OPTIMAL PROJECTION-FREE ADAPTIVE SGD FOR MATRIX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, Jiang et al. (2026) developed Leon, a practical variant of One-sided Shampoo (Xie et al., 2025a; An et al., 2025) algorithm for online convex optimization, which does not require computing a costly quadratic projection at each iteration. Unfortunately, according to the existing analysis, Leon requires tuning an additional hyperparameter in its preconditioner and cannot achieve dimension-independent convergence guarantees for convex optimization problems beyond the bounded gradients assumption. In this paper, we resolve this issue by proving certain stability properties of Leon's preconditioner. Using our improved analysis, we show that tuning the extra hyperparameter can be avoided and, more importantly, develop *the first practical variant of One-sided Shampoo with Nesterov acceleration*, which does not require computing projections at each iteration. As a side contribution, we obtain improved dimension-independent rates in the non-smooth non-convex setting and develop a unified analysis of the proposed algorithm, which yields accelerated projection-free adaptive SGD with (block-)diagonal preconditioners.

1 INTRODUCTION

Stochastic gradient methods with diagonal (i.e., coordinate-wise) preconditioning, such as AdaGrad (Duchi et al., 2011), Adam (Kingma & Ba, 2014), and AdamW (Loshchilov & Hutter, 2017), have been considered the default choice for more than a decade in most deep learning tasks (LeCun et al., 2015). However, recently, the focus of the research community has started to shift toward SGD algorithms with matrix preconditioning. The most notable examples of such algorithms are One-sided Shampoo (Xie et al., 2025a) (also known as ASGO (An et al., 2025)), which incorporates gradient accumulation into the preconditioning matrix, and Muon (Jordan et al., 2024), its accumulation-free alternative. The latter algorithm, Muon, has relatively cheap iterations and has already shown exceptional practical results in large-scale problems (Liu et al., 2025; Team et al., 2025). On the other hand, One-sided Shampoo remains a highly promising algorithm. From a practical perspective, it shows good preliminary experimental results (An et al., 2025) while maintaining almost the same iteration cost as Muon. Moreover, from a theoretical perspective, One-sided Shampoo provably converges in the case of non-smooth objective functions (An et al., 2025) and is provably universal, i.e., it can adapt to different levels of Hölder smoothness with the same choice of hyperparameters (Kovalev, 2026), thanks to the gradient accumulation in its preconditioning matrix. However, there are still important open questions related to One-sided Shampoo that we aim to address in this paper.

One-sided Shampoo vs AdaGrad. One-sided Shampoo was developed by Xie et al. (2025a) as a special instance of the adaptive meta-algorithm (Gupta et al., 2017), which, in turn, is closely related to the original AdaGrad. The key feature of AdaGrad is that it considers the geometric properties of the optimization problems (such as distances between iterates or Lipschitz/smoothness constants of the objective function) with respect to the infinity vector norm, $\|\cdot\|_\infty$, thanks to its diagonal preconditioning. This allows AdaGrad (and other diagonally preconditioned methods such as Adam/AdamW) to adapt to cases where the coordinates of the objective function have very different scales (Liu et al., 2024; Jiang et al., 2024). In contrast, One-sided Shampoo is used for solving optimization problems over a space of matrices and considers the geometric properties of the problems with respect to the spectral matrix norm, $\|\cdot\|_{\mathrm{op}}$, thanks to its more complex matrix preconditioner. This allows One-sided Shampoo to exploit more subtle geometric properties, such as low-rank gradients or approximately block-diagonal Hessians (An et al., 2025; Xie et al., 2025a). Such properties are observed in deep neural networks (Zhao et al., 2021; Yang et al., 2023; Cosson

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.

et al., 2023; Zhang et al., 2024a;b), which may explain the recent success of the accumulation-free variant of One-sided Shampoo, Muon (Kovalev & Borodich, 2025; Xie et al., 2025b).

Projection-free Methods. It is well known that both AdaGrad and One-sided Shampoo require the iterates to be bounded to provably converge (Gupta et al., 2017; Xie et al., 2025a; Kovalev, 2026). To ensure this, a projection onto the appropriate (infinity or spectral) norm ball is typically performed at each iteration. However, while in AdaGrad this projection step is equivalent to a simple coordinate-wise clipping of the iterates, One-sided Shampoo requires solving a complex auxiliary problem of the form $\min_{\|X'\|_{\mathrm{op}}\leq 1}\langle X'-X, Q(X'-X)\rangle$, where $X, X' \in \mathbb{R}^{m\times n}$, and $Q \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix. Recently, Jiang et al. (2026) proposed Leon, a practical variant of One-sided Shampoo based on the Follow-the-Regularized-Leader (FTRL) algorithm (Abernethy et al., 2009; 2016), that completely eliminates the necessity for the costly projection. Unfortunately, this algorithm suffers from an important drawback: it requires tuning an additional scalar hyperparameter $\delta > 0$ (see Section 3.1). According to the existing analysis, poor tuning of this parameter severely hurts the theoretical convergence guarantees and makes the resulting iteration complexity of Leon dimension-dependent. In addition, it requires the bounded gradients assumption, which significantly limits the range of applicability of Leon. Consequently, we arrive at the first main research question considered in this paper:

Q1. *Is it possible to develop a practical projection-free variant of One-sided Shampoo that maintains dimension-independent convergence guarantees beyond the bounded gradients assumption and avoids tuning extra hyperparameters?*

Nesterov acceleration. It is well known that vanilla GD requires $\mathcal{O}(\epsilon^{-2/(1+\nu)})$ iterations to find an ϵ-accurate solution to a convex ν-Hölder smooth minimization problem, where $\nu \in [0, 1]$ (Nesterov, 2015), and that this iteration complexity can be improved to $\mathcal{O}(\epsilon^{-2/(1+3\nu)})$ with the help of Nesterov momentum (Nesterov, 1983; Nesterov et al., 2018). Similarly, the convergence of adaptive gradient methods can be accelerated with Nesterov momentum. In particular, an accelerated version of AdaGrad was developed by Kovalev (2026), and an accelerated variant of One-sided Shampoo was developed by Xie et al. (2025b). Unfortunately, the accelerated One-sided Shampoo of Xie et al. (2025b) is not practical, as it still requires performing the complex projection step at each iteration. Consequently, we arrive at the second main research question considered in this paper:

Q2. *Is it possible to develop a practical projection-free variant of One-sided Shampoo with Nesterov acceleration?*

Summary of contributions. In this paper, we provide positive answers to Questions 1 and 2 and make the following contributions:

- **(i)** We show that the preconditioning operator of Leon satisfies a certain *gradient stability property* in Lemma 4. Using this property, we develop an improved analysis of the FTRL algorithm with Leon's preconditioning, which shows that the parameter δ can be set arbitrarily small. Therefore, we eliminate the necessity of tuning this parameter and obtain dimension-independent convergence guarantees for Leon in online and non-smooth non-convex optimization. Refer to Section 3 for details.
- **(ii)** We develop an accelerated version of the FTRL algorithm with Leon-like preconditioning, in which we incorporate the accumulation of squared gradient distances in a UniXGrad-like fashion (Kavis et al., 2019; Rodomanov et al., 2024). In particular, we improve the analysis of UniXGrad to accommodate general preconditioners, whereas the previous analyses of UniXGrad were limited to scalar stepsizes. As a result, we develop the first accelerated and practical projection-free variant of One-sided Shampoo, which achieves the optimal $\mathcal{O}(\epsilon^{-2/(1+3\nu)})$ iteration complexity for convex ν-Hölder smooth optimization. Refer to Section 4 for details.
- **(iii)** As a side contribution, we provide a unified analysis of Leon and its accelerated version for optimization over general Euclidean spaces $\mathcal{X}$. Therefore, our accelerated projection-free adaptive SGD algorithm is not limited to matrix preconditioners but can also incorporate a wide range of preconditioning operators, including diagonal preconditioners and scalar stepsizes.

Notation. In this paper, we use the following notation: $\dim(\mathcal{X})$ is the dimension of the space $\mathcal{X}$; $\mathbb{L}$ is the space of linear operators $\mathcal{X} \to \mathcal{X}$; $\mathbf{I} \in \mathbb{L}$ and $\mathbf{O} \in \mathbb{L}$ denote the identity and the zero

operators, respectively; for arbitrary operator $\mathbf{A} \in \mathbb{L}$, $\mathbf{A}^* \in \mathbb{L}$ denotes its adjoint operator; $\mathbb{S} \subset \mathbb{L}$ is the space of self-adjoint linear operators, $\mathbb{S}_{++}, \mathbb{S}_+ \subset \mathbb{S}$ are the spaces of positive definite and positive semi-definite self-adjoint operators, respectively; $\prec, \preceq, \succ, \succeq$ denote the standard Löwner order on $\mathbb{S}$; $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the standard inner product and the Euclidean norm on $\mathcal{X}$ or $\mathbb{L}$, depending on the context, in particular, $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{A}\mathbf{B}^*)$ for $\mathbf{A}, \mathbf{B} \in \mathbb{L}$, where $\operatorname{tr}(\cdot)$ is the standard trace functional on $\mathbb{L}$ induced by the inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{X}$; for arbitrary $\mathbf{H} \in \mathbb{S}_{++}$, $\|\cdot\|_{\mathbf{H}}$ denotes the weighted Euclidean norm on $\mathcal{X}$, i.e., $\|x\|_{\mathbf{H}}^2 = \langle x, \mathbf{H}x \rangle$ for $x \in \mathcal{X}$; $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{tr}}$ denote the operator and trace norm on $\mathbb{L}$, respectively, i.e., $\|\mathbf{A}\|_{\mathrm{op}} = \max_{\|x\| \leq 1} \|\mathbf{A}x\|$ and $\|\mathbf{A}\|_{\mathrm{tr}} = \operatorname{tr}(\sqrt{\mathbf{A}\mathbf{A}^*})$ for all $\mathbf{A} \in \mathbb{L}$; for arbitrary $y, z \in \mathcal{X}$, by $z\langle y, \cdot \rangle \in \mathbb{L}$ we denote the rank-1 linear operator $x \mapsto z\langle y, x \rangle$; $\mathbf{out}(z) \in \mathbb{S}$ denotes a symmetric rank-1 linear operator $z\langle z, \cdot \rangle$; $\mathbb{N}$ and $\mathbb{N}_0$ denote the set of positive and non-negative integers; $\mathbb{B}_h(\cdot\,;\cdot)$ denotes the Bregman divergence associated with the function $h(\cdot)$. We also use the convention $0^0 = 1$; whenever an expression defines a function $\psi(t)$ for $t > 0$ that admits a continuous extension to $t = 0$, we set $\psi(0) = \lim_{t \to +0} \psi(t)$.

2 Preliminaries

Leon is a special instance of the FTRL algorithm for online convex optimization. Given a sequence of strongly convex closed regularization functions $\Psi_k(x) \colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and a sequence of linear maps $x \mapsto \langle g_k, x \rangle$, where $g_k \in \mathcal{X}$ are the past gradients, the update rule for FTRL is given as follows:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \Psi_k(-x) + \langle m_k, x \rangle, \quad \text{where} \quad m_k = \sum_{i=0}^{k} g_i. \tag{1}$$

From standard convex analysis, it follows that the Fenchel conjugate functions $\Psi_k^*(m) \colon \mathcal{X} \to \mathbb{R}$ are convex and smooth, and the iterations in eq. (1) can be written in a simplified form:

$$x_{k+1} = -\nabla \Psi_k^*(m_k). \tag{2}$$

Inspired by the choice of the regularization function proposed by Jiang et al. (2026) for Leon, we will further use the following choice of $\Psi_k^*(m)$:

$$\Psi_k^*(m) = \eta \| \sqrt{\operatorname{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k} \|_{\mathrm{tr}}, \tag{3}$$

where $\eta > 0$ is a positive parameter, $\mathbf{S}_k \in \mathcal{H} \cap \mathbb{S}_{++}$ is a positive definite operator, and $\mathcal{H} \subset \mathbb{S}$ is a certain linear subspace of self-adjoint operators. The linear subspace $\mathcal{H}$ in this definition is used for the purpose of unified analysis of the resulting algorithm. It is inspired by the unified analysis of the adaptive meta-algorithm by Gupta et al. (2017), where they restricted the choice of the preconditioners to the space $\mathcal{H}$, e.g., diagonal and matrix preconditioners, as well as scalar stepsizes. Of course, to conduct the unified analysis, it is necessary to impose specific assumptions on the space $\mathcal{H}$, e.g., to ensure that the operator under the square root in eq. (3) is positive definite. There are two closely related sets of assumptions on $\mathcal{H}$ proposed by Xie et al. (2025a;b) and Kovalev & Borodich (2025); Kovalev (2026). We adhere to the latter approach with the following Assumption 1.

Assumption 1. *The linear subspace of operators $\mathcal{H} \subset \mathbb{S}$ satisfies the following properties:*

- **(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: $\operatorname{proj}_{\mathcal{H}}(\mathbf{H}) \in \mathbb{S}_{++}$ for all $\mathbf{H} \in \mathbb{S}_{++}$.*

Following Kovalev & Borodich (2025, Lemma 1), we define the primal norm $\rho(\cdot) \colon \mathcal{X} \to \mathbb{R}_+$ and the dual norm $\rho_*(\cdot) \colon \mathcal{X} \to \mathbb{R}_+$, which are non-Euclidean in general, as follows:

$$\rho(x) = \| \sqrt{\operatorname{proj}_{\mathcal{H}}(\mathbf{out}(x))} \|_{\mathrm{op}} \quad \text{and} \quad \rho_*(x) = \| \sqrt{\operatorname{proj}_{\mathcal{H}}(\mathbf{out}(x))} \|_{\mathrm{tr}}. \tag{4}$$

For instance, One-sided Shampoo and Leon assume $\mathcal{X} = \mathbb{R}^{m \times n}$ and $\rho(\cdot)$ is the spectral matrix norm (up to constants), and diagonal AdaGrad assumes $\mathcal{X} = \mathbb{R}^d$ and $\rho(\cdot)$ is the infinity vector norm. Refer to Kovalev & Borodich (2025); Kovalev (2026); Xie et al. (2025a;b) for more examples. These norms also play an important role in the description of the geometric properties of optimization problems that we will consider in this paper. In particular, we will consider online linear optimization problems,

Algorithm 1 Generalized Leon for Online Optimization

input: $x_0 = 0 \in \mathcal{X}$
parameters: $K \in \mathbb{N}$, $\delta > 0$, $\eta \leq \mathcal{R}$
$\mathbf{S}_{-1} = \delta^2 \mathbf{I}, m_{-1} = 0$
for $k = 0, \dots, K$ **do**
  observe gradient $g_k \in \mathcal{X}$, $m_k = m_{k-1} + g_k$
  $\mathbf{S}_k = \mathbf{S}_{k-1} + \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(g_k))$
  $x_{k+1} = -\nabla \Psi_k^*(m_k)$ ▷ $\Psi_k^*(m)$ *is defined in eq.* (3)
output: $(x_0, \dots, x_K) \in Q_{\mathcal{R}}^{K+1}$

where the decision set $Q_{\mathcal{R}} = \{x \in \mathcal{X} : \rho(x) \leq \mathcal{R}\}$ is the ball of radius $\mathcal{R} > 0$ with respect to the norm $\rho(\cdot)$. The regret for such online problems can be written as follows:

$$\mathrm{Reg}_K = \left[\max_{x \in Q_{\mathcal{R}}} \sum_{k=0}^{K} \langle g_k, x_k - x \rangle \right], \tag{5}$$

where each decision point $x_k \in Q_{\mathcal{R}}$ is selected using only the past gradients $g_0, \dots, g_{k-1}$; refer to Orabona & Pál (2018) for more details on the online learning setting. We will also discuss the geometric properties of optimization problems related to Assumption 1 in Sections 3.3 and 4.1.

3 Improved and Unified Analysis of Leon

3.1 Unified Leon and Existing Convergence Results

We start with computing the gradient of the dual regularization function $\Psi_k^*(m)$ in the following Lemma 1, which generalizes the discussion of Jiang et al. (2026, Section 4.2) for the matrix optimization setting, $\mathcal{X} = \mathbb{R}^{m \times n}$.

Lemma 1 (↓). *The function $\Psi_k^*(m)$ is convex and differentiable, and its gradient is given as follows:*

$$\nabla \Psi_k^*(m) = \eta [\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k]^{-1/2} m. \tag{6}$$

Next, Jiang et al. (2026, Theorem 4.2) derived the feasibility, dominance, upper stability, and smoothness properties of the dual regularization function $\Psi_k^*(m)$ in the matrix optimization setting. We generalize these properties to the case of general $\mathcal{X}$ under Assumption 1 in the following Lemma 2.

Lemma 2 (↓). *Let $\mathbf{S}_{k+1} \succeq \mathbf{S}_k$. Then, for all $m, m' \in \mathcal{X}$, the following inequalities hold:*

(i) *Feasibility:* $\rho(\nabla \Psi_k^*(m)) \leq \eta$,

(ii) *Dominance:* $\eta \rho_*(m) \leq \Psi_k^*(m)$,

(iii) *Upper stability:* $0 \leq \Psi_{k+1}^*(m) - \Psi_k^*(m) \leq \eta \left[\|\sqrt{\mathbf{S}_{k+1}}\|_{\mathrm{tr}} - \|\sqrt{\mathbf{S}_k}\|_{\mathrm{tr}} \right]$,

(iv) *Smoothness:* $\mathbb{B}_{\Psi_k^*}(m; m') \leq \frac{1}{2} \eta \|m - m'\|_{\mathbf{S}_k^{-1/2}}^2$.

Now, we are ready to present Leon for solving online problems of the form (5), which is formalized as Algorithm 1 and is unified according to the discussion in Section 2. Note that the iterations of Algorithm 1 coincide with the FTRL iterations in eqs. (1) and (2), where we use the squared gradient accumulation operator $\mathbf{S}_k = \delta^2 \mathbf{I} + \sum_{i=0}^{k} \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(g_i))$ in the definition of the dual regularization function $\Psi_k^*(m)$ in eq. (3). Also, note that the feasibility condition holds: $x_0, \dots, x_K \in Q_{\mathcal{R}}$, due to Lemma 2. We also restate the main result for this algorithm as Theorem 1. Note that Theorem 1 is also presented in a generalized manner under Assumption 1, similar to the rest of this paper.

Theorem 1 (Generalization of Theorem 3.2, Jiang et al. (2026)). *Let $\rho(g_k) \leq \delta$ for all $k \in \mathbb{N}_0$ and let $\eta = \mathcal{R}$. Then, the output of Algorithm 1 satisfies the following inequality:*

$$\mathrm{Reg}_K \leq \mathcal{O}\left[\delta \mathcal{R} \dim(\mathcal{X}) + \mathcal{R} \rho_*(g_0) + \mathcal{R} \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} \right]. \tag{7}$$

Now, we discuss the key problems with the result in Theorem 1, which are essential to solve in order to provide a positive answer to Question 1. First, Theorem 1 requires the gradients to be

uniformly bounded. This poses issues in the stochastic optimization case, where we typically assume the boundedness of the second moments of the gradients, which does not imply that the gradients are almost surely bounded. Second, the presence of the term $\delta\mathcal{R}\dim(\mathcal{X})$ in the regret upper bound prevents us from obtaining dimension-independent complexities for solving smooth convex problems and from accelerating the convergence of the algorithm using Nesterov momentum.

3.2 Improved Analysis of LEON and FTRL

First, for the FTRL algorithm in eq. (2), instead of bounding the standard regret defined in eq. (5), we will bound the regret-like quantity Reg_K^+ defined as follows:

$$\mathrm{Reg}_K^+ = \max_{x\in Q_{\mathcal{R}}} \left[\sum_{k=0}^{K} \langle g_k, x_{k+1} - x\rangle\right]. \tag{8}$$

We obtain the key identity for the regret-like quantity Reg_K^+ in the following Lemma 3. In contrast to the analogous identity in Lemma 3.1 of Jiang et al. (2026), the identity in Lemma 3 contains negative Bregman divergence terms $-\mathbb{B}_{\Psi_k^*}(m_{k-1}; m_k)$, which we will be able to utilize by applying the smoothness property in Lemma 2.

Lemma 3 (↓). *Let $\eta = \mathcal{R}$. Then, the iterations in eq. (2) satisfy the following relation:*

$$\begin{aligned}\mathrm{Reg}_K^+ &= \Psi_0^*(0) + \eta\rho_*(m_K) - \Psi_K^*(m_K) \\ &\quad + \sum_{k=1}^{K}\left[\Psi_k^*(m_{k-1}) - \Psi_{k-1}^*(m_{k-1})\right] - \sum_{k=0}^{K}\mathbb{B}_{\Psi_k^*}(m_{k-1}; m_k),\end{aligned} \tag{9}$$

Next, we establish the key property of the dual regularization function $\Psi_k^*(m)$, which we call gradient stability, in the following Lemma 4.

Lemma 4 (↓). *Let $\mathbf{S}_{k+1} \succeq \mathbf{S}_k$. Then, for all $m \in \mathcal{X}$, the following inequality holds:*

$$\|\nabla\Psi_{k+1}^*(m) - \nabla\Psi_k^*(m)\|_{\mathbf{S}_{k+1}^{1/2}}^2 \le \eta^2\left[\|\sqrt{\mathbf{S}_{k+1}}\|_{\mathrm{tr}} - \|\sqrt{\mathbf{S}_k}\|_{\mathrm{tr}}\right]. \tag{10}$$

Now, using the previously established properties in Lemma 2, the new property in Lemma 4, and the negative Bregman divergence terms $-\mathbb{B}_{\Psi_k^*}(m_{k-1}; m_k)$, we are ready to establish the main upper bound on the regret-like quantity Reg_K^+ in the following Theorem 2.

Theorem 2 (↓). *Let $\eta = \mathcal{R}$. Let $\mathbf{S}_{k+1} \succeq \mathbf{S}_k$ for all $k \in \mathbb{N}_0$. Let $g_0, \ldots, g_K \in \mathcal{X}$ and $x_0 = 0$. Let $x_1, \ldots, x_{K+1}$ be generated according to eqs. (2) and (3). Then, the following inequality holds:*

$$\mathrm{Reg}_K^+ \le \Psi_0^*(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} - \sum_{k=0}^{K}\tfrac{1}{4\eta}\|x_{k+1} - x_k\|_{\mathbf{S}_k^{1/2}}^2. \tag{11}$$

In the following section, we show how to use the upper-bound in Theorem 2 for obtaining improved guarantees in online convex and non-smooth non-convex optimization.

3.3 Application to Online and Non-smooth Non-convex Optimization

Online optimization. We start with the general improved result for Algorithm 1 for solving online convex optimization, as stated in the following Theorem 3.

Theorem 3 (↓). *Let $\eta = \mathcal{R}$ and $\delta > 0$. Then, the output of Algorithm 1 satisfies the inequality*

$$\mathrm{Reg}_K \le \mathcal{O}\left[\delta\mathcal{R}\dim(\mathcal{X}) + \mathcal{R}\rho_*(g_0) + \mathcal{R}\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right]. \tag{12}$$

The upper bound on the regret in the improved Theorem 3 coincides with the inequality in Theorem 1 by Jiang et al. (2026). However, we significantly simplify the assumptions required to achieve this inequality. First, Theorem 3 does not require the gradients g_k to be uniformly bounded. Second, we do not have to tune the parameter $\delta > 0$, as we can simply choose an arbitrarily small value for δ, thus obtaining dimension-independent guarantees in Theorem 3.[1]

[1] In fact, the results in Theorems 3 and 4 hold true even with $\delta = 0$. Refer to Section A.

Non-smooth non-convex optimization. Now, we show, how to apply the improved result for solving optimization problems of the form

$$\min_{x\in\mathcal{X}} f(x), \tag{13}$$

where $f(x)\colon \mathcal{X}\to\mathbb{R}$ is a differentiable objective function. Our goal is to find an expected-(γ,ϵ)-stationary point $\hat{x}\in\mathcal{X}$, i.e., satisfying the following Definition 1.

Definition 1. *A random point $\hat{x}\in\mathcal{X}$ is called expected-(γ,ϵ)-stationary if $\mathbb{E}[s_\gamma(\hat{x})]\le\epsilon$, where*

$$s_\gamma(\hat{x}) = \inf_{\mathcal{P}}[\rho_*(\mathbb{E}_{z\sim\mathcal{P}}[\nabla f(z)])] \quad \text{s.t.} \quad \begin{array}{l}\mathcal{P}-\text{probability distribution over } \mathcal{X},\\ \mathbb{E}_{z\sim\mathcal{P}}[z]=\hat{x}, \quad \mathbb{E}_{z\sim\mathcal{P}}[\rho(z-\hat{x})]\le\gamma.\end{array} \tag{14}$$

We are naturally interested in solving the problem using a stochastic algorithm. Hence, we assume the existence of an appropriate stochastic gradient estimator in the following Assumption 2.

Assumption 2. *There exists a stochastic estimator $\nabla_\xi f(x)$ of the gradient $\nabla f(x)$, where $\xi\sim\mathcal{D}$ is a random variable sampled from the distribution $\mathcal{D}$, which satisfies the following properties:*

- **(i)** *Unbiased estimator:* $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla_\xi f(x)]=\nabla f(x)$.
- **(ii)** *2nd moment bound:* $\mathbb{E}_{\xi\sim\mathcal{D}}[\|\nabla_\xi f(x)\|^2_{\mathbf{B}^{-1}}]\le\mathcal{G}^2$, *where* $\mathbf{B}\in\mathbb{S}_{++}\cap\mathcal{H}$, $\|\mathbf{B}\|_{\mathrm{tr}}=1$, $\mathcal{G}>0$.

Now, we can state the complexity result for Algorithm 1 for solving non-smooth non-convex minimization problems. It is obtained with the help of the O2NC framework (Zhang & Cutkosky, 2024). The proof and the resulting algorithm are almost identical to those described by Jiang et al. (2026, Section 5, Appendix G); hence, we will not describe them in this paper. Note that, similar to the online optimization case, we obtain improved dimension-independent guarantees in Theorem 4 by choosing a very small value of $\delta>0$.[1]

Theorem 4. *For a radius $\gamma>0$ and a sufficiently small accuracy $\epsilon>0$, Algorithm 1 with the O2NC framework requires the following number of iterations to find an expected-(γ,ϵ)-stationary point:*

$$\mathcal{O}\big[\Delta\mathcal{G}^2\cdot\gamma^{-1}\epsilon^{-3}+\mathcal{G}^2\cdot\epsilon^{-2}+\delta\dim(\mathcal{X})\cdot\epsilon^{-1}\big], \tag{15}$$

where $\Delta=f(x_0)-\inf_x f(x)$ is the initial function gap.

On the geometry of Assumption 2. Note that Assumption 2 is given in terms of the weighted Euclidean norm $\|\cdot\|_{\mathbf{B}^{-1}}$. In the case of matrix optimization, $\mathcal{X}=\mathbb{R}^{m\times n}$, with the standard left-preconditioning subspace $\mathcal{H}=\{G\mapsto BG : B\in\mathbb{R}^{m\times m}, B=B^\top\}$, it is convenient to use this assumption for modeling low-rank gradients, which are often observed in deep neural networks. Indeed, condition **(ii)** in Assumption 2 with $\mathcal{G}=\sqrt{n}\,\mathrm{tr}(C)$ and $\mathbf{B}(G)=\frac{CG}{n\,\mathrm{tr}(C)}$ is implied by the inequality $\mathbb{E}_{\xi\sim\mathcal{D}}[G_\xi G_\xi^\top]\preceq C^2$, where $G_\xi\in\mathbb{R}^{m\times n}$ is a stochastic gradient and $C\in\mathbb{R}^{m\times m}$ is a symmetric positive definite matrix with low effective rank;[2] see also the work of An et al. (2025, Section 4). In addition, Assumption 2 is also closely related to its non-Euclidean analogue: it implies the inequality $\mathbb{E}_{\xi\sim\mathcal{D}}[\rho_*^2(\nabla_\xi f(x))]\le\mathcal{G}^2$, and the optimization under non-Euclidean geometry assumptions has attracted a lot of interest in the research community (Bernstein & Newhouse, 2024; Kovalev, 2025; Pethick et al., 2025; Kovalev & Borodich, 2025; Xie et al., 2025b). Also note that, strictly speaking, the constant $\mathcal{G}$ and the induced norms $\rho(X)=\|X\|_{\mathrm{op}}/\sqrt{n}$ and $\rho_*(X)=\sqrt{n}\|X\|_{\mathrm{tr}}$ contain dimension-dependent factors $n^{\pm1/2}$. However, these factors cancel under the consistent rescaling $(\mathcal{G},\epsilon,\gamma)=(\sqrt{n}\mathcal{G}',\sqrt{n}\epsilon',\gamma'/\sqrt{n})$, where $(\mathcal{G}',\epsilon',\gamma')$ are the corresponding parameters for the unscaled spectral/nuclear norm pair, and therefore introduce no additional dimension dependence in Definition 1, Theorem 4, and further results.

4 Generalized Leon with Nesterov Acceleration for Convex Problems

4.1 Problem and Assumptions

In this section, we focus on solving convex optimization problems of the form

$$f^*=\min_{x\in Q_{\mathcal{R}}} f(x). \tag{16}$$

[2] By the effective rank of a symmetric positive definite matrix C, we mean $\mathrm{tr}(C)/\lambda_{\max}(C)$.

Algorithm 2 Generalized Leon with Nesterov Acceleration

input: $x_0 = \overline{x}_0 = 0 \in Q_{\mathcal{R}}$
parameters: $K \in \mathbb{N}$, $\delta > 0$, $\eta \leq \mathcal{R}$, $\alpha_0, \ldots, \alpha_K \geq 1$
$\mathbf{S}_0 = \delta^2 \mathbf{I}$, $m_{-1} = 0$
for $k = 0, \ldots, K$ **do**
  sample $\xi_k, \xi'_{k+1} \sim \mathcal{D}$
  $g_k = \nabla_{\xi_k} f_k(x_k)$, $m_k = m_{k-1} + g_k$ ▷ $f_k(x)$ *is defined in eq.* (17)
  $x_{k+1} = -\nabla \Psi_k^*(m_k)$ ▷ $\Psi_k^*(m)$ *is defined in eq.* (3)
  $\overline{x}_{k+1} = x_{k+1}/\alpha_k + (1 - 1/\alpha_k)\overline{x}_k$
  $\tilde{g}_{k+1} = \nabla_{\xi'_{k+1}} f_k(x_{k+1})$ ▷ $f_k(x)$ *is defined in eq.* (17)
  $\mathbf{S}_{k+1} = \mathbf{S}_k + \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(\tilde{g}_{k+1} - g_k))$
output: $\overline{x}_{K+1} \in Q_{\mathcal{R}}$

We start by imposing the formal Assumptions 3 and 4 on the objective function $f(x)$. These assumptions can be seen as a generalization of the previously discussed Assumption 2 to the case of stochastic Hölder smooth convex optimization. Consequently, these assumptions can also be used to model low-rank gradients in the matrix optimization case, as discussed in Section 3.3. It is also important to highlight that, in the matrix case, the constraint $x \in Q_{\mathcal{R}}$ implies that the solution is likely to be high-rank, as discussed by Chen et al. (2025), which can be beneficial, as discussed by An et al. (2025, Section 5). Additionally, in the general case, Assumption 3 implies Hölder smoothness of the objective function $f(x)$ with respect to the non-Euclidean norm $\rho(\cdot)$, which makes it relevant to non-Euclidean optimization (Bernstein & Newhouse, 2024; Kovalev, 2025; Pethick et al., 2025; Kovalev & Borodich, 2025).

Assumption 3. *Function $f(x)$ is convex and $(\mathcal{L}, \nu)$-Hölder smooth with respect to the norm $\|\cdot\|_{\mathbf{B}}$ (see Nesterov (2015, eq. (1.1))), where $\mathcal{L} > 0$, $\nu \in [0, 1]$, $\mathbf{B} \in \mathbb{S}_{++} \cap \mathcal{H}$ and $\|\mathbf{B}\|_{\mathrm{tr}} = 1$.*

Assumption 4. *There exists a stochastic estimator $\nabla_\xi f(x) = n_\xi(x) + \nabla f(x)$ of the (sub)gradient $\nabla f(x) \in \partial f(x)$, where $n_\xi(x)$ is the noise and $\xi \sim \mathcal{D}$ is a random variable sampled from the distribution $\mathcal{D}$. The noise $n_\xi(x)$ satisfies the following properties:*

- **(i)** *Zero mean:* $\mathbb{E}_{\xi \sim \mathcal{D}}[n_\xi(x)] = 0$.
- **(ii)** *Variance bound:* $\mathbb{E}_{\xi \sim \mathcal{D}}[\|n_\xi(x)\|^2_{\mathbf{B}^{-1}}] \leq \sigma^2$, *where* $\mathbf{B} \in \mathbb{S}_{++} \cap \mathcal{H}$, $\|\mathbf{B}\|_{\mathrm{tr}} = 1$, *and* $\sigma > 0$.

4.2 Optimal Projection-Free Adaptive SGD with Preconditioning

Now, we are ready to present the accelerated version of Leon with Nesterov acceleration, formalized as Algorithm 2. In order to incorporate Nesterov acceleration into the algorithm, we follow the approach of Kovalev & Borodich (2024). That is, at each iteration k, we apply the FTRL step in eq. (2) to the modified objective function $f_k(x) \colon \mathcal{X} \to \mathbb{R}$, which is defined as follows:

$$f_k(x) = \alpha_k^2 f(x/\alpha_k + (1 - 1/\alpha_k)\overline{x}_k), \quad \text{where} \;\; \alpha_k \geq 1, \;\; \overline{x}_k \in \mathcal{X}, \tag{17}$$

Consequently, using our previously obtained Theorem 2, we obtain the following Lemma 5.

Lemma 5 (↓). *Let $\eta = \mathcal{R}$ and $\delta > 0$. Then, for all $x \in Q_{\mathcal{R}}$, the following inequality holds:*

$$\sum_{k=0}^{K} \mathbb{E}[\mathbb{B}_{f_k}(x_k; x_{k+1}) + f_k(x_{k+1}) - f_k(x)] \leq \mathbb{E}\left[\delta \mathcal{R} \dim(\mathcal{X}) + \tfrac{9}{2}\mathcal{R}\|\sqrt{\mathbf{S}_{K+1}}\|_{\mathrm{tr}}\right]. \tag{18}$$

The reference point $\overline{x}_k$ is updated using line 8, which allows us to obtain the following Lemma 6.

Lemma 6 (↓). *Let $\alpha_0 = 1$ and $\alpha_k^2 \leq \alpha_k + \alpha_{k-1}^2$ for all $k \in \mathbb{N}$. Let $x^* \in Q_{\mathcal{R}}$ be a solution to problem (16). Then, the following inequality holds:*

$$\alpha_K^2[f(\overline{x}_{K+1}) - f^*] \leq \sum_{k=0}^{K}[f_k(x_{k+1}) - f_k(x^*)]. \tag{19}$$

Furthermore, we make another important modification to the algorithm. It is related to the fact that the convergence of standard AdaGrad-like algorithms relies on bounding the squared gradients using

the objective function optimality gap (Orabona, 2023, Corollary 1):

$$\|\nabla f(x_k)\|_{\mathbf{B}^{-1}}^{1+\nu} \leq \left[\frac{1+\nu}{\nu}\right]^{\nu} \cdot \mathcal{L} \cdot [f(x_k) - \min_{x \in \mathcal{X}} f(x)]^{\nu}. \tag{20}$$

However, the minimizer of the objective function $f(x)$ does not coincide with the minimizer of the modified function $f_k(x)$, even in the unconstrained optimization case, which prevents us from using this inequality in an accelerated algorithm. To tackle this issue, we accumulate squared gradient differences in the operator $\mathbf{S}_k$ in line 10, instead of merely accumulating squared gradients, similar to the UniXGrad algorithm. In addition, we can bound the squared gradient differences with the Bregman divergences as follows:

$$\|\nabla f_k(x_{k+1}) - \nabla f_k(x_k)\|_{\mathbf{B}^{-1}}^{1+\nu} \leq \left[\frac{1+\nu}{\nu}\right]^{\nu} \cdot \mathcal{L}\alpha_k^{1-\nu} \cdot [\mathbb{B}_{f_k}(x_k; x_{k+1})]^{\nu}. \tag{21}$$

Consequently, we can bound the squared gradient differences by utilizing the negative Bregman divergence term in Lemma 5. This improves upon the previous analyses by Kavis et al. (2019); Rodomanov et al. (2024), who did not properly utilize this term and achieved convergence only for scalar stepsizes. In accordance with this discussion, we obtain the following Lemma 7.

Lemma 7 (↓). *Let $\alpha_k \geq \alpha_{k-1}$ for all $k \in \mathbb{N}$. Then, for all $c \geq 1$, the following inequality holds:*

$$\begin{aligned}\mathbb{E}\left[\|\sqrt{\mathbf{S}_{K+1}}\|_{\mathrm{tr}}\right] \leq \delta \dim(\mathcal{X}) + \mathbb{E}\left[\frac{1}{c\eta}\sum_{k=0}^{K} \mathbb{B}_{f_k}(x_k; x_{k+1})\right] \\ + 2\sigma\left[[K+1]\alpha_K^2\right]^{\frac{1}{2}} + c\mathcal{L}\eta^{\nu}\left[[K+1]\alpha_K^2\right]^{\frac{1-\nu}{2}}.\end{aligned} \tag{22}$$

Now, it remains to combine Lemmas 5, 6 and 7 in order to derive the final convergence result for Algorithm 2, as stated in the following Theorem 5.

Theorem 5 (↓). *Let $\eta = \mathcal{R}$ and $\delta > 0$, and let $\alpha_k = 1 + k/2$. Let $x^* \in Q_{\mathcal{R}}$ be a solution to problem (16). Then, the output of Algorithm 2 satisfies the following inequality:*

$$\mathbb{E}[f(\overline{x}_{K+1}) - f(x^*)] \leq \mathcal{O}\left[\frac{\mathcal{L}\mathcal{R}^{1+\nu}}{[K+1]^{(3\nu+1)/2}} + \frac{\sigma\mathcal{R}}{[K+1]^{1/2}} + \frac{\delta\mathcal{R}\dim(\mathcal{X})}{[K+1]^2}\right]. \tag{23}$$

Hence, to reach the precision $\mathbb{E}[f(\overline{x}_{K+1}) - f^*] \leq \epsilon$, it is sufficient to perform the following number of iterations of Algorithm 2:

$$K+1 = \mathcal{O}\left[\left[\frac{\mathcal{L}\mathcal{R}^{1+\nu}}{\epsilon}\right]^{\frac{2}{1+3\nu}} + \left[\frac{\sigma\mathcal{R}}{\epsilon}\right]^2 + \left[\frac{\delta\mathcal{R}\dim(\mathcal{X})}{\epsilon}\right]^{\frac{1}{2}}\right]. \tag{24}$$

Theorem 5 implies the optimal complexity $\mathcal{O}(\epsilon^{-2/(3\nu+1)})$ for non-stochastic ν-Hölder smooth convex problems and the optimal complexity $\mathcal{O}(\epsilon^{-2})$ for stochastic convex optimization problems (Nemirovskij & Yudin, 1983). In addition, the complexity result in Theorem 5 is dimension-independent by choosing a very small value of $\delta > 0$.[3] Moreover, it involves the constants $\sigma, \mathcal{L} > 0$ from Assumptions 3 and 4, and the radius $\mathcal{R} > 0$ of the constraint set $Q_{\mathcal{R}}$ with respect to the non-Euclidean norm $\rho(\cdot)$. Therefore, in the matrix optimization case, our Algorithm 2 can exploit low-rank gradients and a high-rank solution to outperform its Euclidean counterparts, just like the original non-accelerated One-sided Shampoo (An et al., 2025, Section 5).

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.

[3] The result in Theorem 5 holds true even with $\delta = 0$. Refer to Section A.

REFERENCES

Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Perturbation techniques in online learning and optimization. *Perturbations, Optimization, and Statistics*, 233:17, 2016.

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, number 110, 2009.

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.

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

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

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

Romain Cosson, Ali Jadbabaie, Anuran Makur, Amirhossein Reisizadeh, and Devavrat Shah. Low-rank gradient descent. *IEEE Open Journal of Control Systems*, 2:380–395, 2023.

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

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

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of adagrad over sgd: Upper and lower bounds in stochastic non-convex optimization. *arXiv preprint arXiv:2406.04592*, 2024.

Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive matrix online learning through smoothing with guarantees for nonsmooth nonconvex optimization. *arXiv preprint arXiv:2602.08232*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Ali Kavis, Kfir Y Levy, Francis Bach, and Volkan Cevher. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *Advances in neural information processing systems*, 32, 2019.

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

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

Dmitry Kovalev. Sgd with adaptive preconditioning: Unified analysis and momentum acceleration. In *International Conference on Learning Representations*, volume 2026, pp. 90473–90497, 2026.

Dmitry Kovalev and Ekaterina Borodich. On linear convergence in smooth convex-concave bilinearly-coupled saddle-point optimization: Lower bounds and optimal algorithms. *arXiv preprint arXiv:2411.14601*, 2024.

Dmitry Kovalev and Ekaterina Borodich. Non-euclidean sgd for structured optimization: Unified analysis and improved rates. *arXiv preprint arXiv:2511.11466*, 2025.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

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

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

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.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Dokl. Akad. Nauk. SSSR*, 269(3):543, 1983.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Francesco Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018.

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.

Anton Rodomanov, Xiaowen Jiang, and Sebastian Stich. Universality of adagrad stepsizes for stochastic optimization: Inexact oracle, acceleration and variance reduction. *Advances in Neural Information Processing Systems*, 37:26770–26813, 2024.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 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*, 2025a.

Shuo Xie, Tianhao Wang, Beining Wu, and Zhiyuan Li. A tale of two geometries: Adaptive optimizers and non-euclidean descent. *arXiv preprint arXiv:2511.20584*, 2025b.

Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. *arXiv preprint arXiv:2405.09742*, 2024.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. *Advances in neural information processing systems*, 37:131786–131823, 2024a.

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

Jiawei Zhao, Florian Schäfer, and Anima Anandkumar. Zero initialization: Initializing neural networks with only zeros and ones. *arXiv preprint arXiv:2110.12661*, 2021.

A Additional Details

A.1 Structural Properties of the Preconditioner Space

In Assumption 1, we require the space of preconditioning operators $\mathcal{H}$ to be closed under symmetric multiplications. However, throughout this paper, we will use a stronger property that $\mathbf{H}^p \in \mathcal{H} \cap \mathbb{S}_{++}$ for all $\mathbf{H} \in \mathcal{H} \cap \mathbb{S}_{++}$ and $p \in \{\pm 1/2, -1\}$. This is implied by the following Lemma 8.

Lemma 8 (↓). *Under Assumption 1, the space $\mathcal{H}$ is closed under arbitrary operator functions (Kovalev, 2026, Definition 1).*

We will also use the fact that $\mathrm{proj}_{\mathcal{H}}(\mathbf{H}) \in \mathbb{S}_+$ for all $\mathbf{H} \in \mathbb{S}_+$. This is implied by property **(iii)** in Assumption 1 and standard limiting arguments using the continuity of the projection.

A.2 Singular Preconditioners

As previously mentioned, the results in Theorems 3 and 5 are already enough to essentially achieve dimension-independent iteration complexities for both Algorithms 1 and 2 by choosing a very small value of δ^2. This aligns well with common practice, for instance, in AdaGrad-like algorithms, where the corresponding parameter is typically chosen as $\sim 10^{-10}$. However, for full mathematical rigor, we will show that the update rules on line 7 of Algorithm 1 and line 7 of Algorithm 2 are well-defined in the case $\delta = 0$ by continuous extension via appropriate pseudoinverse operators. This is justified by the following Lemma 9.

Lemma 9 (↓). *Let $u_0, \ldots, u_k \in \mathcal{X}$ and let $\mathbf{S}_k(\delta) = \delta^2 \mathbf{I} + \sum_{i=0}^{k} \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u_i))$. Then, the following relation holds for all $m \in \mathcal{X}$:*

$$\lim_{\delta \downarrow 0} [\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k(\delta)]^{-\frac{1}{2}} m = [\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k(0)]^{\frac{\dagger}{2}} m. \tag{25}$$

Consequently, by standard limiting arguments, it is straightforward that the results in Theorems 3, 4 and 5 hold true in the case $\delta = 0$.

B Proofs for Section 3

B.1 Proof of Lemma 1

The convexity of the function $\Psi_k^*(m)$ follows from its representation as a finite-valued infimum over $\mathbf{H} \in \mathcal{H} \cap \mathbb{S}_{++}$ of a function jointly convex in $(m, \mathbf{H})$:

$$\Psi_k^*(m) = \inf_{\mathbf{H} \in \mathcal{H} \cap \mathbb{S}_{++}} \tfrac{1}{2}\eta[\|\mathbf{H}\|_{\mathrm{tr}} + \langle \mathbf{S}_k, \mathbf{H}^{-1} \rangle + \|m\|^2_{\mathbf{H}^{-1}}].$$

Next, we compute the differential $\mathrm{d}\Psi_k^*(m)$ as follows:

$$\begin{aligned}
\mathrm{d}\Psi_k^*(m) &\overset{\text{(a)}}{=} \tfrac{1}{2}\eta\langle [\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k]^{-1/2}, \mathrm{d}[\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k] \rangle \\
&\overset{\text{(b)}}{=} \tfrac{1}{2}\eta\langle [\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k]^{-1/2}, \mathrm{proj}_{\mathcal{H}}(\mathrm{d}m\langle m, \cdot \rangle + m\langle \mathrm{d}m, \cdot \rangle) \rangle \\
&\overset{\text{(c)}}{=} \tfrac{1}{2}\eta\langle [\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k]^{-1/2}, [\mathrm{d}m\langle m, \cdot \rangle + m\langle \mathrm{d}m, \cdot \rangle] \rangle \\
&= \langle \eta[\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k]^{-1/2} m, \mathrm{d}m \rangle,
\end{aligned}$$

where (a) uses the differentiation rule for trace functions; (b) uses the linearity of $\mathrm{proj}_{\mathcal{H}}(\cdot)$ and the bilinearity of $\mathbf{out}(\cdot)$; (c) uses the fact that $\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k \in \mathcal{H}$, Assumption 1, and the properties of the projection. □

B.2 Proof of Lemma 2

Feasibility. Lemma 1 implies $\nabla\Psi_k^*(m) = \eta \mathbf{A}_k^{-1/2} m$, where $\mathbf{A}_k = \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k$. Hence, we can upper-bound $\rho^2(\nabla\Psi_k^*(m))$ as follows:

$$\begin{aligned}
\rho^2(\nabla\Psi_k^*(m)) &\overset{\text{(a)}}{=} \|\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(\nabla\Psi_k^*(m)))\|_{\mathrm{op}} = \max_{\|u\| \le 1} \langle \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(\nabla\Psi_k^*(m))), \mathbf{out}(u) \rangle \\
&\overset{\text{(b)}}{=} \max_{\|u\| \le 1} \langle \mathbf{out}(\nabla\Psi_k^*(m)), \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u)) \rangle \\
&\overset{\text{(c)}}{=} \eta^2 \max_{\|u\| \le 1} \langle \mathbf{A}_k^{-1/2}\, \mathbf{out}(m) \mathbf{A}_k^{-1/2}, \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u)) \rangle \\
&= \eta^2 \max_{\|u\| \le 1} \langle \mathbf{out}(m), \mathbf{A}_k^{-1/2}\, \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u)) \mathbf{A}_k^{-1/2} \rangle \\
&\overset{\text{(d)}}{=} \eta^2 \max_{\|u\| \le 1} \langle \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)), \mathbf{A}_k^{-1/2}\, \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u)) \mathbf{A}_k^{-1/2} \rangle \\
&\overset{\text{(e)}}{\le} \eta^2 \max_{\|u\| \le 1} \langle \mathbf{A}_k, \mathbf{A}_k^{-1/2}\, \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u)) \mathbf{A}_k^{-1/2} \rangle \\
&= \eta^2 \max_{\|u\| \le 1} \langle \mathbf{I}, \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u)) \rangle \\
&\overset{\text{(f)}}{=} \eta^2 \max_{\|u\| \le 1} \langle \mathbf{I}, \mathbf{out}(u) \rangle = \eta^2 \max_{\|u\| \le 1} \|u\|^2 = \eta^2,
\end{aligned}$$

where (a) uses the definition in Equation (4); (b) uses the properties of the projection; (c) uses the expression for $\nabla\Psi_k^*(m)$ above; (d) uses the fact that $\mathbf{A}_k^{-1/2} \in \mathcal{H}$, Assumption 1 and the properties of the projection; (e) uses the fact that $\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) \preceq \mathbf{A}_k$ and the fact that $\mathbf{A}_k^{-1/2}\, \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(u)) \mathbf{A}_k^{-1/2} \succeq \mathbf{O}$ due to Assumption 1; (f) uses the fact that $\mathbf{I} \in \mathcal{H}$ due to Assumption 1 and the properties of the projection.

Dominance. We can lower-bound $\Psi_k^*(m)$ as follows:

$$\Psi_k^*(m) = \eta\langle \mathbf{A}_k^{1/2}, \mathbf{I} \rangle \overset{\text{(a)}}{\ge} \eta\langle \sqrt{\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m))}, \mathbf{I} \rangle \overset{\text{(b)}}{=} \eta\rho_*(m),$$

where (a) uses the monotonicity of the trace square root; (b) uses the definition in eq. (4).

Upper stability. The inequality $\Psi_{k+1}^*(m) \ge \Psi_k^*(m)$ is implied by the monotonicity of the trace square root. Furthermore, $\Psi_{k+1}^*(m) - \Psi_k^*(m)$ can be upper-bounded as follows:

$$\begin{aligned}
\Psi_{k+1}^*(m) - \Psi_k^*(m) &\overset{\text{(a)}}{=} \eta\langle \mathbf{I}, \mathbf{A}_{k+1}^{1/2} - \mathbf{A}_k^{1/2} \rangle \\
&\overset{\text{(b)}}{=} \int_0^1 \tfrac{1}{2}\eta\langle [\mathbf{A}_k + t(\mathbf{A}_{k+1} - \mathbf{A}_k)]^{-1/2}, \mathbf{A}_{k+1} - \mathbf{A}_k \rangle \mathrm{d}t
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^1 \tfrac{1}{2}\eta\langle[\mathbf{A}_k + t(\mathbf{S}_{k+1} - \mathbf{S}_k)]^{-1/2}, \mathbf{S}_{k+1} - \mathbf{S}_k\rangle \mathrm{d}t \\
&\overset{(c)}{\leq} \int_0^1 \tfrac{1}{2}\eta\langle[\mathbf{S}_k + t(\mathbf{S}_{k+1} - \mathbf{S}_k)]^{-1/2}, \mathbf{S}_{k+1} - \mathbf{S}_k\rangle \mathrm{d}t \\
&\overset{(d)}{=} \eta\langle \mathbf{I}, \mathbf{S}_{k+1}^{1/2} - \mathbf{S}_k^{1/2}\rangle = \eta\Big[\|\sqrt{\mathbf{S}_{k+1}}\|_{\mathrm{tr}} - \|\sqrt{\mathbf{S}_k}\|_{\mathrm{tr}}\Big],
\end{aligned}
$$

where (a) uses the definition in eq. (3); (b) and (d) use the differentiability properties of the trace square root; (c) uses the assumption $\mathbf{S}_{k+1} \succeq \mathbf{S}_k$, the fact that $\mathbf{A}_k \succeq \mathbf{S}_k$, and the Löwner Heinz theorem (Carlen, 2010, Theorem 2.6).

Smoothness. It is sufficient to upper-bound the second-order differential $\mathrm{d}^2\Psi_k^*(m)$:

$$
\begin{aligned}
\mathrm{d}^2\Psi_k^*(m) &\overset{(a)}{=} \mathrm{d}^2[\eta\langle \mathbf{I}, \mathbf{A}_k^{1/2}\rangle] \\
&\overset{(b)}{=} \tfrac{1}{2}\eta\mathrm{d}[\langle \mathbf{A}_k^{-1/2}, \mathrm{d}\mathbf{A}_k\rangle] = \tfrac{1}{2}\eta\langle \mathbf{A}_k^{-1/2}, \mathrm{d}^2\mathbf{A}_k\rangle + \tfrac{1}{2}\eta\langle \mathrm{d}\mathbf{A}_k^{-1/2}, \mathrm{d}\mathbf{A}_k\rangle \\
&\overset{(c)}{=} \tfrac{1}{2}\eta\langle \mathbf{A}_k^{-1/2}, \mathrm{proj}_{\mathcal{H}}(2\,\mathbf{out}(\mathrm{d}m))\rangle + \tfrac{1}{2}\eta\langle \mathrm{d}\mathbf{A}_k^{-1/2}, \mathrm{d}\mathbf{A}_k\rangle \\
&\overset{(d)}{=} \eta\langle \mathbf{A}_k^{-1/2}, \mathbf{out}(\mathrm{d}m)\rangle + \tfrac{1}{2}\eta\langle \mathrm{d}\mathbf{A}_k^{-1/2}, \mathrm{d}\mathbf{A}_k\rangle \\
&= \eta\|\mathrm{d}m\|^2_{\mathbf{A}_k^{-1/2}} + \tfrac{1}{2}\eta\langle \mathrm{d}\mathbf{A}_k^{-1/2}, \mathrm{d}\mathbf{A}_k\rangle \\
&\overset{(e)}{\leq} \eta\|\mathrm{d}m\|^2_{\mathbf{A}_k^{-1/2}} \overset{(f)}{\leq} \eta\|\mathrm{d}m\|^2_{\mathbf{S}_k^{-1/2}},
\end{aligned}
$$

where (a) uses the definition in eq. (3); (b) uses the differentiability properties of the trace square root; (c) uses the definition of $\mathbf{A}_k$ and the linearity of the projection; (d) uses the fact that $\mathbf{A}_k^{-1/2} \in \mathcal{H}$ due to Assumption 1; (e) uses the fact that $\frac{1}{2}\langle \mathrm{d}\mathbf{A}_k^{-1/2}, \mathrm{d}\mathbf{A}_k\rangle$ is the second-order differential of the concave function $\mathbf{A} \mapsto \langle \mathbf{A}^{1/2}, \mathbf{I}\rangle$; (f) uses the fact that $\mathbf{S}_k \preceq \mathbf{A}_k$ and the Löwner-Heinz theorem (Carlen, 2010, Theorem 2.6). □

B.3 Proof of Lemma 3

We can rewrite Reg_K^+ as follows:

$$
\begin{aligned}
\mathrm{Reg}_K^+ &\overset{(a)}{=} \sum_{k=0}^K \langle g_k, x_{k+1}\rangle + \eta\rho_*(m_K) \\
&\overset{(b)}{=} \sum_{k=0}^K \langle g_k, -\nabla\Psi_k^*(m_k)\rangle + \eta\rho_*(m_K) \\
&\overset{(c)}{=} \sum_{k=0}^K \langle m_{k-1} - m_k, \nabla\Psi_k^*(m_k)\rangle + \eta\rho_*(m_K) \\
&= \sum_{k=0}^K \big[\Psi_k^*(m_{k-1}) - \Psi_k^*(m_k) - \mathbb{B}_{\Psi_k^*}(m_{k-1}; m_k)\big] + \eta\rho_*(m_K) \\
&= \Psi_0^*(m_{-1}) - \Psi_0^*(m_0) + \sum_{k=1}^K \big[\Psi_k^*(m_{k-1}) - \Psi_{k-1}^*(m_{k-1}) + \Psi_{k-1}^*(m_{k-1}) - \Psi_k^*(m_k)\big] \\
&\quad + \eta\rho_*(m_K) - \sum_{k=0}^K \mathbb{B}_{\Psi_k^*}(m_{k-1}; m_k) \\
&= \Psi_0^*(0) + \eta\rho_*(m_K) - \Psi_K^*(m_K) + \sum_{k=1}^K \big[\Psi_k^*(m_{k-1}) - \Psi_{k-1}^*(m_{k-1})\big] \\
&\quad - \sum_{k=0}^K \mathbb{B}_{\Psi_k^*}(m_{k-1}; m_k),
\end{aligned}
$$

where (a) uses the definition in eq. (8) duality between the norms defined in eq. (4); (b) uses eq. (2); (c) uses eq. (1). □

B.4 Proof of Lemma 4

Let $\mathbf{A}_k = \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) + \mathbf{S}_k$. Then, we can obtain the following:

$$
\begin{aligned}
&\|\nabla\Psi^*_{k+1}(m) - \nabla\Psi^*_k(m)\|^2_{\mathbf{S}^{1/2}_{k+1}} \\
&\quad\overset{(a)}{=} \eta^2\|(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k)m\|^2_{\mathbf{S}^{1/2}_{k+1}} \\
&\quad\overset{(b)}{\leq} \eta^2\|(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k)m\|^2_{\mathbf{A}^{1/2}_{k+1}} \\
&\quad= \eta^2\langle(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k)\mathbf{A}^{1/2}_{k+1}(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k), \mathbf{out}(m)\rangle \\
&\quad\overset{(c)}{=} \eta^2\langle(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k)\mathbf{A}^{1/2}_{k+1}(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k), \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m))\rangle \\
&\quad\overset{(d)}{\leq} \eta^2\langle(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k)\mathbf{A}^{1/2}_{k+1}(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k), \mathbf{A}_k\rangle \\
&\quad= \eta^2\langle\mathbf{A}^{1/2}_{k+1} - \mathbf{A}^{1/2}_k, \mathbf{I} - \mathbf{A}^{-1/2}_{k+1}\mathbf{A}^{1/2}_k\rangle = \eta^2\langle\mathbf{A}^{1/2}_{k+1} - \mathbf{A}^{1/2}_k, \mathbf{I}\rangle + \eta^2\langle\mathbf{A}_k, \mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k\rangle \\
&\quad\overset{(e)}{\leq} \eta^2\Big[\|\mathbf{A}^{1/2}_{k+1}\|_{\mathrm{tr}} - \|\mathbf{A}^{1/2}_k\|_{\mathrm{tr}}\Big] \overset{(f)}{=} \eta\big[\Psi^*_{k+1}(m) - \Psi^*_k(m)\big] \overset{(g)}{\leq} \eta^2\Big[\|\sqrt{\mathbf{S}_{k+1}}\|_{\mathrm{tr}} - \|\sqrt{\mathbf{S}_k}\|_{\mathrm{tr}}\Big],
\end{aligned}
$$

where (a) uses Lemma 1; (b) uses the fact that $\mathbf{A}_{k+1} \succeq \mathbf{S}_{k+1}$ and the Löwner-Heinz theorem (Carlen, 2010, Theorem 2.6); (c) uses the properties of the projection and the inclusion $(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k)\mathbf{A}^{1/2}_{k+1}(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k) \in \mathcal{H}$ due to the definition of $\mathbf{A}_k$, Assumption 1, and the inclusion $\mathbf{S}_k \in \mathcal{H}$; (d) uses the fact that $(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k)\mathbf{A}^{1/2}_{k+1}(\mathbf{A}^{-1/2}_{k+1} - \mathbf{A}^{-1/2}_k) \succeq \mathbf{O}$ and $\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m)) \preceq \mathbf{A}_k$; (e) uses the fact that $\mathbf{O} \preceq \mathbf{A}_k \preceq \mathbf{A}_{k+1}$ and the Löwner-Heinz theorem (Carlen, 2010, Theorem 2.6); (f) uses the definition in eq. (3); (g) uses Lemma 2. □

B.5 Proof of Theorem 2

We can upper-bound Reg^+_K as follows:

$$
\begin{aligned}
\mathrm{Reg}^+_K &\overset{(a)}{\leq} \Psi^*_0(0) + \eta\rho_*(m_K) - \Psi^*_K(m_K) \\
&\quad+ \sum_{k=1}^K\big[\Psi^*_k(m_{k-1}) - \Psi^*_{k-1}(m_{k-1})\big] - \sum_{k=0}^K \mathbb{B}_{\Psi^*_k}(m_{k-1}; m_k), \\
&\overset{(b)}{\leq} \Psi^*_0(0) + \sum_{k=1}^K\big[\Psi^*_k(m_{k-1}) - \Psi^*_{k-1}(m_{k-1})\big] - \sum_{k=0}^K \mathbb{B}_{\Psi^*_k}(m_{k-1}; m_k), \\
&\overset{(c)}{\leq} \Psi^*_0(0) + \sum_{k=1}^K \eta\Big[\|\sqrt{\mathbf{S}_k}\|_{\mathrm{tr}} - \|\sqrt{\mathbf{S}_{k-1}}\|_{\mathrm{tr}}\Big] - \sum_{k=0}^K \mathbb{B}_{\Psi^*_k}(m_{k-1}; m_k), \\
&\overset{(d)}{\leq} \Psi^*_0(0) + \eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} - \sum_{k=0}^K \tfrac{1}{2\eta}\|\nabla\Psi^*_k(m_{k-1}) - \nabla\Psi^*_k(m_k)\|^2_{\mathbf{S}^{1/2}_k} \\
&\overset{(e)}{=} \Psi^*_0(0) + \eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} - \tfrac{1}{2\eta}\|x_1 - x_0\|^2_{\mathbf{S}^{1/2}_0} - \sum_{k=1}^K \tfrac{1}{2\eta}\|\nabla\Psi^*_k(m_{k-1}) + x_{k+1}\|^2_{\mathbf{S}^{1/2}_k} \\
&\overset{(f)}{\leq} \Psi^*_0(0) + \eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} - \tfrac{1}{2\eta}\|x_1 - x_0\|^2_{\mathbf{S}^{1/2}_0} - \sum_{k=1}^K \tfrac{1}{4\eta}\|x_{k+1} - x_k\|^2_{\mathbf{S}^{1/2}_k} \\
&\quad+ \sum_{k=1}^K \tfrac{1}{2\eta}\|\nabla\Psi^*_k(m_{k-1}) - \nabla\Psi^*_{k-1}(m_{k-1})\|^2_{\mathbf{S}^{1/2}_k} \\
&\overset{(g)}{\leq} \Psi^*_0(0) + \eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} - \sum_{k=0}^K \tfrac{1}{4\eta}\|x_{k+1} - x_k\|^2_{\mathbf{S}^{1/2}_k} + \sum_{k=1}^K \tfrac{1}{2}\eta\Big[\|\sqrt{\mathbf{S}_k}\|_{\mathrm{tr}} - \|\sqrt{\mathbf{S}_{k-1}}\|_{\mathrm{tr}}\Big] \\
&\leq \Psi^*_0(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} - \sum_{k=0}^K \tfrac{1}{4\eta}\|x_{k+1} - x_k\|^2_{\mathbf{S}^{1/2}_k},
\end{aligned}
$$

where (a) uses Lemma 3; (b) uses the dominance property in Lemma 2; (c) uses the upper stability property in Lemma 2; (d) uses the smoothness property in Lemma 2; (e) uses eqs. (1) and (2), Lemma 1, and the assumption $x_0 = 0$; (f) uses eq. (2); (g) uses Lemma 4. □

B.6 Proof of Theorem 3

We can upper-bound Reg_K as follows:

$$
\begin{aligned}
\mathrm{Reg}_K &= \mathrm{Reg}_K^+ + \sum_{k=0}^{K} \langle g_k, x_k - x_{k+1} \rangle \\
&\overset{(a)}{\leq} \mathrm{Reg}_K^+ + \sum_{k=0}^{K} \left[\eta \|g_k\|^2_{\mathbf{S}_k^{-1/2}} + \frac{1}{4\eta} \|x_{k+1} - x_k\|^2_{\mathbf{S}_k^{1/2}} \right] \\
&\overset{(b)}{\leq} \Psi_0^*(0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} + \sum_{k=0}^{K} \eta \|g_k\|^2_{\mathbf{S}_k^{-1/2}} \\
&\overset{(c)}{\leq} \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} + \sum_{k=0}^{K} \eta \|g_k\|^2_{\mathbf{S}_k^{-1/2}} \\
&\overset{(d)}{=} \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} + \sum_{k=0}^{K} \eta \langle \mathbf{S}_k^{-1/2}, \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(g_k)) \rangle \\
&\overset{(e)}{=} \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} + \sum_{k=0}^{K} \eta \langle \mathbf{S}_k^{-1/2}, \mathbf{S}_k - \mathbf{S}_{k-1} \rangle \\
&= \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} \\
&\quad + \sum_{k=0}^{K} \eta \langle \mathbf{S}_k^{-1/2}, (\mathbf{S}_k^{1/2} - \mathbf{S}_{k-1}^{1/2})(\mathbf{S}_k^{1/2} + \mathbf{S}_{k-1}^{1/2}) + \mathbf{S}_{k-1}^{1/2}\mathbf{S}_k^{1/2} - \mathbf{S}_k^{1/2}\mathbf{S}_{k-1}^{1/2} \rangle \\
&= \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} + \sum_{k=0}^{K} \eta \langle \mathbf{I} + \mathbf{S}_k^{-1/2}\mathbf{S}_{k-1}^{1/2}, \mathbf{S}_k^{1/2} - \mathbf{S}_{k-1}^{1/2} \rangle \\
&\overset{(f)}{\leq} \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} + \sum_{k=0}^{K} \eta \left[1 + \|\mathbf{S}_k^{-1/2}\mathbf{S}_{k-1}^{1/2}\|_{\mathrm{op}} \right] \|\mathbf{S}_k^{1/2} - \mathbf{S}_{k-1}^{1/2}\|_{\mathrm{tr}} \\
&\overset{(g)}{\leq} \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{3}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}} + \sum_{k=0}^{K} 2\eta \left[\|\sqrt{\mathbf{S}_k}\|_{\mathrm{tr}} - \|\sqrt{\mathbf{S}_{k-1}}\|_{\mathrm{tr}} \right] \\
&\leq \eta\delta \dim(\mathcal{X}) + \eta\rho_*(g_0) + \tfrac{7}{2}\eta \|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}},
\end{aligned}
$$

where (a) uses Young's inequality; (b) uses Theorem 2; (c) uses lines 3 and 6, the definition in eq. (3), Lemma 3 of An et al. (2025), and the definition in eq. (4); (d) uses the fact that $\mathbf{S}_k^{-1/2} \in \mathcal{H}$ due to Assumption 1, and the properties of the projection; (e) uses line 6; (f) uses Hölder's inequality for Schatten norms and the triangle inequality; (g) uses the fact that $\mathbf{S}_k \succeq \mathbf{S}_{k-1}$ due to line 6 and Assumption 1. □

C Proofs for Section 4

C.1 Proof of Lemma 5

First, we can lower-bound $\mathbb{E}[\langle g_k, x_{k+1} - x\rangle]$ for $k \in \mathbb{N}_0$ as follows:

$$
\begin{aligned}
&\mathbb{E}[\langle g_k, x_{k+1} - x\rangle] \\
&\quad = \mathbb{E}[\langle g_k, x_{k+1} - x_k\rangle] + \mathbb{E}[\langle g_k, x_k - x\rangle] \\
&\quad \overset{\text{(a)}}{=} \mathbb{E}[\langle g_k, x_{k+1} - x_k\rangle] + \mathbb{E}[\langle \nabla f_k(x_k), x_k - x\rangle] \\
&\quad \overset{\text{(b)}}{\geq} \mathbb{E}[\langle g_k, x_{k+1} - x_k\rangle] + \mathbb{E}[f_k(x_k) - f_k(x)] \\
&\quad = \mathbb{E}[\langle g_k, x_{k+1} - x_k\rangle] + \mathbb{E}[f_k(x_{k+1}) + \langle \nabla f_k(x_{k+1}), x_k - x_{k+1}\rangle + \mathbb{B}_{f_k}(x_k; x_{k+1}) - f_k(x)] \\
&\quad = \mathbb{E}[\langle g_k - \nabla f_k(x_{k+1}), x_{k+1} - x_k\rangle] + \mathbb{E}[\mathbb{B}_{f_k}(x_k; x_{k+1}) + f_k(x_{k+1}) - f_k(x)] \\
&\quad \overset{\text{(c)}}{=} \mathbb{E}[\langle g_k - \tilde{g}_{k+1}, x_{k+1} - x_k\rangle] + \mathbb{E}[\mathbb{B}_{f_k}(x_k; x_{k+1}) + f_k(x_{k+1}) - f_k(x)] \\
&\quad \overset{\text{(d)}}{\geq} -\mathbb{E}\left[\tfrac{1}{4}\eta^{-1}\|x_{k+1} - x_k\|^2_{\mathbf{S}^{1/2}_{k+1}} + \eta\|g_k - \tilde{g}_{k+1}\|^2_{\mathbf{S}^{-1/2}_{k+1}}\right] \\
&\quad\quad + \mathbb{E}[\mathbb{B}_{f_k}(x_k; x_{k+1}) + f_k(x_{k+1}) - f_k(x)]
\end{aligned}
$$

where (a) uses Assumption 4 and the fact that x_k is independent of ξ_k; (b) uses the convexity in Assumption 3; (c) uses line 9, Assumption 4, and the fact that x_{k+1} and x_k are independent of ξ'_{k+1}; (d) uses Young's inequality. After rearranging and summing these inequalities for $k = 0, \ldots, K$, we obtain the following:

$$
\begin{aligned}
&\sum_{k=0}^{K} \mathbb{E}[\mathbb{B}_{f_k}(x_k; x_{k+1}) + f_k(x_{k+1}) - f_k(x)] \\
&\quad \leq \mathbb{E}\left[\sum_{k=0}^{K}\langle g_k, x_{k+1} - x\rangle\right] + \sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{4\eta}\|x_{k+1} - x_k\|^2_{\mathbf{S}^{1/2}_{k+1}} + \eta\|g_k - \tilde{g}_{k+1}\|^2_{\mathbf{S}^{-1/2}_{k+1}}\right] \\
&\quad \overset{\text{(a)}}{\leq} \mathbb{E}\left[\mathrm{Reg}^+_K\right] + \sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{4\eta}\|x_{k+1} - x_k\|^2_{\mathbf{S}^{1/2}_{k+1}} + \eta\|g_k - \tilde{g}_{k+1}\|^2_{\mathbf{S}^{-1/2}_{k+1}}\right] \\
&\quad \overset{\text{(b)}}{\leq} \mathbb{E}\left[\Psi^*_0(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{4\eta}\|x_{k+1} - x_k\|^2_{\mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k}} + \eta\|g_k - \tilde{g}_{k+1}\|^2_{\mathbf{S}^{-1/2}_{k+1}}\right] \\
&\quad \overset{\text{(c)}}{=} \mathbb{E}\left[\Psi^*_0(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{4\eta}\langle \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(x_{k+1} - x_k)), \mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k}\rangle\right] \\
&\quad\quad + \sum_{k=0}^{K}\mathbb{E}\left[\eta\langle \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(g_k - \tilde{g}_{k+1})), \mathbf{S}^{-1/2}_{k+1}\rangle\right] \\
&\quad \overset{\text{(d)}}{\leq} \mathbb{E}\left[\Psi^*_0(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{4\eta}\rho^2(x_{k+1} - x_k)\|\mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k}\|_{\mathrm{tr}}\right] \\
&\quad\quad + \sum_{k=0}^{K}\mathbb{E}\left[\eta\langle \mathbf{S}_{k+1} - \mathbf{S}_k, \mathbf{S}^{-1/2}_{k+1}\rangle\right] \\
&\quad \overset{\text{(e)}}{\leq} \mathbb{E}\left[\Psi^*_0(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[\eta\|\mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k}\|_{\mathrm{tr}} + \eta\langle \mathbf{S}_{k+1} - \mathbf{S}_k, \mathbf{S}^{-1/2}_{k+1}\rangle\right] \\
&\quad = \mathbb{E}\left[\Psi^*_0(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[\eta\|\mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k}\|_{\mathrm{tr}}\right] \\
&\quad\quad + \sum_{k=0}^{K}\mathbb{E}\left[\eta\langle(\mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k})(\mathbf{S}^{1/2}_{k+1} + \mathbf{S}^{1/2}_{k}) + \mathbf{S}^{1/2}_{k}\mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k+1}\mathbf{S}^{1/2}_{k}, \mathbf{S}^{-1/2}_{k+1}\rangle\right] \\
&\quad = \mathbb{E}\left[\Psi^*_0(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[\eta\|\mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k}\|_{\mathrm{tr}}\right] \\
&\quad\quad + \sum_{k=0}^{K}\mathbb{E}\left[\eta\langle \mathbf{S}^{1/2}_{k+1} - \mathbf{S}^{1/2}_{k}, \mathbf{I} + \mathbf{S}^{-1/2}_{k+1}\mathbf{S}^{1/2}_{k}\rangle\right]
\end{aligned}
$$

$$\overset{(f)}{\leq} \mathbb{E}\left[\Psi_0^*(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[\left[2\eta + \eta\|\mathbf{S}_{k+1}^{-1/2}\mathbf{S}_k^{1/2}\|_{\mathrm{op}}\right]\|\mathbf{S}_{k+1}^{1/2} - \mathbf{S}_k^{1/2}\|_{\mathrm{tr}}\right]$$

$$\overset{(g)}{\leq} \mathbb{E}\left[\Psi_0^*(0) + \tfrac{3}{2}\eta\|\sqrt{\mathbf{S}_K}\|_{\mathrm{tr}}\right] + \sum_{k=0}^{K}\mathbb{E}\left[3\eta\|\mathbf{S}_{k+1}^{1/2} - \mathbf{S}_k^{1/2}\|_{\mathrm{tr}}\right]$$

$$\overset{(h)}{\leq} \mathbb{E}\left[\Psi_0^*(0) + \tfrac{9}{2}\eta\|\sqrt{\mathbf{S}_{K+1}}\|_{\mathrm{tr}}\right]$$

$$\overset{(i)}{\leq} \mathbb{E}\left[\eta\delta\dim(\mathcal{X}) + \tfrac{9}{2}\eta\|\sqrt{\mathbf{S}_{K+1}}\|_{\mathrm{tr}}\right]$$

where (a) is obtained by taking the maximum over $x \in Q_{\mathcal{R}}$ and using the definition in eq. (8); (b) uses Theorem 2; (c) uses the properties of the projection and the fact that $\mathbf{S}_k^{-1/2} \in \mathcal{H}$ and $\mathbf{S}_{k+1}^{1/2} - \mathbf{S}_k^{1/2} \in \mathcal{H}$ due to lines 3 and 10 and Assumption 1; (d) uses line 10, Hölder's inequality for Schatten norms, and the definition in eq. (4); (e) uses the triangle inequality, eq. (2), and the feasibility property in Lemma 2; (f) uses Hölder's inequality for Schatten norms and the triangle inequality; (g) uses the fact that $\mathbf{S}_{k+1} \succeq \mathbf{S}_k$, which is implied by line 10 and Assumption 1; (h) uses the fact that $\mathbf{S}_{k+1}^{1/2} \succeq \mathbf{S}_k^{1/2}$ due to line 10, Assumption 1, and the Löwner Heinz theorem (Carlen, 2010, Theorem 2.6); (i) uses line 3. □

C.2 Proof of Lemma 6

We can lower-bound $\sum_{k=0}^{K}[f_k(x_{k+1}) - f_k(x^*)]$ as follows:

$$\begin{aligned}
\sum_{k=0}^{K}[f_k(x_{k+1}) - f_k(x^*)] &\overset{(a)}{=} \sum_{k=0}^{K}\alpha_k^2[f(\overline{x}_{k+1}) - f(x^*/\alpha_k + (1-1/\alpha_k)\overline{x}_k)] \\
&\overset{(b)}{\geq} \sum_{k=0}^{K}\left[\alpha_k^2[f(\overline{x}_{k+1}) - f(x^*)] - (\alpha_k^2 - \alpha_k)[f(\overline{x}_k) - f(x^*)]\right] \\
&\overset{(c)}{\geq} \alpha_K^2[f(\overline{x}_{K+1}) - f(x^*)] - (\alpha_0^2 - \alpha_0)[f(\overline{x}_0) - f(x^*)] \\
&\overset{(d)}{=} \alpha_K^2[f(\overline{x}_{K+1}) - f(x^*)],
\end{aligned}$$

where (a) uses line 8 and the definition in eq. (17); (b) uses Assumption 3; (c) uses the fact that $f(\overline{x}_k) - f(x^*) \geq 0$ which is implied by $\rho(\overline{x}_k) \leq \eta$ and the assumption $\alpha_k^2 \leq \alpha_k + \alpha_{k-1}^2$; (d) uses the assumption $\alpha_0 = 1$. □

C.3 Proof of Lemma 7

We can upper-bound $\mathbb{E}\left[\|\sqrt{\mathbf{S}_{K+1}}\|_{\mathrm{tr}}\right]$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\|\sqrt{\mathbf{S}_{K+1}}\|_{\mathrm{tr}}\right] &\overset{(a)}{=} \mathbb{E}\left[\langle\sqrt{\mathbf{S}_{K+1}}, \mathbf{I}\rangle\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\langle\sqrt{\mathbf{S}_0}, \mathbf{I}\rangle\right] + \mathbb{E}\left[\langle\sqrt{\mathbf{S}_{K+1} - \mathbf{S}_0}, \mathbf{I}\rangle\right] \\
&\overset{(c)}{=} \delta\dim(\mathcal{X}) + \mathbb{E}\left[\langle\sqrt{\mathbf{S}_{K+1} - \mathbf{S}_0}\mathbf{B}^{-1/2}, \mathbf{B}^{1/2}\rangle\right] \\
&\overset{(d)}{\leq} \delta\dim(\mathcal{X}) + \mathbb{E}\left[\sqrt{\langle\mathbf{S}_{K+1} - \mathbf{S}_0, \mathbf{B}^{-1}\rangle}\right] \\
&\overset{(e)}{=} \delta\dim(\mathcal{X}) + \mathbb{E}\left[\sqrt{\sum_{k=0}^{K}\langle\mathrm{proj}_{\mathcal{H}}(\mathbf{out}(\tilde{g}_{k+1} - g_k)), \mathbf{B}^{-1}\rangle}\right] \\
&\overset{(f)}{=} \delta\dim(\mathcal{X}) + \mathbb{E}\left[\sqrt{\sum_{k=0}^{K}\langle\mathbf{out}(\tilde{g}_{k+1} - g_k), \mathbf{B}^{-1}\rangle}\right] \\
&= \delta\dim(\mathcal{X}) + \mathbb{E}\left[\sqrt{\sum_{k=0}^{K}\|\tilde{g}_{k+1} - g_k\|_{\mathbf{B}^{-1}}^2}\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(g)}{\leq} \delta \dim(\mathcal{X}) + \sqrt{\sum_{k=0}^{K} \mathbb{E}\left[\|\tilde{g}_{k+1} - g_k\|^2_{\mathbf{B}^{-1}}\right]} \\
&\overset{(h)}{\leq} \delta \dim(\mathcal{X}) + \sqrt{\sum_{k=0}^{K} 2\sigma^2\alpha_k^2} + \sqrt{\mathbb{E}\left[\sum_{k=0}^{K} \|\nabla f_k(x_{k+1}) - \nabla f_k(x_k)\|^2_{\mathbf{B}^{-1}}\right]} \\
&\overset{(i)}{\leq} \delta \dim(\mathcal{X}) + \sqrt{\sum_{k=0}^{K} 2\sigma^2\alpha_k^2} + \sqrt{\mathbb{E}\left[\sum_{k=0}^{K} \left[\tfrac{1+\nu}{\nu}\right]^{\frac{2\nu}{1+\nu}} \left[\mathcal{L}\alpha_k^{1-\nu}\right]^{\frac{2}{1+\nu}} \left[\mathbb{B}_{f_k}(x_k; x_{k+1})\right]^{\frac{2\nu}{1+\nu}}\right]} \\
&\leq \delta \dim(\mathcal{X}) + \sigma\sqrt{\sum_{k=0}^{K} 2\alpha_k^2} + \left[\tfrac{1+\nu}{\nu}\right]^{\frac{\nu}{1+\nu}} \mathcal{L}^{\frac{1}{1+\nu}} \sqrt{\mathbb{E}\left[\sum_{k=0}^{K} [\alpha_k^2]^{\frac{(1-\nu)}{1+\nu}} \left[\mathbb{B}_{f_k}(x_k; x_{k+1})\right]^{\frac{2\nu}{1+\nu}}\right]} \\
&\overset{(j)}{\leq} \delta \dim(\mathcal{X}) + \sigma\sqrt{\sum_{k=0}^{K} 2\alpha_k^2} + \left[\tfrac{1+\nu}{\nu}\right]^{\frac{\nu}{1+\nu}} \mathcal{L}^{\frac{1}{1+\nu}} \left[\sum_{k=0}^{K} \alpha_k^2\right]^{\frac{1-\nu}{2(1+\nu)}} \left[\mathbb{E}\left[\sum_{k=0}^{K} \mathbb{B}_{f_k}(x_k; x_{k+1})\right]\right]^{\frac{\nu}{1+\nu}} \\
&\overset{(k)}{\leq} \delta \dim(\mathcal{X}) + \sigma\left[\sum_{k=0}^{K} 2\alpha_k^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\frac{1}{c\eta}\sum_{k=0}^{K} \mathbb{B}_{f_k}(x_k; x_{k+1})\right] + \frac{(c\eta)^\nu}{1+\nu}\mathcal{L}\left[\sum_{k=0}^{K} \alpha_k^2\right]^{\frac{1-\nu}{2}} \\
&\overset{(l)}{\leq} \delta \dim(\mathcal{X}) + 2\sigma\left[[K+1]\alpha_K^2\right]^{\frac{1}{2}} + \mathbb{E}\left[\frac{1}{c\eta}\sum_{k=0}^{K} \mathbb{B}_{f_k}(x_k; x_{k+1})\right] + c\mathcal{L}\eta^\nu\left[[K+1]\alpha_K^2\right]^{\frac{1-\nu}{2}},
\end{aligned}
$$

where (a) uses the fact that $\mathbf{S}_{K+1} \in \mathbb{S}_+$; (b) uses Lemma 3 of An et al. (2025); (c) uses line 3; (d) uses Young's inequality and the fact that $\|\mathbf{B}\|_{\text{tr}} = 1$; (e) uses lines 3 and 10; (f) uses the properties of the projection and the fact that $\mathbf{B}^{-1} \in \mathcal{H}$ due to Assumption 1; (g) uses Jensen's inequality; (h) uses lines 6 and 9, Minkowski's inequality, the definition in eq. (17), Assumption 4, and the fact that ξ_k, ξ'_{k+1} are independent; (i) uses eq. (21); (j) uses Hölder's inequality; (k) uses Young's inequality; (l) uses the assumptions $\alpha_k \leq \alpha_{k+1}$ and $c \geq 1$. □

C.4 Proof of Theorem 5

We combine Lemma 5 and Lemma 7 with $c = 9/2$ and obtain the following inequality:

$$
\begin{aligned}
\sum_{k=0}^{K} \mathbb{E}[\mathbb{B}_{f_k}(x_k; x_{k+1}) + f_k(x_{k+1}) - f_k(x^*)] \leq \mathbb{E}\left[\sum_{k=0}^{K} \mathbb{B}_{f_k}(x_k; x_{k+1})\right] \\
+ \mathcal{O}\left[\mathcal{R}\delta \dim(\mathcal{X}) + \sigma\mathcal{R}\left[[K+1]\alpha_K^2\right]^{\frac{1}{2}} + \mathcal{L}\mathcal{R}^{1+\nu}\left[[K+1]\alpha_K^2\right]^{\frac{1-\nu}{2}}\right].
\end{aligned}
$$

After rearranging, we get the following inequality

$$
\sum_{k=0}^{K} \mathbb{E}[f_k(x_{k+1}) - f_k(x^*)] \leq \mathcal{O}\left[\mathcal{R}\delta \dim(\mathcal{X}) + \sigma\mathcal{R}\left[[K+1]\alpha_K^2\right]^{\frac{1}{2}} + \mathcal{L}\mathcal{R}^{1+\nu}\left[[K+1]\alpha_K^2\right]^{\frac{1-\nu}{2}}\right].
$$

Next, we use Lemma 6 and obtain the following inequality:

$$
\alpha_K^2 \mathbb{E}[f(\overline{x}_{K+1}) - f(x^*)] \leq \mathcal{O}\left[\mathcal{R}\delta \dim(\mathcal{X}) + \sigma\mathcal{R}\left[[K+1]\alpha_K^2\right]^{\frac{1}{2}} + \mathcal{L}\mathcal{R}^{1+\nu}\left[[K+1]\alpha_K^2\right]^{\frac{1-\nu}{2}}\right].
$$

It remains to use the assumption $\alpha_K = \frac{1}{2}(K+2)$. □

D PROOFS FOR SECTION A

D.1 PROOF OF LEMMA 8

First, we show that $\mathbf{H}^p \in \mathcal{H}$ for all $\mathbf{H} \in \mathcal{H}$ and $p \in \mathbb{N}_0$. We can show this by induction. The base case, $p \in \{0, 1\}$, is implied by property **(i)** in Assumption 1 and the fact that $\mathbf{H} \in \mathcal{H}$. Then, assuming that $p \geq 2$ and $\mathbf{H}^q \in \mathcal{H}$ for all $q \in \{0, \ldots, p-1\}$, we conclude that $\mathbf{H}^p = \mathbf{H}^{\lfloor p/2 \rfloor} \mathbf{H}^{p \bmod 2} \mathbf{H}^{\lfloor p/2 \rfloor} \in \mathcal{H}$ by property **(ii)** in Assumption 1 and the induction hypothesis. Now, the closedness of $\mathcal{H}$ under arbitrary operator functions follows by a polynomial interpolation argument. □

D.2 PROOF OF LEMMA 9

To prove the statement of the lemma, it is sufficient to show that $m \in \mathrm{range}(\mathbf{M} + \mathbf{S}_k(0))$, where $\mathbf{M} = \mathrm{proj}_{\mathcal{H}}(\mathbf{out}(m))$. Since $\mathbf{S}_k(0) \in \mathbb{S}_+$, it is sufficient to show that $m \in \mathrm{range}(\mathbf{M})$, or that $\mathrm{proj}_{\ker(\mathbf{M})}(m) = 0$. Let $\mathbf{P} \in \mathbb{S}_+$ be the orthogonal projector onto $\ker(\mathbf{M})$. By Lemma 8, we have $\mathbf{P} \in \mathcal{H}$. Hence, by properties of the projection, we have $\|\mathbf{P}m\|^2 = \langle \mathbf{P}, \mathbf{out}(m) \rangle = \langle \mathbf{P}, \mathbf{M} \rangle = 0$, which implies $\mathbf{P}m = 0$. □