\alpha)^2\|\mathbf{L}\|^2_{\mathrm{tr}}} \\
&\overset{(c)}{\leq} 2\big[(1-\alpha/2)^k+\sqrt{\alpha}\big]\|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4\|\mathbf{L}\|_{\mathrm{tr}}\eta/\alpha,
\end{aligned}$$

where (a) uses Jensen's inequality for expectations; (b) uses the upper bound above; (c) uses the inequality $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ and the inequality $\sqrt{1-\alpha} \leq 1-\alpha/2$. □

E.3 PROOF OF THEOREM 1

We can upper-bound $\mathbb{E}\big[\min_{k\in\{0,\ldots,K-1\}}\mathcal{G}_{\mathcal{R}}(x_k)\big]$ as follows:

$$\begin{aligned}
&\mathbb{E}\big[\min\nolimits_{k\in\{0,\ldots,K-1\}}\mathcal{G}_{\mathcal{R}}(x_k)\big] \leq \\
&\qquad\leq \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\mathbb{E}[\mathcal{G}_{\mathcal{R}}(x_k)] \\
&\qquad\overset{(a)}{\leq} \tfrac{1}{\eta K}\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[f(x_k)-f(x_{k+1})+4\eta\|\nabla f(x_k)-m_k\|_{\mathbf{B}^{-1}} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2\big]
\end{aligned}$$

$$\overset{\text{(b)}}{\leq} \tfrac{\Delta_0}{\eta K} + 2\|\mathbf{L}\|_{\text{tr}}\eta + \tfrac{4}{K}\textstyle\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}]$$
$$\overset{\text{(c)}}{\leq} \tfrac{\Delta_0}{\eta K} + 2\|\mathbf{L}\|_{\text{tr}}\eta + \tfrac{4}{K}\textstyle\sum_{k=0}^{K-1}\big[2\big[(1-\alpha/2)^k + \sqrt{\alpha}\big]\|\mathbf{\Sigma}\|_{\text{tr}} + 4\|\mathbf{L}\|_{\text{tr}}\eta/\alpha\big]$$
$$\overset{\text{(d)}}{\leq} \tfrac{\Delta_0}{\eta K} + \tfrac{16\|\mathbf{\Sigma}\|_{\text{tr}}}{\alpha K} + \tfrac{16\|\mathbf{L}\|_{\text{tr}}\eta}{\alpha} + 2\|\mathbf{L}\|_{\text{tr}}\eta + 8\sqrt{\alpha}\|\mathbf{\Sigma}\|_{\text{tr}},$$

where (a) uses Lemma 7 and the definition of η in eq. (22); (b) uses the definition of Δ_0; (c) uses Corollary 1; (d) uses the inequality $\sum_{k=0}^{K-1}(1-\alpha/2)^k \leq 2/\alpha$. It remains to use the definitons of the parameters in eqs. (22) and (23).

Furthermore, in the convex case, we define the following Lyapunov function:

$$\Psi_k = f(x_k) - f^* + c\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}, \tag{42}$$

where $c > 0$. We can upper-bound $\mathbb{E}[\Psi_{k+1}]$ as follows:

$$\begin{aligned}
\mathbb{E}[\Psi_{k+1}] &\overset{\text{(a)}}{\leq} \mathbb{E}[f(x_k) - f^*] - \eta\mathbb{E}[\mathcal{G}_{\eta/\beta}(x_k)] + 4\eta\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}] + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 \\
&\quad + c(1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] + \tfrac{16c\eta^2}{\alpha}\|\mathbf{L}\|^2_{\text{tr}} + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\text{tr}} \\
&\overset{\text{(b)}}{\leq} (1-\beta)\mathbb{E}[f(x_k) - f^*] + c(1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] \\
&\quad + 4\eta\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}] + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 + \tfrac{16c\eta^2}{\alpha}\|\mathbf{L}\|^2_{\text{tr}} + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\text{tr}} \\
&\overset{\text{(c)}}{\leq} (1-\beta)\mathbb{E}[f(x_k) - f^*] + c(1-\alpha/2)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] + \tfrac{8\eta^2}{c\alpha} \\
&\quad + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 + \tfrac{16c\eta^2}{\alpha}\|\mathbf{L}\|^2_{\text{tr}} + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\text{tr}} \\
&\overset{\text{(d)}}{\leq} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 + \tfrac{16c\eta^2}{\alpha}\|\mathbf{L}\|^2_{\text{tr}} + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\text{tr}} + \tfrac{8\eta^2}{c\alpha} \\
&\overset{\text{(e)}}{=} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 + 2\sqrt{\tfrac{8\eta^2}{c\alpha}}\sqrt{\tfrac{16c\eta^2}{\alpha}\|\mathbf{L}\|^2_{\text{tr}} + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\text{tr}}} \\
&\overset{\text{(f)}}{\leq} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\text{tr}}\eta^2 + \tfrac{16\sqrt{2}\eta^2}{\alpha}\|\mathbf{L}\|_{\text{tr}} + 8\sqrt{2\alpha}\eta\|\mathbf{\Sigma}\|_{\text{tr}} \\
&\overset{\text{(g)}}{\leq} \big(1 - \tfrac{1}{2}\beta\big)\mathbb{E}[\Psi_k] + \tfrac{1}{4}\beta\epsilon,
\end{aligned}$$

where (a) uses Lemmas 7 and 8; (b) uses the definition of η in eq. (24) and Lemma 4; (c) uses Young's inequality; (d) uses the definition of Ψ_k in eq. (42); (e) is obtained by minimizing in $c > 0$; (f) uses the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$; (g) uses the parameters in eq. (24). Hence, we can upper-bound $\mathbb{E}[f(x_K) - f^*]$ as follows:

$$\mathbb{E}[f(x_K) - f^*] \overset{\text{(a)}}{\leq} \mathbb{E}[\Psi_K] \overset{\text{(b)}}{\leq} \big(1 - \tfrac{1}{2}\beta\big)^K\mathbb{E}[\Psi_0] + \tfrac{1}{2}\epsilon \overset{\text{(c)}}{\leq} \epsilon,$$

where (a) uses the definition of Ψ_k in eq. (42); (b) uses the upper bound on $\mathbb{E}[\Psi_k]$ above and the inequality $\sum_{k=0}^{K-1}(1-\beta/2)^k \leq 2/\beta$; (c) uses the choice of K in eq. (25). □

F Proofs for Section 3.2 (Option 12b)

F.1 Proof of Lemma 9

We can upper-bound $\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}]$ as follows:

$$
\begin{aligned}
&\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}] = \\
&\quad= \mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[m_{k+1}]-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}] + \mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[m_{k+1}]\|^2_{\mathbf{B}^{-1}}] \\
&\quad\overset{\text{(a)}}{\leq} \|(1-\alpha)m_k+\alpha\nabla f(\overline{x}_{k+1})-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} + \alpha^2\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|n(\overline{x}_{k+1};\xi_{k+1})\|^2_{\mathbf{B}^{-1}}] \\
&\quad\overset{\text{(b)}}{\leq} \|(1-\alpha)(m_k-\nabla f(x_k))+\alpha\nabla f(\overline{x}_{k+1})+(1-\alpha)\nabla f(x_k)-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{\text{(c)}}{\leq} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + (1/\alpha)\|\alpha\nabla f(\overline{x}_{k+1})+(1-\alpha)\nabla f(x_k)-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} \\
&\quad\quad + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{\text{(d)}}{\leq} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + (3/\alpha)\|\nabla^2 f(x_{k+1})(\alpha\overline{x}_{k+1}+(1-\alpha)x_k-x_{k+1})\|^2_{\mathbf{B}^{-1}} \\
&\quad\quad + (3(1-\alpha)^2/\alpha)\|\nabla f(x_k)-\nabla f(x_{k+1})-\nabla^2 f(x_{k+1})(x_k-x_{k+1})\|^2_{\mathbf{B}^{-1}} \\
&\quad\quad + 3\alpha\|\nabla f(\overline{x}_{k+1})-\nabla f(x_{k+1})-\nabla^2 f(x_{k+1})(\overline{x}_{k+1}-x_{k+1})\|^2_{\mathbf{B}^{-1}} + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{\text{(e)}}{=} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\quad + (3(1-\alpha)^2/\alpha)\|\nabla f(x_k)-\nabla f(x_{k+1})-\nabla^2 f(x_{k+1})(x_k-x_{k+1})\|^2_{\mathbf{B}^{-1}} \\
&\quad\quad + 3\alpha\|\nabla f(\overline{x}_{k+1})-\nabla f(x_{k+1})-\nabla^2 f(x_{k+1})(\overline{x}_{k+1}-x_{k+1})\|^2_{\mathbf{B}^{-1}} \\
&\quad\overset{\text{(f)}}{\leq} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + \tfrac{3(1-\alpha)^2}{\alpha}\|x_k-x_{k+1}\|^4_{\mathbf{T}} + 3\alpha\|\overline{x}_{k+1}-x_{k+1}\|^4_{\mathbf{T}} \\
&\quad\quad + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{\text{(g)}}{=} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + \tfrac{3(1-\alpha)^2}{\alpha}\Big[1+\tfrac{(1-\alpha)^2}{\alpha^2}\Big]\|x_k-x_{k+1}\|^4_{\mathbf{T}} + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\quad\overset{\text{(h)}}{\leq} (1-\alpha)\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + 48\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}
\end{aligned}
$$

where (a) uses eq. (12b) and the definition of $n(x;\xi)$ in Assumption 2; (b) uses Lemma 6 and Assumption 2; (c) uses the convexity of $\|\cdot\|^2_{\mathbf{B}^{-1}}$; (d) uses the triangle inequality; (e) and (g) use the definition of $\overline{x}_{k+1}$ in eq. (12b); (f) is implied by Assumption 3 and Lemma 6; (h) uses Lemmas 2 and 5. It remains to take the full expectation. □

F.2 Proof of Corollary 2

We can upper-bound $\mathbb{E}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}]$ as follows:

$$
\begin{aligned}
&\mathbb{E}\big[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}\big] \leq \\
&\qquad\overset{\text{(a)}}{\leq} (1-\alpha)^{k+1}\mathbb{E}\big[\|m_0-\nabla f(x_0)\|^2_{\mathbf{B}^{-1}}\big] + \big[48\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\big]\textstyle\sum_{i=0}^k(1-\alpha)^i \\
&\qquad\overset{\text{(b)}}{\leq} (1-\alpha)^{k+1}\mathbb{E}\big[\|n(x_0;\xi_0)\|^2_{\mathbf{B}^{-1}}\big] + \big[48\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\big]\textstyle\sum_{i=0}^k(1-\alpha)^i \\
&\qquad\overset{\text{(c)}}{\leq} 4(1-\alpha)^{k+1}\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + \big[48\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\big]\textstyle\sum_{i=0}^k(1-\alpha)^i \\
&\qquad\overset{\text{(d)}}{\leq} 4\big[(1-\alpha)^{k+1}+\alpha\big]\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + 48(\eta/\alpha)^4\|\mathbf{T}\|^2_{\mathrm{tr}}
\end{aligned}
$$

where (a) uses Lemma 9; (b) uses the definition $m_0 = g(x_0;\xi_0)$ on line 3 and the definition of $n(x;\xi)$ in Assumption 2; (c) uses Lemma 6 and Assumption 2; (d) uses the inequality $\sum_{i=0}^k(1-\alpha)^i \leq 1/\alpha$. Now, we can upper-bound $\mathbb{E}[\|m_k-\nabla f(x_k)\|_{\mathbf{B}^{-1}}]$ as follows:

$$
\begin{aligned}
\mathbb{E}[\|m_k-\nabla f(x_k)\|_{\mathbf{B}^{-1}}] &\overset{\text{(a)}}{\leq} \sqrt{\mathbb{E}\big[\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}}\big]} \\
&\overset{\text{(b)}}{\leq} \sqrt{4[(1-\alpha)^k+\alpha]\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + 48(\eta/\alpha)^4\|\mathbf{T}\|^2_{\mathrm{tr}}}
\end{aligned}
$$

$$\overset{(c)}{\leq} 2\big[(1-\alpha/2)^k + \sqrt{\alpha}\big]\|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4\sqrt{3}(\eta/\alpha)^2\|\mathbf{T}\|_{\mathrm{tr}},$$

where (a) uses Jensen's inequality for expectations; (b) uses the upper bound above; (c) uses the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and the inequality $\sqrt{1-\alpha} \leq 1 - \alpha/2$. □

F.3 Proof of Theorem 2

We can upper-bound $\mathbb{E}\big[\min_{k\in\{0,\dots,K-1\}} \mathcal{G}_{\mathcal{R}}(x_k)\big]$ as follows:

$$\begin{aligned}
&\mathbb{E}\big[\min_{k\in\{0,\dots,K-1\}}\mathcal{G}_{\mathcal{R}}(x_k)\big] \leq \\
&\quad\leq \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\mathbb{E}[\mathcal{G}_{\mathcal{R}}(x_k)] \\
&\quad\overset{(a)}{\leq} \tfrac{1}{\eta K}\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[f(x_k) - f(x_{k+1}) + 4\eta\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2\big] \\
&\quad\overset{(b)}{\leq} \tfrac{\Delta_0}{\eta K} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta + \tfrac{4}{K}\textstyle\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}] \\
&\quad\overset{(c)}{\leq} \tfrac{\Delta_0}{\eta K} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta + \tfrac{4}{K}\textstyle\sum_{k=0}^{K-1}\Big[2\big[(1-\alpha/2)^k + \sqrt{\alpha}\big]\|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4\sqrt{3}(\eta/\alpha)^2\|\mathbf{T}\|_{\mathrm{tr}}\Big] \\
&\quad\overset{(d)}{\leq} \tfrac{\Delta_0}{\eta K} + \tfrac{16\|\mathbf{\Sigma}\|_{\mathrm{tr}}}{\alpha K} + \tfrac{16\sqrt{3}\|\mathbf{T}\|_{\mathrm{tr}}\eta^2}{\alpha^2} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta + 8\sqrt{\alpha}\|\mathbf{\Sigma}\|_{\mathrm{tr}},
\end{aligned}$$

where (a) uses Lemma 7 and the definition of η in eq. (26); (b) uses the definition of Δ_0; (c) uses Corollary 2; (d) uses the inequality $\sum_{k=0}^{K-1}(1-\alpha/2)^k \leq 2/\alpha$. It remains to use the definitons of the parameters in eqs. (26) and (27).

Furthermore, in the convex case, we define the following Lyapunov function:

$$\Psi_k = f(x_k) - f^* + c\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}, \tag{43}$$

where $c > 0$. We can upper-bound $\mathbb{E}[\Psi_{k+1}]$ as follows:

$$\begin{aligned}
\mathbb{E}[\Psi_{k+1}] &\overset{(a)}{\leq} \mathbb{E}[f(x_k) - f^*] - \eta\mathbb{E}[\mathcal{G}_{\eta/\beta}(x_k)] + 4\eta\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 \\
&\quad + c(1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] + 48c\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\overset{(b)}{\leq} (1-\beta)\mathbb{E}[f(x_k) - f^*] + c(1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] \\
&\quad + 4\eta\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 48c\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\overset{(c)}{\leq} (1-\beta)\mathbb{E}[f(x_k) - f^*] + c(1-\alpha/2)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] + \tfrac{8\eta^2}{c\alpha} \\
&\quad + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 48c\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\overset{(d)}{\leq} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 48c\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + \tfrac{8\eta^2}{c\alpha} \\
&\overset{(e)}{=} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 2\sqrt{\tfrac{8\eta^2}{c\alpha}}\sqrt{48c\|\mathbf{T}\|^2_{\mathrm{tr}}\eta^4/\alpha^3 + 4c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}} \\
&\overset{(f)}{\leq} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + \tfrac{16\sqrt{6}\eta^3}{\alpha^2}\|\mathbf{T}\|_{\mathrm{tr}} + 8\sqrt{2c\alpha}\eta\|\mathbf{\Sigma}\|_{\mathrm{tr}} \\
&\overset{(g)}{\leq} \big(1 - \tfrac{1}{2}\beta\big)\mathbb{E}[\Psi_k] + \tfrac{1}{4}\beta\epsilon,
\end{aligned}$$

where (a) uses Lemmas 7 and 9; (b) uses the definition of η in eq. (28) and Lemma 4; (c) uses Young's inequality; (d) uses the definition of Ψ_k in eq. (43); (e) is obtained by minimizing in $c > 0$; (f) uses the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$; (g) uses the parameters in eq. (28). Hence, we can upper-bound $\mathbb{E}[f(x_K) - f^*]$ as follows:

$$\mathbb{E}[f(x_K) - f^*] \overset{(a)}{\leq} \mathbb{E}[\Psi_K] \overset{(b)}{\leq} \big(1 - \tfrac{1}{2}\beta\big)^K\mathbb{E}[\Psi_0] + \tfrac{1}{2}\epsilon \overset{(c)}{\leq} \epsilon,$$

where (a) uses the definition of Ψ_k in eq. (43); (b) uses the upper bound on $\mathbb{E}[\Psi_k]$ above and the inequality $\sum_{k=0}^{K-1}(1-\beta/2)^k \leq 2/\beta$; (c) uses the choice of K in eq. (29). □

G Proofs for Section 3.2 (Option 12c)

G.1 Proof of Lemma 10

We can upper-bound $\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}]$ as follows:

$$
\begin{aligned}
&\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}] = \\
&= \mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|m_{k+1}-\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[m_{k+1}]\|^2_{\mathbf{B}^{-1}}] + \mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[m_{k+1}]-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}] \\
&\overset{(a)}{=} \mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|(1-\alpha)(g(x_k;\xi_{k+1})-\nabla f(x_k)) - (g(x_{k+1};\xi_{k+1})-\nabla f(x_{k+1}))\|^2_{\mathbf{B}^{-1}}] \\
&+ (1-\alpha)^2\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} \\
&\overset{(b)}{\leq} 2(1-\alpha)^2\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|g(x_{k+1};\xi_{k+1})-g(x_k;\xi_{k+1})\|^2_{\mathbf{B}^{-1}}] + 2\alpha^2\mathbb{E}_{\xi_{k+1}\sim\mathcal{D}}[\|n(x_{k+1};\xi_{k+1})\|^2_{\mathbf{B}^{-1}}] \\
&+ (1-\alpha)^2\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} \\
&\overset{(c)}{\leq} 8(1-\alpha)^2\|\mathbf{M}\|_{\mathrm{tr}}\|x_{k+1}-x_k\|^2_{\mathbf{M}} + 8\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + (1-\alpha)^2\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} \\
&\overset{(d)}{\leq} (1-\alpha)^2\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}} + 32\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}},
\end{aligned}
$$

where (a) uses eq. (12c); (b) uses the triangle inequality and the definition of $n(x;\xi)$ in Assumption 2; (c) is implied by Assumptions 2 and 4 and Lemma 6; (d) uses Lemma 2 and the definition of $\mathbf{B}$ in eq. (13). It remains to take the full expectation. □

G.2 Proof of Corollary 3

We can upper-bound $\mathbb{E}[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}]$ as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\|m_{k+1}-\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}}\right] \leq \\
&\quad\overset{(a)}{\leq} (1-\alpha)^{2(k+1)}\mathbb{E}\left[\|m_0-\nabla f(x_0)\|^2_{\mathbf{B}^{-1}}\right] + \left[32\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\right]\textstyle\sum_{i=0}^{k}(1-\alpha)^{2i} \\
&\quad\overset{(b)}{\leq} (1-\alpha)^{2(k+1)}\mathbb{E}\left[\|n(x_0;\xi_0)\|^2_{\mathbf{B}^{-1}}\right] + \left[32\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\right]\textstyle\sum_{i=0}^{k}(1-\alpha)^{2i} \\
&\quad\overset{(c)}{\leq} 4(1-\alpha)^{2(k+1)}\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + \left[32\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}\right]\textstyle\sum_{i=0}^{k}(1-\alpha)^{2i} \\
&\quad\overset{(d)}{\leq} 4[(1-\alpha)^{2(k+1)} + 2\alpha]\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + 32\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2/\alpha,
\end{aligned}
$$

where (a) uses Lemma 10; (b) uses the definition $m_0 = g(x_0;\xi_0)$ on line 3 and the definition of $n(x;\xi)$ in Assumption 2; (c) uses Lemma 6 and Assumption 2; (d) uses the inequality $\sum_{i=0}^{k}(1-\alpha)^{2i} \leq 1/\alpha$. Now, we can upper-bound $\mathbb{E}[\|m_k-\nabla f(x_k)\|_{\mathbf{B}^{-1}}]$ as follows:

$$
\begin{aligned}
\mathbb{E}[\|m_k-\nabla f(x_k)\|_{\mathbf{B}^{-1}}] &\overset{(a)}{\leq} \sqrt{\mathbb{E}\left[\|m_k-\nabla f(x_k)\|^2_{\mathbf{B}^{-1}}\right]} \\
&\overset{(b)}{\leq} \sqrt{4[(1-\alpha)^{2k}+2\alpha]\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + 32\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2/\alpha} \\
&\overset{(c)}{\leq} 2\left[(1-\alpha)^k + \sqrt{2\alpha}\right]\|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4\sqrt{2/\alpha}\|\mathbf{M}\|_{\mathrm{tr}}\eta,
\end{aligned}
$$

where (a) uses Jensen's inequality for expectations; (b) uses the upper bound above; (c) uses the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. □

G.3 Proof of Theorem 3

One can observe that Assumption 4 implies Assumption 1 with $\mathbf{L} = \mathbf{M}$. Furthermore, we can upper-bound $\mathbb{E}\left[\min_{k\in\{0,\ldots,K-1\}}\mathcal{G}_{\mathcal{R}}(x_k)\right]$ as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\min\nolimits_{k\in\{0,\ldots,K-1\}}\mathcal{G}_{\mathcal{R}}(x_k)\right] \leq \\
&\qquad\leq \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\mathbb{E}[\mathcal{G}_{\mathcal{R}}(x_k)]
\end{aligned}
$$

$$\overset{\text{(a)}}{\leq} \tfrac{1}{\eta K}\sum_{k=0}^{K-1}\mathbb{E}\big[f(x_k) - f(x_{k+1}) + 4\eta\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2\big]$$
$$\overset{\text{(b)}}{\leq} \tfrac{\Delta_0}{\eta K} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta + \tfrac{4}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}]$$
$$\overset{\text{(c)}}{\leq} \tfrac{\Delta_0}{\eta K} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta + \tfrac{4}{K}\sum_{k=0}^{K-1}\Big[2[(1-\alpha)^k + \sqrt{2\alpha}]\|\mathbf{\Sigma}\|_{\mathrm{tr}} + 4\sqrt{2/\alpha}\|\mathbf{M}\|_{\mathrm{tr}}\eta\Big]$$
$$\overset{\text{(d)}}{\leq} \tfrac{\Delta_0}{\eta K} + \tfrac{8\|\mathbf{\Sigma}\|_{\mathrm{tr}}}{\alpha K} + \tfrac{16\sqrt{2}\|\mathbf{M}\|_{\mathrm{tr}}\eta}{\sqrt{\alpha}} + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta + 8\sqrt{2\alpha}\|\mathbf{\Sigma}\|_{\mathrm{tr}},$$

where (a) uses Lemma 7 and the definition of η in eq. (30); (b) uses the definition of Δ_0; (c) uses Corollary 3; (d) uses the inequality $\sum_{k=0}^{K-1}(1-\alpha)^k \leq 1/\alpha$. It remains to use the definitons of the parameters in eqs. (30) and (31).

Furthermore, in the convex case, we define the following Lyapunov function:

$$\Psi_k = f(x_k) - f^* + c\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}, \tag{44}$$

where $c > 0$. We can upper-bound $\mathbb{E}[\Psi_{k+1}]$ as follows:

$$\begin{aligned}
\mathbb{E}[\Psi_{k+1}] &\overset{\text{(a)}}{\leq} \mathbb{E}[f(x_k) - f^*] - \eta\mathbb{E}[\mathcal{G}_{\eta/\beta}(x_k)] + 4\eta\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 \\
&\quad + c(1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] + 32c\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\overset{\text{(b)}}{\leq} (1-\beta)\mathbb{E}[f(x_k) - f^*] + c(1-\alpha)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] \\
&\quad + 4\eta\mathbb{E}[\|\nabla f(x_k) - m_k\|_{\mathbf{B}^{-1}}] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 32c\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\overset{\text{(c)}}{\leq} (1-\beta)\mathbb{E}[f(x_k) - f^*] + c(1-\alpha/2)\mathbb{E}[\|m_k - \nabla f(x_k)\|^2_{\mathbf{B}^{-1}}] + \tfrac{8\eta^2}{c\alpha} \\
&\quad + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 32c\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} \\
&\overset{\text{(d)}}{\leq} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 32c\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}} + \tfrac{8\eta^2}{c\alpha} \\
&\overset{\text{(e)}}{=} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + 2\sqrt{\tfrac{8\eta^2}{c\alpha}}\sqrt{32c\|\mathbf{M}\|^2_{\mathrm{tr}}\eta^2 + 8c\alpha^2\|\mathbf{\Sigma}\|^2_{\mathrm{tr}}} \\
&\overset{\text{(f)}}{\leq} \big(1 - \tfrac{1}{2}\min\{\alpha,\beta\}\big)\mathbb{E}[\Psi_k] + 2\|\mathbf{L}\|_{\mathrm{tr}}\eta^2 + \tfrac{32\eta^2}{\sqrt{\alpha}}\|\mathbf{M}\|_{\mathrm{tr}} + 16\sqrt{\alpha}\eta\|\mathbf{\Sigma}\|_{\mathrm{tr}} \\
&\overset{\text{(g)}}{\leq} \big(1 - \tfrac{1}{2}\beta\big)\mathbb{E}[\Psi_k] + \tfrac{1}{4}\beta\epsilon,
\end{aligned}$$

where (a) uses Lemmas 7 and 10; (b) uses the definition of η in eq. (32) and Lemma 4; (c) uses Young's inequality; (d) uses the definition of Ψ_k in eq. (44); (e) is obtained by minimizing in $c > 0$; (f) uses the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$; (g) uses the parameters in eq. (32). Hence, we can upper-bound $\mathbb{E}[f(x_K) - f^*]$ as follows:

$$\mathbb{E}[f(x_K) - f^*] \overset{\text{(a)}}{\leq} \mathbb{E}[\Psi_K] \overset{\text{(b)}}{\leq} \big(1 - \tfrac{1}{2}\beta\big)^K\mathbb{E}[\Psi_0] + \tfrac{1}{2}\epsilon \overset{\text{(c)}}{\leq} \epsilon,$$

where (a) uses the definition of Ψ_k in eq. (44); (b) uses the upper bound on $\mathbb{E}[\Psi_k]$ above and the inequality $\sum_{k=0}^{K-1}(1-\beta/2)^k \leq 2/\beta$; (c) uses the choice of K in eq. (33). □