Structured Preconditioners in Adaptive Optimization: A Unified Analysis

Shuo Xie [1] **Tianhao Wang** [1] **Sashank Reddi** [2] **Sanjiv Kumar** [2] **Zhiyuan Li** [1 2]

Abstract

We present a novel unified analysis for a broad class of adaptive optimization algorithms with structured (e.g., layerwise, diagonal, and kronecker-factored) preconditioners for both online regret minimization and offline convex optimization. Our analysis not only provides matching rate to several important structured preconditioned algorithms including diagonal AdaGrad, full-matrix AdaGrad, and AdaGrad-Norm, but also gives an improved convergence rate for a one-sided variant of Shampoo over that of original Shampoo. Interestingly, more structured preconditioners (e.g., diagonal Adagrad, AdaGrad-Norm which use less space and compute) are often presented as computationally efficient approximations to full-matrix Adagrad, aiming for improved optimization performance through better approximations. Our unified analysis challenges this prevailing view and reveals, perhaps surprisingly, that more structured preconditioners, despite using less space and computation per step, can outperform their less structured counterparts. To demonstrate this, we show that one-sided Shampoo, which is relatively much cheaper than full-matrix AdaGrad could outperform it both theoretically and experimentally.

1. Introduction

Adaptive optimization algorithms (Streeter & McMahan, 2010; Duchi et al., 2011; Kingma & Ba, 2014) play a pivotal role in modern machine learning, especially in the expensive training of large foundation models. Within the machine learning community, full-matrix AdaGrad is considered as an ideal adaptive preconditioner for fast convergence in terms of number of steps. The computation of full-matrix AdaGrad precondtioner typically involves inverse square root of a $d \times d$ matrix where d is the number of parameters. Thus, for large-scale settings, the huge computation and memory cost makes it prohibitively expensive. This has inspired works on designing more efficient adaptive optimizers by using a structured preconditioner, such as coordinate-wise adaptivity employed by AdaGrad (Streeter & McMahan, 2010; Duchi et al., 2011), Kronecker product based preconditioner employed by Shampoo (Gupta et al., 2018; Anil et al., 2020) and layerwise adaptivity employed by LARS (You et al., 2017) and LAMB (You et al., 2019)), which all aim to provide a computational and memory efficient approximation of full-matrix AdaGrad. These algorithms have been shown to be very effective for general deep learning settings. It is often assumed that a better approximation of full-matrix preconditioner usually results in better optimization convergence, as seen with methods like Shampoo (Gupta et al., 2018). In contrast, preconditioners with more structure such as diagonal preconditioner have inferior performance. In this paper, we challenge these prevailing notions by providing theoretical and empirical evidence against them.

Conceptually, one could equate the degree of structure in a preconditioner to the ease of its computation and storage. In this view, full-matrix AdaGrad can be considered as the least structured and most expensive preconditioner while AdaGrad-Norm (Ward et al., 2020), which only maintains a scalar and uses the same preconditioning for every direction, is the most structured and least expensive. The conventional wisdom here is that using a less structured preconditioner, which requires more space and compute per step, reduces the number of steps needed for training. Thus, choosing the right structure balances this trade-off between convergence speed and training step cost.

Among these preconditioning methods, Shampoo (Gupta et al., 2018) has gained notable attention for its Kronecker-factored preconditioning approach, which promises improved convergence in large-scale optimization tasks (Dahl et al., 2023). Shampoo was originally proposed as a computationally efficient surrogate for full-matrix AdaGrad (Duchi et al., 2011). Despite its popularity, existing analyses of Shampoo (Gupta et al., 2018) are limited and do not provide a full justification for its effectiveness. In particular, we argue that the best-known regret bounds for both Shampoo as well as full-matrix AdaGrad are consistently worse

[1]Toyota Technological Institute at Chicago [2]Google Research. Correspondence to: Shuo Xie <shuox@ttic.edu>, Zhiyuan Li <zhiyuanli@ttic.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Algorithm	Subalgebra $\mathcal{K}$	$\|\boldsymbol{x}\|_{\mathcal{H}}$	$\|\|\|\boldsymbol{g}_{1:T}\|\|\|_{\mathcal{H}}$	Regret Bound $= \|\mathcal{X}\|_{\mathcal{H}} \cdot \|\|\|\boldsymbol{g}_{1:T}\|\|\|_{\mathcal{H}}$
AdaGrad-Norm	$c \cdot \boldsymbol{I}_d$ for $c \in \mathbb{R}$	$\frac{\|\boldsymbol{x}\|_2}{\sqrt{d}}$	$\sqrt{d}\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|_2^2}$	$\|\mathcal{X}\|_2 \sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|_2^2}$ (Streeter & McMahan, 2010)
AdaGrad	Diagonal Matrices	$\|\boldsymbol{x}\|_\infty$	$\sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}$	$\|\mathcal{X}\|_\infty \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}$ (Streeter & McMahan, 2010)
Full-matrix AdaGrad	All Matrices	$\|\boldsymbol{x}\|_2$	$\mathrm{Tr}[(\sum_{t=1}^{T} \boldsymbol{g}_t \boldsymbol{g}_t^\top)^{\frac{1}{2}}]$	$\|\mathcal{X}\|_2 \mathrm{Tr}\left[\left(\sum_{t=1}^{T} \boldsymbol{g}_t \boldsymbol{g}_t^\top\right)^{\frac{1}{2}}\right]$ (Duchi et al., 2011)
One-sided Shampoo	$\mathbb{R}^{d_L \times d_L} \otimes \boldsymbol{I}_{d_R}$	$\frac{\|\boldsymbol{X}\|_{\mathrm{op}}}{\sqrt{d_R}}$	$\mathrm{Tr}[\left(d_R \sum_{t=1}^{T} \boldsymbol{G}_t \boldsymbol{G}_t^\top\right)^{\frac{1}{2}}]$	$\|\mathcal{X}\|_{\mathrm{op}} \mathrm{Tr}\left[\left(\sum_{t=1}^{T} \boldsymbol{G}_t \boldsymbol{G}_t^\top\right)^{\frac{1}{2}}\right]$ (Theorem 4.2)

Table 1. Regret bound obtained by Theorem 3.4 for different algorithms. Theorem 4.2 proves the rate specifically for one-sided shampoo and other rows match existing results in literature. $\mathcal{H}$ is chosen as $\mathcal{K} \cap \mathcal{S}_+^d$. In the third column, $\boldsymbol{x}$ denotes the parameter in vector form and $\boldsymbol{X}$ denotes the parameter in matrix form with $\boldsymbol{x} = \overline{\mathrm{vec}}(\boldsymbol{X})$. In the fourth column, $\boldsymbol{g}_t$ denotes the gradient in vector form and $\boldsymbol{G}_t$ denotes the gradient in matrix form with $\boldsymbol{g}_t = \overline{\mathrm{vec}}(\boldsymbol{G}_t)$. We omit $O(\cdot)$ in complexity measures and regret bound for convenience.

than a more memory- and computationally efficient variant of these methods like diagonal AdaGrad and AdaGrad-Norm (Streeter & McMahan, 2010; Duchi et al., 2011; Ward et al., 2020). This demonstrates that more structure on the preconditioner may not necessarily restrict its optimization performance. In fact, our unified analysis uncovers an interesting finding: we show a simpler, more structured variant of Shampoo, called one-sided Shampoo, actually has substantially better regret bound compared to original Shampoo and full-matrix Adagrad (Section 3.3).

1.1. Main contributions

In light of the above discussion, we highlight the main contributions of the paper.

- We present a comprehensive and unified theoretical framework (Theorem 2.1) for adaptive optimization with structured preconditioners, encompassing several popular methods including diagonal AdaGrad, full-matrix AdaGrad, AdaGrad-Norm, and layerwise adaptive methods for both online convex optimization and stochastic smooth convex optimization. In particular, our analysis integrates and extends key insights from existing work (Gupta et al., 2017), which presents a unified and elegant way (Algorithm 1) to derive the aforementioned structured preconditioners, but only allows analysis on a case-by-case basis. To our knowledge, this is the *first* truly unified analysis of a large family of adaptive optimization algorithms.

- To enable a unified analysis, we identify a novel sufficient condition, *well-structured preconditioners* (Definition 3.1), which overcomes a key technical barrier in Gupta et al. (2017); thereby, allowing a unified analysis for several important adaptive algorithms. A more detailed discussion is provided in Section 2 and a summary of our results is presented in Table 1.

- Our unified analysis provides a new regret bound (Theorem 4.2) for a one-sided variant of Shampoo, which is always better than the existing bound for two-sided Shampoo (Gupta et al., 2018) and could be smaller by a multiplicative factor of d, where we assume the matrix-shaped parameter is of size $\sqrt{d}$-by-$\sqrt{d}$. This also leads to a novel convergence rate of one-sided shampoo for stochastic convex optimization (Theorem 4.4) via a standard offline-to-online reduction (Levy et al., 2018).

- Conceptually, our findings challenge the conventional wisdom that using a larger set of preconditioners which require more memory and compute leads to better optimization performance in terms of number of steps. In particular, while using a larger set of preconditioners reduces the gradient term $\|\|\|\boldsymbol{g}_{1:t}\|\|\|_{\mathcal{H}}$ in our regret bound, it increases the other term involving the magnitude of the optimal solution by increasing the norm metric; thus, leading to a worse regret bound (Theorem 2.1 and Table 1). For instance, we demonstrate that one-sided Shampoo can outperform full-matrix AdaGrad both theoretically (Section 4.3) and experimentally (Section 5). This suggest that one-sided Shampoo is not just a computational-efficient surrogate of full-matrix AdaGrad, but could also be fundamentally better, depending on the optimization problem.

1.2. Notations

Let $\mathcal{M}^d$ be the set of all d-by-d matrices, and $\mathcal{S}^d \subset \mathcal{M}^d$ be the subset of all symmetric matrices. We use $\mathcal{S}_+^d$ to denote the set of positive semi-definite matrices, and $\mathcal{S}_{++}^d$ to denote the set of positive definite matrices. $\mathcal{D}^d$ is the set of all d-dimensional diagonal matrices, and $\mathcal{D}_+^d = \mathcal{D}^d \cap \mathcal{S}_+^d$. We denote by $\boldsymbol{I}_d$ the d-by-d identity matrix. For matrices $\boldsymbol{A}, \boldsymbol{B}$, we denote their inner product by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$.

For any $\boldsymbol{H} \in \mathcal{S}_+^d$ such that $\boldsymbol{H} \neq 0$, we denote $\overline{H} = \boldsymbol{H} / \mathrm{Tr}(\boldsymbol{H})$. For $\boldsymbol{H} \in \mathcal{S}_+^d$, $\|\boldsymbol{x}\|_{\boldsymbol{H}} := \sqrt{\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x}}$ is the

(semi-)norm of $\boldsymbol{x} \in \mathbb{R}^d$ with respect to $\boldsymbol{H}$. For a convex set $\mathcal{H} \subseteq \mathcal{S}_+^d$, we define

$$\|\boldsymbol{x}\|_{\mathcal{H}} := \sup_{\boldsymbol{H} \in \mathcal{H}, \mathrm{Tr}(\boldsymbol{H}) \le 1} \|\boldsymbol{x}\|_{\boldsymbol{H}}. \tag{1}$$

For a convex set $\mathcal{X} \subseteq \mathbb{R}^d$ and any norm $\|\cdot\|$, we define $\|\mathcal{X}\| := \sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|$ and $\|\mathcal{X}\|_{\mathcal{H}} := \sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|_{\mathcal{H}}$. For any $\boldsymbol{H} \succ 0$, the projection of $\boldsymbol{x}$ onto $\mathcal{X}$ with respect to $\|\cdot\|_{\boldsymbol{H}}$ is defined as $\Pi_{\mathcal{X}}^{\boldsymbol{H}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|_{\boldsymbol{H}}$.

Throughout the paper, we consider the factorization $d = d_L d_R$, and denote the corresponding matrix form of $\boldsymbol{x} \in \mathbb{R}^d$ by $\boldsymbol{X} \in \mathbb{R}^{d_L \times d_R}$. We denote $\boldsymbol{x} = \overline{\mathrm{vec}}(\boldsymbol{X})$ and $\boldsymbol{X} = \overline{\mathrm{vec}}^{-1}(\boldsymbol{x})$ for conversion between the vector form and the matrix form. Then for a function $L(\boldsymbol{x})$ defined on $\boldsymbol{x} \in \mathbb{R}^d$, we extend its definition to matrices by letting $L(\boldsymbol{X})$ denote $L(\overline{\mathrm{vec}}(\boldsymbol{X}))$, and we will use $L(\boldsymbol{x})$ and $L(\boldsymbol{X})$ interchangeably when the context is clear. We define the gradient as $\boldsymbol{g}_t = \nabla L(\boldsymbol{x}_t)$ in vector form and $\boldsymbol{G}_t = \nabla L(\boldsymbol{X}_t)$ in matrix form, so $\boldsymbol{g}_t = \overline{\mathrm{vec}}(\boldsymbol{G}_t)$ and $\boldsymbol{G}_t = \overline{\mathrm{vec}}^{-1}(\boldsymbol{g}_t)$.

For a matrix $\boldsymbol{X} \in \mathbb{R}^{d_L \times d_R}$, we denote its operator norm as $\|\boldsymbol{X}\|_{\mathrm{op}}$ and its Frobenious norm as $\|\boldsymbol{X}\|_{\mathrm{F}}$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we denotes its ℓ_∞ norm by $\|\boldsymbol{x}\|_\infty = \max_{i \in [d]} |x_i|$ and ℓ_2 norm by $\|\boldsymbol{x}\|_2$.

2. Background: Unified Adaptive Regularization with Non-Unified Analysis

Seminal work Gupta et al. (2017) presented an adaptive regularization meta-algorithm, AdaReg (Algorithm 1), which can be used to derive various adaptive optimization algorithms known at that time in a unified approach. For example, AdaReg becomes full-matrix AdaGrad (Duchi et al., 2011), diagonal AdaGrad (Duchi et al., 2011), and AdaGrad-Norm (Streeter & McMahan, 2010; Ward et al., 2020) by choosing the set of preconditioners as the set of all PSD matrices, diagonal PSD matrices, and mutipliers of identity matrix respectively (see Table 1). The original AdaReg also allows other choices of potential function Φ, e.g., $\Phi(\cdot) = \log\det(\cdot)$ for Online Newton Step (Hazan et al., 2007), while we are only interested in the case of $\Phi(\cdot) = \eta^2 \mathrm{Tr}(\cdot)$ in this work.

In addition to the unified approach to deriving various adaptive optimization algorithms, Gupta et al. (2017) also attempts to give a unified analysis for the convergence rate or regret of these adaptive algorithms, which can be summarized by the following theorem.

Theorem 2.1 (Gupta et al. (2017))**.** *Let $\{\boldsymbol{x}_t\}_{t=1}^T$ be the iterates of Algorithm 1. Then for any $\boldsymbol{x}^* \in \mathcal{X}$,*

$$\begin{aligned}\sum_{t=1}^T L_t(\boldsymbol{x}_t) - \sum_{t=1}^T L_t(\boldsymbol{x}^*) &\le \frac{1}{2}\left(\langle \boldsymbol{M}_T, \boldsymbol{H}_T^{-1}\rangle + \eta^2 \mathrm{Tr}(\boldsymbol{H}_T) - \eta^2 \mathrm{Tr}(\boldsymbol{H}_0)\right) \\ &\quad + \frac{1}{2}\sum_{t=1}^T \left(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\boldsymbol{H}_t}^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_{\boldsymbol{H}_t}^2\right).\end{aligned} \tag{2}$$

Algorithm 1 Adaptive Regularization Meta-Algorithm AdaReg (Gupta et al., 2017)

Hyperparam: $\epsilon > 0$, convex set $\mathcal{X} \subseteq \mathbb{R}^d$, learning rate η, preconditioners $\mathcal{H} \subset \mathcal{S}_+^d$
Input: initialization $\boldsymbol{x}_1$, loss functions $\{L_t\}_{t=1}^T : \mathbb{R}^d \to \mathbb{R}$
$\boldsymbol{M}_0 \leftarrow \epsilon \boldsymbol{I}_d$
for $t = 1, 2, \ldots, T$ **do**
  $\boldsymbol{g}_t \leftarrow \nabla L_t(\boldsymbol{x}_t)$
  $\boldsymbol{M}_t \leftarrow \boldsymbol{M}_{t-1} + \boldsymbol{g}_t \boldsymbol{g}_t^\top$
  $\boldsymbol{H}_t \leftarrow \arg\min_{\boldsymbol{H} \in \mathcal{H}} \langle \boldsymbol{M}_t, \boldsymbol{H}^{-1}\rangle + \eta^2 \mathrm{Tr}(\boldsymbol{H})$
  $\boldsymbol{x}_{t+1} \leftarrow \Pi_{\mathcal{X}}^{\boldsymbol{H}_t}\left(\boldsymbol{x}_t - \boldsymbol{H}_t^{-1}\boldsymbol{g}_t\right)$
Return $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$

The above bound is obtained by first applying a standard bound for online mirror descent to get a bound in the form of $\sum_{t=1}^T \|\boldsymbol{g}_t\|_{\boldsymbol{H}_t}$ plus the second term on the RHS of (2). Then the choice of $\boldsymbol{H}_t$ in Algorithm 1 enables the application of FTL-BTL lemma (Kalai & Vempala, 2005) to further bound $\sum_{t=1}^T \|\boldsymbol{g}_t\|_{\boldsymbol{H}_t}$ by the first term on the RHS of (2).

To proceed from Theorem 2.1, Gupta et al. (2017) relies on a crucial assumption that $\boldsymbol{H}_{t-1} \preceq \boldsymbol{H}_t$ for each t. Or more generally, for any $\boldsymbol{M} \succ 0$, define

$$P_{\mathcal{H}}(\boldsymbol{M}) := \arg\min_{\boldsymbol{H} \in \mathcal{H}} \langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \eta^2 \mathrm{Tr}(\boldsymbol{H}). \tag{3}$$

Then we hope it holds that

$$P_{\mathcal{H}}(\boldsymbol{M}) \preceq P_{\mathcal{H}}(\boldsymbol{M}') \quad \text{for any} \quad 0 \prec \boldsymbol{M} \preceq \boldsymbol{M}' \tag{4}$$

In other words, we need $P_{\mathcal{H}} : \mathcal{S}_{++}^d \to \mathcal{S}_{++}^d$ to be *operator monotone* under the semi-definite ordering. With this assumption, a critical step in the derivation of the regret bound in Gupta et al. (2017) is to further rewrite and upper bound the second term on the RHS of Equation (2) by

$$\begin{aligned}&\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|_{\boldsymbol{H}_1}^2 + \sum_{t=2}^T \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\boldsymbol{H}_t - \boldsymbol{H}_{t-1}}^2 \\ &\le 4\|\mathcal{X}\|_{\boldsymbol{H}_1}^2 + 4\sum_{t=2}^T \|\mathcal{X}\|_{\boldsymbol{H}_t - \boldsymbol{H}_{t-1}}^2.\end{aligned} \tag{5}$$

Note that we need $\boldsymbol{H}_{t-1} \preceq \boldsymbol{H}_t$ to ensure that $\|\cdot\|_{\boldsymbol{H}_t - \boldsymbol{H}_{t-1}}^2$ is indeed a (pseudo) norm and that the last inequality holds. Such analysis has been done for a few notable variants of AdaGrad in Gupta et al. (2017), where the condition $\boldsymbol{H}_{t-1} \preceq \boldsymbol{H}_t$ is verified in a case-by-case way for specific choice of $\mathcal{H}$. However, the following question remains unclear for optimizers described by general $\mathcal{H}$:

Question 1. *For a cone $\mathcal{H} \subseteq \mathcal{S}_+^d$, does $P_{\mathcal{H}}(\boldsymbol{M}) \preceq P_{\mathcal{H}}(\boldsymbol{M}')$ hold whenever $0 \prec \boldsymbol{M} \preceq \boldsymbol{M}'$?*

Indeed, the answer is *no* for "ill-structured" $\mathcal{H}$, and we mention two negative examples here for illustration.

Example 2.2. Let $\mathcal{H} = \{\boldsymbol{A} \otimes \boldsymbol{B} \succeq 0 \mid \boldsymbol{A} \in \mathcal{M}^{d_L}, \boldsymbol{B} \in \mathcal{M}^{d_R}\}$ with $d_L d_R = d$, i.e., the set of preconditioners for

two-sided Shampoo (Algorithm 3). We show that in the special case where $d_L = d_R = 2$, for $\boldsymbol{M} = \mathrm{diag}(1, \epsilon, \epsilon, \epsilon)$ and $\boldsymbol{M}' = \mathrm{diag}(1, \epsilon, \epsilon, 1)$, $P_{\mathcal{H}}(\boldsymbol{M}) \preceq P_{\mathcal{H}}(\boldsymbol{M}')$ does not hold for sufficiently small $\epsilon > 0$, although we have $\boldsymbol{M} \preceq \boldsymbol{M}'$. See Appendix A.4.1 for a detailed proof.

Example 2.3. The second example of ill-structured $\mathcal{H}$ involves tridiagonal PSD matrices, i.e., matrices that only have nonzero elements on the main diagonal, the first diagonal above the main diagonal, and the first diagonal below the main diagonal. Specifically, for $\mathcal{H}$ containing all 3-dimensional PSD matrices, we provide numerical evidence demonstrating that the desired condition Equation (4) breaks for very simple instances. See details in Appendix A.4.2.

These failure modes naturally lead to the second question:

Question 2. *Is there a sufficient yet general condition (which covers all existing examples) on $\mathcal{H}$ for the inequality to hold?*

As one of the main contributions of our work, we give an affirmative answer to Question 2 in Section 3 by proposing a notion of well-structured preconditioners (Definition 3.1) and deriving a unified analysis correspondingly.

3. Unified Analysis for Well-Structured Preconditioners

We establish a unified framework for adaptive optimization with structured preconditioners. In Section 3.1, we propose the notion of well-structured preconditioners and show that they satisfy the desired condition Equation (4). Then in Section 3.2, we present a unified analysis for adaptive optimization with well-structured preconditioners. We discuss several prominent examples in Section 3.3.

3.1. Well-structured preconditioners

For a set of d-by-d matrices $\mathcal{K} \subseteq \mathcal{M}^d$, we say that $\mathcal{K}$ is a *subalgebra* if it is closed under scalar multiplication, matrix addition, and matrix multiplication. More concretely, we require that for any $\alpha \in \mathbb{R}$ and $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{K}$, it holds that $\alpha\boldsymbol{A}, \boldsymbol{A}\boldsymbol{B}, \boldsymbol{A} + \boldsymbol{B} \in \mathcal{K}$. Based on this, we propose the following core concept of our paper.

Definition 3.1 (Well-structured preconditioner sets)**.** $\mathcal{H} \subseteq \mathcal{S}_+^d$ is said to be a *well-structured preconditioner set* if $\mathcal{H} = \mathcal{S}_+^d \cap \mathcal{K}$ for some matrix subalgebra $\mathcal{K} \subseteq \mathcal{M}^d$ with $\boldsymbol{I}_d \in \mathcal{K}$.

As a positive response to Question 2, the following proposition shows that our notion of well-structured preconditioner sets provides a sufficient condition for $P_{\mathcal{H}}(\cdot)$ to be operator monotone. See Appendix A for a proof.

Proposition 3.2. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. For any $\boldsymbol{M} \succ 0$, there exists a unique solution $P_{\mathcal{H}}(\boldsymbol{M}) \succ 0$ to the optimization problem in (3). Furthermore, for any $\boldsymbol{M} \succ 0$, $P_{\mathcal{H}}(\boldsymbol{M})$ satisfies the following properties:*

(a) $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1} \rangle = \eta^2 \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$.

(b) $\overline{P_{\mathcal{H}}(\boldsymbol{M})} = \arg\min_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \langle \boldsymbol{M}, \boldsymbol{H}^{-1} \rangle$, *where we recall that* $\overline{P_{\mathcal{H}}(\boldsymbol{M})} = P_{\mathcal{H}}(\boldsymbol{M}) / \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$.

Moreover, for any $0 \prec \boldsymbol{M} \preceq \boldsymbol{M}'$, $P_{\mathcal{H}}(\boldsymbol{M}') - P_{\mathcal{H}}(\boldsymbol{M}) \in \mathcal{H}$ holds. In particular, it implies that $P_{\mathcal{H}}(\boldsymbol{M}) \preceq P_{\mathcal{H}}(\boldsymbol{M}')$.

The closure under both matrix addition and matrix multiplication is crucial in the proof of Proposition 3.2. Violation of any of the two properties could lead to problems: for the preconditioner set of two-sided Shampoo in Example 2.2, it is not closed under matrix addition; while for Example 2.3 regarding tridiagonal matrices, we note that the set of tridiagonal matrices is not closed under matrix multiplication.

3.2. Unified analysis for adaptive optimization

To proceed with the regret bound, recall from Proposition 3.2 that $\overline{\boldsymbol{H}}_t = \boldsymbol{H}_t / \operatorname{Tr}(\boldsymbol{H}_t)$ is a solution to the following constrained optimization problem

$$\overline{\boldsymbol{H}}_t = \operatorname*{arg\,min}_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \langle \boldsymbol{M}_t, \boldsymbol{H}^{-1} \rangle. \tag{6}$$

Also recall that $\boldsymbol{M}_t = \sum_{s=1}^t \boldsymbol{g}_s \boldsymbol{g}_s^\top$ (assuming $\epsilon = 0$ for illustration). We interpret the optimal value of this constrained optimization problem as the magnitude of the sequence of gradients $\boldsymbol{g}_{1:t} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_t)$ with respect to the best preconditioner in hindsight. Motivated by this, for any sequence of gradients $\boldsymbol{g}_{1:t}$, we define their *adaptive gradient norm* with respect to $\mathcal{H}$ to be

$$\|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}} := \inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \sqrt{\left\langle \sum_{s=1}^t \boldsymbol{g}_s \boldsymbol{g}_s^\top, \boldsymbol{H}^{-1} \right\rangle}. \tag{7}$$

Indeed, this definition of adaptive gradient norm corresponds to the dual norm of the norm $\|\cdot\|_{\mathcal{H} \otimes \boldsymbol{I}_t}$, denoted as $\|\cdot\|_{\mathcal{H} \otimes \boldsymbol{I}_t}^*$. Specifically, we define

$$\|\,\overline{\mathrm{vec}}(\boldsymbol{g}_{1:t})\|_{\mathcal{H} \otimes \boldsymbol{I}_t}^* = \sup_{\boldsymbol{w} \in \mathbb{R}^{td} : \|\boldsymbol{w}\|_{\mathcal{H} \otimes \boldsymbol{I}_t} \le 1} \overline{\mathrm{vec}}(\boldsymbol{g}_{1:t})^\top \boldsymbol{w}.$$

Then we have $\|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}} = \|\,\overline{\mathrm{vec}}(\boldsymbol{g}_{1:t})\|_{\mathcal{H} \otimes \boldsymbol{I}_t}^* / \sqrt{t}$ by the following lemma, which is proved in Appendix A.3.

Lemma 3.3. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Then for any $t \ge 1$ and $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_t \in \mathbb{R}^d$, it holds that*

$$\inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \sqrt{\left\langle \sum_{s=1}^t \boldsymbol{g}_s \boldsymbol{g}_s^\top, \boldsymbol{H}^{-1} \right\rangle} = \frac{1}{\sqrt{t}} \|\,\overline{\mathrm{vec}}(\boldsymbol{g}_{1:t})\|_{\mathcal{H} \otimes \boldsymbol{I}_t}^*.$$

Given this definition of adaptive gradient norm, combining (6) and the fact that $\operatorname{Tr}(\boldsymbol{H}_t) = \eta^{-2} \langle \boldsymbol{M}_t, \boldsymbol{H}_t^{-1} \rangle$ by Proposition 3.2, we obtain $\operatorname{Tr}(\boldsymbol{H}_t) = \eta^{-1} \|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}}$.

Now recall the upper bound in (5) for the second term in the regret bound (2) from Theorem 2.1. Note that for each t, Proposition 3.2 guarantees that $\boldsymbol{H}_t - \boldsymbol{H}_{t-1} \in \mathcal{H}$, so we can further bound $\|\mathcal{X}\|^2_{\boldsymbol{H}_t - \boldsymbol{H}_{t-1}} \leq \|\mathcal{X}\|^2_{\mathcal{H}} \cdot \mathrm{Tr}(\boldsymbol{H}_t - \boldsymbol{H}_{t-1})$. This allows us to telescope the sum in Equation (5) to get

$$\|\mathcal{X}\|^2_{\boldsymbol{H}_1} + \sum_{t=2}^{T} \|\mathcal{X}\|^2_{\boldsymbol{H}_t - \boldsymbol{H}_{t-1}} = \frac{\|\mathcal{X}\|^2_{\mathcal{H}}}{\eta} \|\boldsymbol{g}_{1:T}\|_{\mathcal{H}} \tag{8}$$

Observe that the upper bound is factored into two parts: 1) $\|\mathcal{X}\|_{\mathcal{H}}$, the diameter of the domain under norm $\|\cdot\|_{\mathcal{H}}$; and 2) $\|\boldsymbol{g}_{1:T}\|_{\mathcal{H}}$, the adaptive gradient norm with respect to $\mathcal{H}$. We pause here for an important remark. Note that $\|\mathcal{X}\|^2_{\boldsymbol{H}_t - \boldsymbol{H}_{t-1}}$ is defined by taking the supremum over $\boldsymbol{x} \in \mathcal{X}$, which precludes telescoping the sum over $t \in [T]$ at first glance. We address this issue by proposing the norm $\|\mathcal{X}\|_{\mathcal{H}}$, which allows us to extract the factor $\mathrm{Tr}(\boldsymbol{H}_t - \boldsymbol{H}_{t-1})$. Again, this unified analysis is possible thanks to Proposition 3.2 for well-structured preconditioner sets, in contrast to the case-by-case analysis done by Gupta et al. (2017). Furthermore, we remark that the above factored bound is crucial for us to identify the correct norm metric for Shampoo, leading to an improved analysis. See Section 4.1 for details.

Finally, combining (8) and the original regret bound in (2) yields the final regret bound for Algorithm 1. This is summarized in the following Theorem 3.4. The complete proof can be found in Appendix D.

Theorem 3.4. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Then for any convex loss functions $L_1, \ldots, L_T$, the regret of Algorithm 1 compared to any $\boldsymbol{x}^* \in \mathcal{X}$ can be bounded as*

$$\sum_{t=1}^{T} L_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} L_t(\boldsymbol{x}^*) \leq \left(\frac{D^2}{2\eta} + \eta\right)\left(G + d\sqrt{\epsilon}\right)$$

where $G = \|\boldsymbol{g}_{1:T}\|_{\mathcal{H}}$, $D = \max_{t\in[T]} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\mathcal{H}}$.

Corollary 3.5. *Under the setting of Theorem 2.1, further suppose that $\mathcal{X}$ is a bounded set in $\mathbb{R}^d$. Then choosing $\eta = \sqrt{2}\|\mathcal{X}\|_{\mathcal{H}}$, the regret bound for Algorithm 1 becomes*

$$\sum_{t=1}^{T} L_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} L_t(\boldsymbol{x}^*) \leq 2\sqrt{2}\|\mathcal{X}\|_{\mathcal{H}}\left(G + d\sqrt{\epsilon}\right).$$

Ignoring ϵ, the bound reveals an intrinsic trade-off between $\|\mathcal{X}\|_{\mathcal{H}}$ and $\|\boldsymbol{g}_{1:t}\|_{\mathcal{H}}$: as $\mathcal{H}$ gets larger, $\|\mathcal{X}\|_{\mathcal{H}}$ increases while $\|\boldsymbol{g}_{1:t}\|_{\mathcal{H}}$ decreases. The previous common belief that more adaptivity (larger $\mathcal{H}$) helps optimization could be largely due to the loose upper bound on $\|\mathcal{X}\|_{\mathcal{H}}$, i.e., always measuring the size of domain $\mathcal{X}$ by Frobenius norm instead of potentially much smaller $\|\cdot\|_{\mathcal{H}}$. The fact that Kroneckered-factored subalgebra induces smaller $\|\cdot\|_{\mathcal{H}}$ than the entire matrix subalgebra is the key reason behind our surprising finding that one-sided Shampoo can outperform full-matrix AdaGrad in terms of the number of steps. See a more detailed analysis in Section 4.3.

Next, we present the convergence rate of Algorithm 1 for stochastic convex smooth loss functions. We use Definition 3.6 to characterize the smoothness of a loss function. It is an extension of Φ-smoothness in (Xie et al., 2025), which only applies to block-diagonal preconditioners (corresponding to blockwise Adam).

Definition 3.6 ($\mathcal{H}$-smoothness)**.** For a set $\mathcal{H} \subseteq \mathcal{S}^d_+$ and any loss function $L : \mathbb{R}^d \to \mathbb{R}$, we define the $\mathcal{H}$-smoothness of L, denoted by $H(L, \mathcal{H})$, as the smallest number $H \geq 0$ such that there exists a matrix $\boldsymbol{H}^* \in \mathcal{H}$ satisfying $H = \mathrm{Tr}(\boldsymbol{H}^*)$ and for any $\boldsymbol{x} \in \mathbb{R}^d$, it holds that $-\boldsymbol{H}^* \preceq \nabla^2 L(\boldsymbol{x}) \preceq \boldsymbol{H}^*$. In the case of convex L, this requirement becomes $\nabla^2 L(\boldsymbol{x}) \preceq \boldsymbol{H}^*$. Furthermore, we extend the notation to matrices $\boldsymbol{A} \in \mathcal{M}^d$ by defining $H(\boldsymbol{A}, \mathcal{H})$ as $H(\boldsymbol{x} \mapsto \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \mathcal{H})$.

We need the following assumption on the stochastic noise.

Assumption 3.7. For any $t \in [T]$ and any $\boldsymbol{x} \in \mathcal{X}$, $\mathbb{E}[L_t(\boldsymbol{x})] = L(\boldsymbol{x})$ and there exists some $\boldsymbol{\Sigma} \in \mathcal{S}^d_+$ such that $\mathbb{E}[(\nabla L_t(\boldsymbol{x}) - \nabla L(\boldsymbol{x}))(\nabla L_t(\boldsymbol{x}) - \nabla L(\boldsymbol{x}))^\top] \preceq \boldsymbol{\Sigma}$.

Now we are ready to state our main results on the convergence rate of Algorithm 1 for stochastic convex functions.

Theorem 3.8. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Consider any independent stochastic convex loss functions $L_1, \ldots, L_T$ satisfying Assumption 3.7, and let $H(L, \mathcal{H})$ be the $\mathcal{H}$-smoothness of their expectation L. Suppose the global minimizer of L, denoted by $\boldsymbol{x}^*$, is in $\mathcal{X}$. Then for the iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ of Algorithm 1, denoting $\bar{\boldsymbol{x}}_{1:T} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$, it holds that*

$$\begin{aligned}&\mathbb{E}\left[L(\bar{\boldsymbol{x}}_{1:T}) - L(\boldsymbol{x}^*)\right]\\ \leq &\frac{16}{T}\|\mathcal{X}\|^2_{\mathcal{H}} H(L, \mathcal{H}) + \frac{4\sqrt{2}}{\sqrt{T}}\|\mathcal{X}\|_{\mathcal{H}}\sigma + \frac{4\sqrt{2}d\sqrt{\epsilon}}{T}\|\mathcal{X}\|_{\mathcal{H}}\end{aligned}$$

where $\sigma = \inf_{\boldsymbol{H}\in\mathcal{H}, \mathrm{Tr}(\boldsymbol{H})\leq 1} \sqrt{\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1}\rangle}$.

Our analysis naturally extends to Algorithm 4, which replaces the direct sum of past gradient outer products in Algorithm 1 with an exponential moving average (EMA). This modification is widely used in adaptive optimizers, including Adam (Kingma & Ba, 2014) and AdaSGD (Wang & Wiens, 2020), which is an ema version of AdaGrad. The detailed discussion is in Appendix C.

3.3. Examples of well-structured preconditioner sets

Next, we demonstrate that our Definition 3.1 is general enough to cover existing examples, by discussing several important matrix subalgebra $\mathcal{K}$. For each associated well-structured $\mathcal{H} = \mathcal{K} \cap \mathcal{S}^d_+$, recall the optimization problem

over $\mathcal{H}$ defined in (3). We will show that every minimizer $P_{\mathcal{H}}(\boldsymbol{M})$ corresponds to the preconditioner used in a specific adaptive optimization algorithm. We list the correspondence relationship below and the results are summarized in Table 1. The detailed derivation and calculation of the norms $\|\boldsymbol{x}\|_{\mathcal{H}}$ and $\|\|\boldsymbol{g}_{1:t}\|\|_{\mathcal{H}}$ can be found in Appendix B.

Example 3.9 (AdaGrad-Norm: scalar matrices)**.** For the scalar matrix subalgebra $\mathcal{K} = \{c \cdot \boldsymbol{I}_d \mid c \in \mathbb{R}\}$, we have $\mathcal{H} = \{c \cdot \boldsymbol{I}_d \mid c \geq 0\}$. Then solving (3) for $\boldsymbol{M}_t$ yields

$$\boldsymbol{H}_t = \frac{1}{\eta}\sqrt{\mathrm{Tr}(\boldsymbol{M}_t)/d} \cdot \boldsymbol{I}_d = \frac{1}{\eta}\sqrt{\epsilon + \sum_{s=1}^{t} \|\boldsymbol{g}_s\|_2^2 / d} \cdot \boldsymbol{I}_d.$$

This is the preconditioner used in AdaGrad-Norm.

Example 3.10 (Diagonal AdaGrad: diagonal matrices)**.** For the diagonal matrix subalgebra $\mathcal{K} = \mathcal{D}^d$, we have $\mathcal{H} = \mathcal{D}_+^d$. Correspondingly,

$$\boldsymbol{H}_t = \frac{1}{\eta}\mathrm{diag}\left(\sqrt{\epsilon + \sum_{s=1}^{t} \|g_{s,i}\|_2^2} : i \in [d]\right).$$

This is the preconditioner used in diagonal AdaGrad.

Example 3.11 (Full-matrix AdaGrad: all matrices)**.** For $\mathcal{K} = \mathcal{M}^d$, we have $\mathcal{H} = \mathcal{S}_+^d$. In this case, solving (3) for $\boldsymbol{M}_t$ yields that $\boldsymbol{H}_t = \frac{1}{\eta}\boldsymbol{M}_t^{\frac{1}{2}}$, which corresponds to the update rule of full-matrix AdaGrad.

Example 3.12 (One-sided Shampoo: factored matrices)**.** Let d be factored as $d = d_L d_R$. Then for the factored matrix algebra $\mathcal{K} = \mathbb{R}^{d_L \times d_L} \otimes \boldsymbol{I}_{d_R}$, we have $\mathcal{H} = \mathcal{S}_+^{d_L} \otimes \boldsymbol{I}_{d_R}$. Now writing $\boldsymbol{G}_t \in \mathbb{R}^{d_L \times d_R}$ as the matricized version of $\boldsymbol{g}_t \in \mathbb{R}^d$, solving the corresponding problem in (3) leads to

$$\boldsymbol{H}_t = \frac{1}{\eta}\left(\epsilon \cdot \boldsymbol{I}_{d_L} + \frac{1}{d_R}\sum_{s=1}^{t} \boldsymbol{G}_s\boldsymbol{G}_s^\top\right)^{\frac{1}{2}} \otimes \boldsymbol{I}_{d_R},$$

which updates $\boldsymbol{x}_t$ the same as one-sided Shampoo[1] displayed in Algorithm 2, where we write the algorithm in the matrix form for convenience. More specifically, note that $\boldsymbol{g}_t = \overline{\mathrm{vec}}(\boldsymbol{G}_t)$ and $\boldsymbol{H}_t = \boldsymbol{L}_t^{\frac{1}{2}} \otimes \boldsymbol{I}_{d_R}$. Moreover, $\boldsymbol{M}_t$ corresponds to $\boldsymbol{L}_t$ by the fact that $\langle \boldsymbol{M}_t, (\boldsymbol{H}_L \otimes \boldsymbol{I}_{d_R})^{-1}\rangle = \langle \boldsymbol{L}_t, \boldsymbol{H}_L^{-1}\rangle$ for any $\boldsymbol{H}_L \in \mathcal{S}_{++}^{d_L}$.

The detailed calculations of the norms $\|\cdot\|_{\mathcal{H}}$ and $\|\|\cdot\|\|_{\mathcal{H}}$ for the above examples can be found in Appendix B.

Generate new well-structured preconditioner sets. Beyond the previous examples, it is also possible to generate new well-structured preconditioner sets based on existing ones. We discuss in particular an example of layerwise combination for parameters in a neural network. Specifically, for d parameters of an N-layer neural network, decompose $\mathbb{R}^d = \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_N}$ where $d = \sum_{n=1}^N d_n$, and each $\mathbb{R}^{d_n}$ corresponds to the d_n parameters in the n-th layer. For each $n \in [N]$, let $\mathcal{K}_n \subseteq \mathcal{M}^{d_n}$ be a matrix subalgebra. Then we define $\mathcal{K} = \oplus_{n=1}^N \mathcal{K}_n = \{\oplus_{n=1}^N \boldsymbol{A}_n \mid \boldsymbol{A}_n \in \mathcal{K}_n, n \in [N]\}$, and it is easy to verify that $\mathcal{K}$ is also a subalgebra. Then the corresponding well-structured preconditioner set $\mathcal{H} = \oplus_{n=1}^N (\mathcal{K}_n \cap \mathcal{S}_+^{d_n}) = \oplus_{n=1}^N \mathcal{H}_n$ contains preconditioners that apply individual types of transforms to gradient of parameters in different layers[2]. Such $\mathcal{H}$ made up by direct sum of smaller cones also has very compositional property in its induced complexity metrics, namely, $\|\cdot\|_{\mathcal{H}} = \max_{1\leq n\leq N} \|\cdot\|_{\mathcal{H}_n}$ and $\|\|\cdot\|\|_{\mathcal{H}} = \sum_{n=1}^N \|\|\cdot\|\|_{\mathcal{H}_n}$. This operation provides a useful tool for designing layerwise preconditioning methods (Bernstein & Newhouse, 2024a).

Other possible operations that can generate new matrix subalgebra $\mathcal{K}'$ from the original subalgebra $\mathcal{K}$ include taking Kronecker product with the identity matrix, i.e. $\mathcal{K}' = \{\boldsymbol{A}' = \boldsymbol{A} \otimes \boldsymbol{I}_{d'} \mid \boldsymbol{A} \in \mathcal{K}\}$, and rotation by an orthogonal matrix, i.e. $\mathcal{K}' = \{\boldsymbol{A}' = \boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U} \mid \boldsymbol{A} \in \mathcal{K}\}$ where $\boldsymbol{U}$ is an orthogonal matrix.

4. Improved Convergence Analysis for One-sided Shampoo

We now turn to one-sided Shampoo (Algorithm 2), a special example of Algorithm 1. In Section 4.1, we present our main results on its regret bound and convergence rate, and then compare to the previous results for two-sided Shampoo in Section 4.2. In Section 4.3, we present a comprehensive comparison between the regret bound of one-sided Shampoo and those of the AdaGrad variants, which suggests why Shampoo can outperform other adaptive algorithms on some real tasks.

4.1. Our results for one-sided Shampoo

We first characterize the norm $\|\cdot\|_{\mathcal{H}}$ for one-sided Shampoo.

Lemma 4.1 ($\|\cdot\|_{\mathcal{H}}$ for one-sided Shampoo)**.** *Recall from Example 3.12 that for one-sided Shampoo (Algorithm 2), $\mathcal{H} = \mathcal{S}_+^{d_L} \otimes \boldsymbol{I}_{d_R}$ where $d_L d_R = d$. Then for any $\boldsymbol{x} \in \mathbb{R}^d$, it holds that $\|\boldsymbol{x}\|_{\mathcal{H}} = \frac{1}{\sqrt{d_R}}\|\boldsymbol{X}\|_{\mathrm{op}}$ with $\boldsymbol{X} = \overline{\mathrm{vec}}^{-1}(\boldsymbol{x})$, and thus $\|\mathcal{X}\|_{\mathcal{H}} = \frac{1}{\sqrt{d_R}}\|\mathcal{X}\|_{\mathrm{op}}$.*

See Appendix B.4 for the proof.

With this, we can apply Theorem 3.4 to get the regret bound for one-sided Shampoo, as summarized below in Theorem 4.2. See Appendix E.2 for its proof.

Theorem 4.2 (Regret bound for one-sided Shampoo)**.** *For convex functions $L_1, \ldots, L_T$, the regret of one-sided Sham-*

[1] We add a normalized factor $\frac{1}{d_R}$ compared to the $\boldsymbol{L}_t$ in Algorithm 3 so that it can be exactly derived from Algorithm 1.

[2] This can be applied to any partition of the parameters, not only for partition based on layers.

Algorithm 2 One-sided Shampoo

Hyperparam: learning rate $\eta > 0$, convex set $\mathcal{X} \subseteq \mathbb{R}^{d_L \times d_R}$, $\boldsymbol{L}_0 = \epsilon \boldsymbol{I}_{d_L}$ for $\epsilon \geq 0$
Input: initialization $\boldsymbol{x}_0$, stochastic loss functions $\{L_t\}_{t=1}^T : \mathbb{R}^{d_L \times d_R} \to \mathbb{R}$
for $t = 1, 2, \cdots, T$ **do**
  $\boldsymbol{G}_t \leftarrow \nabla L_t(\boldsymbol{X}_{t-1})$
  $\boldsymbol{L}_t \leftarrow \boldsymbol{L}_{t-1} + \frac{1}{d_R} \boldsymbol{G}_t \boldsymbol{G}_t^\top$
  $\boldsymbol{X}_t \leftarrow \Pi_{\mathcal{X}}^{\boldsymbol{L}_t^{\frac{1}{2}} \otimes \boldsymbol{I}_{d_R}} \left(\boldsymbol{X}_{t-1} - \eta_t \boldsymbol{L}_t^{-\frac{1}{2}} \boldsymbol{G}_t \right)$
Return $\boldsymbol{x}_T$

Algorithm 3 Two-sided Shampoo (Gupta et al., 2018)

Hyperparam: learning rate η, convex set $\mathcal{X} \subseteq \mathbb{R}^{d_L \times d_R}$, $\boldsymbol{L}_0 = \epsilon \boldsymbol{I}_{d_L}$, $\boldsymbol{R}_0 = \epsilon \boldsymbol{I}_{d_R}$ for $\epsilon \geq 0$
Input: initialization $\boldsymbol{x}_0$, stochastic loss functions $\{L_t\}_{t=1}^T : \mathbb{R}^{d_L \times d_R} \to \mathbb{R}$
for $t = 1, 2, \cdots, T$ **do**
  $\boldsymbol{G}_t \leftarrow \nabla L_t(\boldsymbol{X}_{t-1})$
  $\boldsymbol{L}_t \leftarrow \boldsymbol{L}_{t-1} + \boldsymbol{G}_t \boldsymbol{G}_t^\top$
  $\boldsymbol{R}_t \leftarrow \boldsymbol{R}_{t-1} + \boldsymbol{G}_t^\top \boldsymbol{G}_t$
  $\boldsymbol{X}_t \leftarrow \Pi_{\mathcal{X}}^{\boldsymbol{L}_t^{\frac{1}{4}} \otimes \boldsymbol{R}_t^{\frac{1}{4}}} \left(\boldsymbol{X}_{t-1} - \eta_t \boldsymbol{L}_t^{-\frac{1}{4}} \boldsymbol{G}_t \boldsymbol{R}_t^{-\frac{1}{4}} \right)$
Return $\boldsymbol{x}_T$

poo (Algorithm 2) compared to any $\boldsymbol{X}^* \in \mathbb{R}^{d_L \times d_R}$ satisfies

$$\sum_{t=1}^T L_t(\boldsymbol{X}_t) - \sum_{t=1}^T L_t(\boldsymbol{X}^*) \leq \left(\frac{D_{\text{op}}^2}{2 d_R \eta} + \eta \right) \left(G + d\sqrt{\epsilon} \right),$$

where $D_{\text{op}} = \max_{t \in [T]} \|\boldsymbol{X}_t - \boldsymbol{X}^*\|_{\text{op}}$ and $G = \sqrt{d_R} \operatorname{Tr}\left[\left(\sum_{t=1}^T \boldsymbol{G}_t \boldsymbol{G}_t^\top \right)^{\frac{1}{2}} \right]$. When the domain $\mathcal{X}$ is bounded in operator norm, i.e., $\|\mathcal{X}\|_{\text{op}} < \infty$, further choosing $\eta = \sqrt{2/d_R} \|\mathcal{X}\|_{\text{op}}$, it holds

$$\begin{aligned} &\sum_{t=1}^T L_t(\boldsymbol{X}_t) - \sum_{t=1}^T L_t(\boldsymbol{X}^*) \\ \leq & 2\sqrt{2} \|\mathcal{X}\|_{\text{op}} \left(\operatorname{Tr}\left[\left(\sum_{t=1}^T \boldsymbol{G}_t \boldsymbol{G}_t^\top \right)^{\frac{1}{2}} \right] + \frac{d}{\sqrt{d_R}} \sqrt{\epsilon} \right). \end{aligned}$$

Next, before presenting the convergence rate of one-sided Shampoo, we first provide a more interpretable formulation of the $\mathcal{H}$-smoothness for one-sided Shampoo, which we call *left smoothness*. See Appendix E.1 for a proof of Lemma 4.3.

Lemma 4.3 (Left smoothness for one-sided Shampoo)**.** *Let $\mathcal{H} = \mathcal{S}_+^{d_L} \otimes \boldsymbol{I}_{d_R}$ be the well-structured preconditioner set for one-sided Shampoo. Then the $\mathcal{H}$-smoothness $H(L, \mathcal{H})$ defined in Definition 3.6 is equal to the smallest number $H \geq 0$ such that there exists $\boldsymbol{H}_{d_L}^* \in \mathbb{R}^{d_L \times d_L}$ satisfying that $H = d_R \operatorname{Tr}(\boldsymbol{H}_{d_L}^*)$ and that for any $\boldsymbol{X}, \boldsymbol{\Delta} \in \mathbb{R}^{d_L \times d_R}$,*

$$\left| \nabla^2 L(\boldsymbol{X})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] \right| \leq \left\langle \boldsymbol{H}_{d_L}^*, \boldsymbol{\Delta} \boldsymbol{\Delta}^\top \right\rangle.$$

In this case, the $\mathcal{H}$-smoothness is denoted by $H_{\text{left}}(L)$.

Then the convergence rate of one-sided Shampoo can be obtained by specializing Theorem 3.8 to $\|\mathcal{X}\|_{\mathcal{H}} = \frac{1}{\sqrt{d_R}} \|\mathcal{X}\|_{\text{op}}$ from Lemma 4.1 and the left smoothness from Lemma 4.3.

Theorem 4.4 (Convergence rate of one-sided Shampoo)**.** *Let $L_1, \ldots, L_T$ be stochastic convex loss functions satisfying Assumption 3.7, and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ be the corresponding iterates of one-sided Shampoo (Algorithm 2) with learning rate $\eta = \sqrt{2} \|\mathcal{X}\|_{\text{op}}$. Then for $\bar{\boldsymbol{x}}_{1:T} = \frac{1}{T} \sum_{t=1}^T \boldsymbol{x}_t$,*

$$\begin{aligned} &\mathbb{E}[L(\bar{\boldsymbol{x}}_{1:T}) - L(\boldsymbol{x}^*)] \\ \leq & \frac{16}{T d_R} \|\mathcal{X}\|_{\text{op}}^2 H_{\text{left}}(L) + \frac{4\sqrt{2}\sigma}{\sqrt{T d_R}} \|\mathcal{X}\|_{\text{op}} + \frac{4\sqrt{2} d \sqrt{\epsilon}}{T} \|\mathcal{X}\|_{\text{op}} \end{aligned}$$

where $\sigma = \inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \leq 1} \sqrt{\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1} \rangle}$, and $H_{\text{left}}(L)$ is the left smoothness of the expected loss L identified in Lemma 4.3.

4.2. Comparison with previous results on Shampoo

We compare our main results for one-sided Shampoo with the original results in Gupta et al. (2018). Here, we restate their original regret bound for easier comparison.

Theorem 4.5 (Regret bound of two-sided Shampoo (Gupta et al., 2018))**.** *For convex functions $\{L_t\}_{t=1}^T$, suppose their gradients $(\boldsymbol{G}_t = \nabla L_t(\boldsymbol{X}_t))_{t=1}^T$ are matrices of rank at most r. Then the regret of two-sided Shampoo (Algorithm 3[3]) compared to any $\boldsymbol{X}^* \in \mathbb{R}^{d_L \times d_R}$ is bounded as*

$$\sum_{t=1}^T L_t(\boldsymbol{X}_t) - \sum_{t=1}^T L_t(\boldsymbol{X}^*) \leq \left(\frac{D_{\text{F}}^2}{2\eta} + r\eta \right) \operatorname{Tr}(\boldsymbol{L}_T^{\frac{1}{4}}) \operatorname{Tr}(\boldsymbol{R}_T^{\frac{1}{4}}),$$

where $D_{\text{F}} = \max_{t \in [T]} \|\boldsymbol{X}_t - \boldsymbol{X}^*\|_{\text{F}}$, $\boldsymbol{L}_T = \epsilon \boldsymbol{I}_{d_L} + \sum_{t=1}^T \boldsymbol{G}_t \boldsymbol{G}_t^\top$, and $\boldsymbol{R}_T = \epsilon \boldsymbol{I}_{d_R} + \sum_{t=1}^T \boldsymbol{G}_t^\top \boldsymbol{G}_t$. When $\|\mathcal{X}\|_{\text{F}} < \infty$, we further choose $\eta = \sqrt{2/r} \|\mathcal{X}\|_{\text{F}}$, then

$$\sum_{t=1}^T L_t(\boldsymbol{X}_t) - \sum_{t=1}^T L_t(\boldsymbol{X}^*) \leq \sqrt{2r} \|\mathcal{X}\|_{\text{F}} \operatorname{Tr}(\boldsymbol{L}_T^{\frac{1}{4}}) \operatorname{Tr}(\boldsymbol{R}_T^{\frac{1}{4}}).$$

We now compare our regret bound in Theorem 4.2 and the original regret bound in Theorem 4.5 by Gupta et al. (2018) when $\epsilon = 0$, i.e., $\boldsymbol{L}_T = \sum_{t=1}^T \boldsymbol{G}_t \boldsymbol{G}_t^\top$ and $\boldsymbol{R}_T = \sum_{t=1}^T \boldsymbol{G}_t^\top \boldsymbol{G}_t$. For a matrix $\boldsymbol{M} \in \mathcal{S}_+^d$, it always holds that $\sqrt{\operatorname{Tr}(\boldsymbol{M})} \leq \operatorname{Tr}(\boldsymbol{M}^{\frac{1}{2}})$. Therefore,

$$\begin{aligned} \operatorname{Tr}(\boldsymbol{L}_T^{\frac{1}{4}}) \operatorname{Tr}(\boldsymbol{R}_T^{\frac{1}{4}}) &\geq \operatorname{Tr}(\boldsymbol{L}_T^{\frac{1}{4}}) \operatorname{Tr}(\boldsymbol{R}_T)^{\frac{1}{4}} = \operatorname{Tr}(\boldsymbol{L}_T^{\frac{1}{4}}) \operatorname{Tr}(\boldsymbol{L}_T)^{\frac{1}{4}} \\ &\geq \operatorname{Tr}(\boldsymbol{L}_T^{\frac{1}{4}}) \|\boldsymbol{L}_T\|_{\text{op}}^{\frac{1}{4}} \geq \operatorname{Tr}(\boldsymbol{L}_T^{\frac{1}{2}}). \end{aligned}$$

[3]The original two-sided Shampoo analysis in (Gupta et al., 2017) is without per step projection to the bounded domain. We adapt their theorem into the projected version in a standard way.

This implies that regret bound in Theorem 4.5 is always no smaller than the regret bound in Theorem 4.2 because $r \geq 1$ and $\|\mathcal{X}\|_{\mathrm{F}} \geq \|\mathcal{X}\|_{\mathrm{op}}$ as $\|\boldsymbol{X}\|_{\mathrm{F}} \geq \|\boldsymbol{X}\|_{\mathrm{op}}$ for any $\boldsymbol{X} \in \mathcal{X}$.

Moreover, in the worst case, the Frobenius norm can be $\sqrt{\min(d_L, d_R)}$ times larger than the operator norm. As a concrete example, suppose $\boldsymbol{G}_t$ satisfies that $\boldsymbol{G}_t[i,j] = 1$ only for $(j-i) \equiv t \pmod{\min(d_L, d_R)}$ and all other elements are zero. Then each $\boldsymbol{G}_t$ has rank $r = \min(d_L, d_R)$. At step $T = d_L d_R$, $\boldsymbol{L}_T = d_R \cdot \min(d_L, d_R) \cdot \boldsymbol{I}_{d_L}$ and $\boldsymbol{R}_T = d_L \cdot \min(d_L, d_R) \cdot \boldsymbol{I}_{d_R}$, and thus $\mathrm{Tr}(\boldsymbol{L}_T^{\frac{1}{4}}) \mathrm{Tr}(\boldsymbol{R}_T^{\frac{1}{4}}) = d_L^{\frac{5}{4}} d_R^{\frac{5}{4}}$ while $\mathrm{Tr}(\boldsymbol{L}_T^{\frac{1}{2}}) = d_L d_R^{\frac{1}{2}}$. In this case, the regret bound of two-sided Shampoo is $\min(d_L, d_R) d_L^{\frac{1}{4}} d_R^{\frac{3}{4}}$ times larger than the our regret bound for one-sided Shampoo.

Duvvuri et al. (2024) introduced CASPR as an alternative to Shampoo for approximating full-matrix AdaGrad and achieved the same regret bound as Theorem 4.5. Our previous comparison thus applies to both Shampoo and CASPR.

4.3. Comparison with AdaGrad variants

Next we show one-sided Shampoo can achieve the best theoretical upper bound of the suboptimality gap for a specific class of functions, where each loss function has the form

$$L(\boldsymbol{X}) = \langle \boldsymbol{H}, (\boldsymbol{X} - \boldsymbol{X}^*)(\boldsymbol{X} - \boldsymbol{X}^*)^\top \rangle \tag{9}$$

where $\boldsymbol{X} \in \mathbb{R}^{d_L \times d_R}$, $\boldsymbol{H} \in \mathcal{S}_+^{d_L}$ with $\mathrm{Tr}(\boldsymbol{H}) \leq 1$, and $\|\boldsymbol{X}^*\|_{\mathrm{op}} \leq 1$. For this function class, we compare the largest possible value of convergence rate of different algorithms given by Theorem 3.8, which is summarized in Table 2.

For each algorithm defined by Algorithm 1 with specific $\mathcal{H}$, we will pick $\mathcal{X} = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_{\mathcal{H}} \leq \|\boldsymbol{x}^*\|_{\mathcal{H}}\}$ so that the global minimizer $\boldsymbol{x}^*$ is reachable. To get the convergence rate, it boils down to calculate $\|\mathcal{X}\|_{\mathcal{H}} = \|\boldsymbol{x}^*\|_{\mathcal{H}}$ and $H(L, \mathcal{H})$ associated with each algorithm. We have already derived the explicit form of $\|\mathcal{X}\|_{\mathcal{H}}$ for each algorithm in Section 3.3, as shown in the third column of Table 2. For the $\mathcal{H}$-smoothness, note that $\nabla^2 L(\boldsymbol{X}) = \boldsymbol{H} \otimes \boldsymbol{I}_{d_R}$ for the loss L in (9), and then it is straightforward to calculate $H(L, \mathcal{H})$ according to Definition 3.6, as shown in the fourth column of Table 2.

Below we present the worst case of the convergence rate on this problem class. We can see that one-sided Shampoo is strictly better the other three adaptive algorithms.

One-sided Shampoo. The worst case of convergence rate for one-sided Shampoo is $\frac{1}{T}$.

AdaGrad-Norm. Since $\max_{\mathrm{Tr}(\boldsymbol{H}) \leq 1} \lambda_{\max}(\boldsymbol{H}) = 1$ and $\max_{\|\boldsymbol{X}^*\|_{\mathrm{op}} \leq 1} \|\overline{\mathrm{vec}}(\boldsymbol{X}^*)\|_2 = \max_{\|\boldsymbol{X}^*\|_{\mathrm{op}} \leq 1} \|(\boldsymbol{X}^*)\|_{\mathrm{F}} = \sqrt{\min\{d_L, d_R\}}$, the worst case of convergence rate is $\min(d_L, d_R)/T$.

AdaGrad. For psd matrix $\boldsymbol{H} \in \mathbb{R}^{d_L \times d_L}$ with $\mathrm{Tr}(\boldsymbol{H}) \leq 1$, it holds that $\boldsymbol{H} \preceq \boldsymbol{I}_{d_L}$. Then we have

$$\max_{\mathrm{Tr}(\boldsymbol{H}) \leq 1} H(\boldsymbol{H}, \mathcal{D}_+^{d_L}) = \max_{\mathrm{Tr}(\boldsymbol{H}) \leq 1} \min_{\mathrm{diag}(\boldsymbol{D}) \succeq \boldsymbol{H}} \mathrm{Tr}(\boldsymbol{D}) \leq d_L.$$

On the other hand, if we choose $\boldsymbol{H} = \mathbf{1}_{d_L} \mathbf{1}_{d_L}^\top / d_L$, for any diagonal $\boldsymbol{D} \succeq \boldsymbol{H}$, it holds that $\mathrm{Tr}(\boldsymbol{D}) = \mathbf{1}_{d_L}^\top \boldsymbol{D} \mathbf{1}_{d_L} \geq \mathbf{1}_{d_L}^\top \boldsymbol{H} \mathbf{1}_{d_L} = d_L$. So we prove that $\max_{\mathrm{Tr}(\boldsymbol{H}) \leq 1} H(\boldsymbol{H}, \mathcal{D}_+^{d_L}) = d_L$.

We also know that $\max_{\|\boldsymbol{X}^*\|_{\mathrm{op}} \leq 1} \|\overline{\mathrm{vec}}(\boldsymbol{X}^*)\|_\infty \leq 1$ since $\|\boldsymbol{A}\|_\infty \leq \|\boldsymbol{A}\|_{\mathrm{op}}$. If the only nonzero entry of $\boldsymbol{X}^*$ is $\boldsymbol{X}^*_{1,1} = 1$, then $\|\boldsymbol{X}^*\|_\infty = \|\boldsymbol{X}^*\|_{\mathrm{op}} = 1$. Overall, the worst case of convergence rate is $\frac{d_L d_R}{T}$.

Full-matrix AdaGrad. We have seen in AdaGrad-Norm that $\max_{\|\boldsymbol{X}^*\|_{\mathrm{op}} \leq 1} \|\overline{\mathrm{vec}}(\boldsymbol{X}^*)\|_2 = \sqrt{\min\{d_L, d_R\}}$, so the worst case of convergence rate is $\frac{\min(d_L, d_R) d_R}{T}$.

To summarize, we have identified a class of optimization problems, namely loss functions like Equation (9), with Hessian of bounded trace and optimizer of bounded spectral norm, for which one-sided Shampoo has much better worst case convergence rate than any other adaptive algorithms that could be derived from our unified analysis.

5. Experiments

In this section we empirically demonstrate the superior performance of 1-sided shampoo over other variants of AdaReg (Algorithm 1) on a simple but natural setting. Moreover, such superior performance is predicted by our theoretical analysis in Section 4.3, which in turn validates the practical utility of our theory in guiding optimizer selection.

Setup. We consider a linear regression problem $\|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{y}\|_2^2$ where $\boldsymbol{A}$ is the data matrix and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{X}^*$ is the label vector generated by ground-truth $\boldsymbol{X}^*$. Thus, the loss function can be equivalently written as

$$f(\boldsymbol{X}) = \left\langle \boldsymbol{H}, (\boldsymbol{X} - \boldsymbol{X}^*)(\boldsymbol{X} - \boldsymbol{X}^*)^\top \right\rangle,$$

which is the same function that we studied in Section 4.3 and show that 1-sided shampoo outperforms other adaptive algorithms. We consider $\boldsymbol{X} \in \mathbb{R}^{d \times d}$ with $d = 10^3$. We set the eigenvalues of $\boldsymbol{H}$ by $\sigma_1 = \cdots = \sigma_{10} = 1$ and $\sigma_i = \frac{1}{(i-10)^2}$ for $11 \leq i \leq 10^3$. Each element of the solution $\boldsymbol{X}^*$ is independently sampled from $\mathcal{N}(0, \frac{1}{d})$. We run AdaGrad-Norm, AdaGrad, one-sided Shampoo and full-matrix AdaGrad for 100 steps from initialization $\boldsymbol{X}_0 = 0$. We also run the original Shampoo algorithm for comparison. Full-matrix AdaGrad is run in a memory-efficient way and the detail is in Appendix F.1. We will compare the last iterate loss and the average iterate loss separately. The learning rate is tuned over five seeds for last iterate loss and average iterate loss respectfully, selecting the one with the

Algorithm	Subalgebra $\mathcal{K}$	$\|\boldsymbol{x}^*\|_{\mathcal{H}}$	$H(L,\mathcal{H})$	Convergence Rate
AdaGrad-Norm	$\{c\boldsymbol{I}_d \mid c\in\mathbb{R}\}$	$\frac{1}{\sqrt{d}}\|\boldsymbol{x}^*\|_2$	$d\cdot\lambda_{\max}(\boldsymbol{H})$	$\|\mathcal{X}\|_2^2\,\lambda_{\max}(\boldsymbol{H})/T$
AdaGrad	Diagonal matrices, $\mathcal{D}^d(\mathbb{R})$	$\|\boldsymbol{x}^*\|_\infty$	$d_R\cdot H(\boldsymbol{H},\mathcal{D}_+^{d_L})$	$\|\mathcal{X}\|_\infty^2\, d_R\cdot H(\boldsymbol{H},\mathcal{D}_+^{d_L})/T$
Full-Matrix AdaGrad	All matrices, $\mathbb{R}^{d\times d}$	$\|\boldsymbol{x}^*\|_2$	$d_R\,\mathrm{Tr}(\boldsymbol{H})$	$\|\mathcal{X}\|_2^2\, d_R\,\mathrm{Tr}(\boldsymbol{H})/T$
One-Sided Shampoo	$\mathbb{R}^{d_L\times d_L}\otimes\boldsymbol{I}_{d_R}$	$\frac{1}{\sqrt{d_R}}\|\boldsymbol{X}^*\|_{\mathrm{op}}$	$d_R\,\mathrm{Tr}(\boldsymbol{H})$	$\|\mathcal{X}\|_{\mathrm{op}}^2\,\mathrm{Tr}(\boldsymbol{H})/T$

Table 2. Convergence rate for the loss function $L(\boldsymbol{X})=\langle\boldsymbol{H},(\boldsymbol{X}-\boldsymbol{X}^*)(\boldsymbol{X}-\boldsymbol{X}^*)^\top\rangle$. The results can be obtained by Theorem 3.8 with $\sigma=0$ and omitting ϵ. For each algorithm we pick the smallest domain which still ensures the minimizer lies in the domain. Here recall that $\mathcal{D}^d$ is the set of all d-dimensional diagonal matrices, and $\mathcal{D}_+^d=\mathcal{D}^d\cap\mathcal{S}_+^d$.

lowest average loss. We use precision float32 and set $\epsilon=0$ for all the experiments.

We also run the EMA version of the adaptive algorithms, whose results are consistent as shown in Appendix F.2.

Results. The overall results are shown in Section 5. We can see that one-sided Shampoo greatly outperforms other algorithms, corroborating the theoretical analysis in Section 4.3. Moreover, full-matrix AdaGrad is apparently the worst, suggesting more or even full adaptivity does not always help optimization for fixed budget of training steps. One-sided Shampoo is also better than the original Shampoo algorithm.

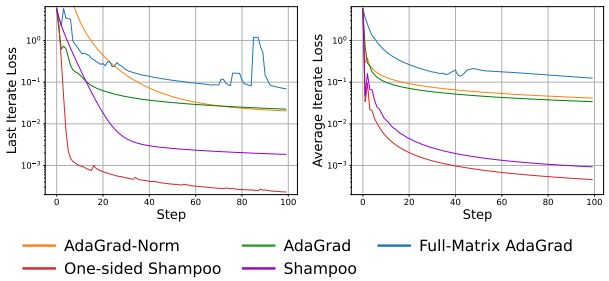

Figure 1. For both the last iterate loss $f(\boldsymbol{X}_t)$ and the average iterate training $f(\frac{1}{t}\sum_{s=1}^t\boldsymbol{X}_s)$, one-sided Shampoo performs the best and Full-matrix AdaGrad performs the worst. Here $f(\boldsymbol{X})=\langle\boldsymbol{H},(\boldsymbol{X}-\boldsymbol{X}^*)(\boldsymbol{X}-\boldsymbol{X}^*)^\top\rangle$.

6. Related Work

Adaptive optimizers and structured preconditioners. The extensive costs for training large-scale deep learning models have motivated the development of efficient optimization algorithms, among which adaptive optimizers have been widely studied because of their ability to exploit the rich geometry of the loss landscape (Pascanu & Bengio, 2013; Martens & Grosse, 2015; Dozat, 2016; Loshchilov & Hutter, 2018; Shazeer & Stern, 2018; Reddi et al., 2019; You et al., 2019; Zhuang et al., 2020; Liu et al., 2023; Yuan et al., 2024). For the sake of memory and computational efficiency, many works involve approximations to full-matrix preconditioners (Ba et al., 2017; George et al., 2018; Martens et al., 2018; Yao et al., 2021; Jahani et al., 2021; Zhang et al., 2022; Duvvuri et al., 2024). Indeed, our results suggest that such compromises might not harm the performance of adaptive optimizers in practice, because more adaptivity is not always helpful, as discussed in Section 4.3 and Section 5.

Understanding Shampoo. There are also recent efforts to understand the Shampoo optimizer from various perspectives. Bernstein & Newhouse (2024b) interpret Shampoo as the steepest descent with respect to the spectral norm of the layerwise matrix-form parameters of the neural network. Recognizing such structures of matrix-form parameters and role of spectral-norm geometry in deep learning has led to development of new optimizer such as Muon (Jordan et al., 2024). In addition, the second-order perspective on Shampoo (Anil et al., 2020) has also led to fruitful results: Morwani et al. (2024) connect the preconditioner in Shampoo to the optimal Kronecker product approximation of the Gauss-Newton component of the Hessian, and Vyas et al. (2024) propose to view Shampoo as Adafactor in the eigenbasis of the preconditioner of Shampoo.

7. Conclusion and Future Works

We present a unified analysis for a broad class of adaptive optimization algorithms with well-structured preconditioners (Definition 3.1) for both online regret minimization and smooth convex optimization. Our analysis not only provides matching rate to several important algorithms including diagonal AdaGrad, full-matrix AdaGrad, and AdaGradNorm, but also gives an improved convergence rate for a one-sided variant of Shampoo over that of the original Shampoo. We reveal a novel trade-off in final convergence rate between domain metric and adaptive gradient norm (Equation (7)) for regret minimization or adaptive smoothness (Definition 3.6) for smooth convex optimization. We hope this insight could be useful towards design of future adaptive optimizers. One important future direction is to identify more subalgebras or other structures that are useful for improving the performance by better adapting to the domain or loss smoothness.

Impact Statement

The goal of this paper is to advance the field of Machine Learning by providing theoretical and experimental insights for adaptive optimization algorithms. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgment

The authors sincerely thank Matt Streeter for his valuable discussions and insightful comments throughout this work.

References

Anil, R., Gupta, V., Koren, T., Regan, K., and Singer, Y. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020. URL https://arxiv.org/pdf/2002.09018.

Ba, J., Grosse, R., and Martens, J. Distributed second-order optimization using kronecker-factored approximations. In *International Conference on Learning Representations*, 2017. URL https://jimmylba.github.io/papers/nsync.pdf.

Bernstein, J. and Newhouse, L. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024a. URL https://arxiv.org/pdf/2410.21265.

Bernstein, J. and Newhouse, L. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024b. URL https://arxiv.org/pdf/2409.20325.

Dahl, G. E., Schneider, F., Nado, Z., Agarwal, N., Sastry, C. S., Hennig, P., Medapati, S., Eschenhagen, R., Kasimbeg, P., Suo, D., et al. Benchmarking neural network training algorithms. *arXiv preprint arXiv:2306.07179*, 2023. URL https://arxiv.org/pdf/2306.07179.

Dozat, T. Incorporating nesterov momentum into adam. 2016. URL https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 2011. URL https://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf.

Duvvuri, S. S., Devvrit, F., Anil, R., Hsieh, C.-J., and Dhillon, I. S. Combining axes preconditioners through kronecker approximation for deep learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=8j9hz8DVi8.

George, T., Laurent, C., Bouthillier, X., Ballas, N., and Vincent, P. Fast approximate natural gradient descent in a kronecker factored eigenbasis. *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/pdf/1806.03884.

Gupta, V., Koren, T., and Singer, Y. A unified approach to adaptive regularization in online and stochastic optimization. *arXiv preprint arXiv:1706.06569*, 2017. URL https://arxiv.org/pdf/1706.06569.

Gupta, V., Koren, T., and Singer, Y. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, 2018. URL https://arxiv.org/pdf/1802.09568.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 2007. URL https://link.springer.com/article/10.1007/s10994-007-5016-8.

Jahani, M., Rusakov, S., Shi, Z., Richtárik, P., Mahoney, M. W., and Takáč, M. Doubly adaptive scaled algorithm for machine learning using second-order information. *arXiv preprint arXiv:2109.05198*, 2021. URL https://arxiv.org/pdf/2109.05198.

Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., and Bernstein, J. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://web.archive.org/web/20250122060345/https://kellerjordan.github.io/posts/muon/.

Kalai, A. and Vempala, S. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 2005. URL https://www.sciencedirect.com/science/article/pii/S0022000004001394.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/pdf/1412.6980.pdf.

Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. *Advances in neural information processing systems*, 2018. URL https://arxiv.org/pdf/1809.02864.

Liu, H., Li, Z., Hall, D. L. W., Liang, P., and Ma, T. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *The Twelfth International Conference on Learning Representations*, 2023. URL https://arxiv.org/pdf/2305.14342.pdf.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/pdf/1711.05101.pdf.

Martens, J. and Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, 2015. URL http://arxiv.org/pdf/1503.05671.

Martens, J., Ba, J., and Johnson, M. Kronecker-factored curvature approximations for recurrent neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/pdf?id=HyMTkQZAb.

Morwani, D., Shapira, I., Vyas, N., Malach, E., Kakade, S., and Janson, L. A new perspective on shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*, 2024. URL https://arxiv.org/pdf/2406.17748.

Pascanu, R. and Bengio, Y. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013. URL https://arxiv.org/pdf/1301.3584.

Petersen, K. B., Pedersen, M. S., et al. The matrix cookbook. *Technical University of Denmark*, 2008. URL https://ece.uwaterloo.ca/~ece602/MISC/matrixcookbook.pdf.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. URL https://arxiv.org/pdf/1904.09237.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, 2018. URL https://arxiv.org/pdf/1804.04235.

Streeter, M. and McMahan, H. B. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010. URL https://arxiv.org/pdf/1002.4862.

Vyas, N., Morwani, D., Zhao, R., Shapira, I., Brandfonbrener, D., Janson, L., and Kakade, S. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024. URL https://arxiv.org/pdf/2409.11321.

Wang, J. and Wiens, J. Adasgd: Bridging the gap between sgd and adam. *arXiv preprint arXiv:2006.16541*, 2020. URL https://arxiv.org/pdf/2006.16541.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 2020. URL https://www.jmlr.org/papers/volume21/18-352/18-352.pdf.

Xie, S., Mohamadi, M. A., and Li, Z. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PUnD86UEK5.

Yao, Z., Gholami, A., Shen, S., Mustafa, M., Keutzer, K., and Mahoney, M. Adahessian: An adaptive second order optimizer for machine learning. In *proceedings of the AAAI conference on artificial intelligence*, 2021. URL https://arxiv.org/pdf/2006.00719.

You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. URL https://arxiv.org/pdf/1708.03888.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019. URL https://arxiv.org/pdf/1904.00962.

Yuan, H., Liu, Y., Wu, S., Zhou, X., and Gu, Q. Mars: Unleashing the power of variance reduction for training large models. *arXiv preprint arXiv:2411.10438*, 2024. URL https://arxiv.org/pdf/2411.10438.

Zhang, L., Shi, S., and Li, B. Eva: Practical second-order optimization with kronecker-vectorized approximation. In *The Eleventh International Conference on Learning Representations*, 2022. URL https://arxiv.org/pdf/2308.02123.

Zhuang, J., Tang, T., Ding, Y., Tatikonda, S. C., Dvornek, N., Papademetris, X., and Duncan, J. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 2020. URL https://arxiv.org/pdf/2010.07468.

A. Proof for Well-Structured Preconditioner Sets

A.1. Definition of $\mathcal{H}$ and its properties

Recall that $\mathcal{K}$ is a subalgebra of d-by-d real-valued matrices, and the corresponding well-structured preconditioner set is $\mathcal{H} = \mathcal{K} \cap \mathcal{S}_+^d$.

Lemma A.1. *For any $\boldsymbol{A} \in \mathcal{K}$ and any polynomial p, $p(\boldsymbol{A}) \in \mathcal{K}$. Furthermore, for any invertible $\boldsymbol{A} \in \mathcal{K}$, its inverse $\boldsymbol{A}^{-1} \in \mathcal{K}$. Also, for any symmetric $\boldsymbol{A} \in \mathcal{K}$, its pseudo inverse $\boldsymbol{A}^{\dagger} \in \mathcal{K}$.*

Proof of Lemma A.1. The first statement follows from the fact that $\mathcal{K}$ is a subalgebra, and the second and third statement are consequences of the Cayley-Hamilton theorem. □

Lemma A.2. *For $\mathcal{H} = \mathcal{K} \cap \mathcal{S}_+^d$ where $\mathcal{K}$ is a subalgebra of d-by-d real-valued matrices, define*

$$\mathcal{H}^* = \{\boldsymbol{A} \in \mathcal{K} : \langle \boldsymbol{H}, \boldsymbol{A} \rangle \geq 0, \forall \boldsymbol{H} \in \mathcal{H}\}. \tag{10}$$

Then $\mathcal{H}^* = \mathcal{H}$. Consequently, for any $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathcal{H}$, if $\langle \boldsymbol{H}_1 - \boldsymbol{H}_2, \boldsymbol{H} \rangle \geq 0$ for all $\boldsymbol{H} \in \mathcal{H}$, then $\boldsymbol{H}_1 \succeq \boldsymbol{H}_2$.

Proof of Lemma A.2. Suppose there exists $\boldsymbol{A} \in \mathcal{H}^*$ such that $\boldsymbol{A}$ has a negative eigenvalue. Let $\lambda_1(\boldsymbol{A}), \ldots, \lambda_d(\boldsymbol{A})$ be the eigenvalues of $\boldsymbol{A}$ in the decreasing order. Consider the matrix $\boldsymbol{B} = (\boldsymbol{A} - 2\max(\lambda_1(\boldsymbol{A}), 1)\boldsymbol{I}_d)^2 \succ 0$. The leading eigenspace of $\boldsymbol{B}$ is the same as the eigenspace of $\boldsymbol{A}$ corresponding to its smallest eigenvalue, which is negative. Consequently, for large enough integer n, $\boldsymbol{B}^n/\|\boldsymbol{B}^n\|_{\mathrm{op}}$ is approximately the projection matrix onto the eigenspace of the smallest negative eigenvalue of $\boldsymbol{A}$. Therefore, for large enough integer n, $\langle \boldsymbol{B}^n/\|\boldsymbol{B}^n\|_{\mathrm{op}}, \boldsymbol{A} \rangle < 0$. However, since $\boldsymbol{B}^n/\|\boldsymbol{B}^n\|_{\mathrm{op}}$ is a polynomial of $\boldsymbol{A}$, we know that $\boldsymbol{B}^n/\|\boldsymbol{B}^n\|_{\mathrm{op}} \in \mathcal{H}$. This is a contradiction to the definition of $\mathcal{H}^*$. Hence, we conclude that for any $\boldsymbol{A} \in \mathcal{H}^*$, it holds that $\boldsymbol{A} \succeq 0$, and thus $\mathcal{H}^* \subseteq \mathcal{H}$. Moreover, for any $\boldsymbol{A} \in \mathcal{H}$, it holds that $\langle \boldsymbol{H}, \boldsymbol{A} \rangle \geq 0$ for any $\boldsymbol{H} \in \mathcal{H}$ because both $\boldsymbol{A}$ and $\boldsymbol{H}$ are positive semi-definite. This shows that $\mathcal{H} \subseteq \mathcal{H}^*$, and hence $\mathcal{H}^* = \mathcal{H}$. This completes the proof. □

A.2. Properties of $P_{\mathcal{H}}(\cdot)$

For any $\boldsymbol{M} \succ 0$, recall the regularized optimization problem in (3). Note that we can assume $\eta = 1$ without loss of generality, because the original problem is equivalent to solve for $\boldsymbol{M}/\eta^2$ with regularizer $\mathrm{Tr}(\boldsymbol{H})$ in place of $\eta^2 \mathrm{Tr}(\boldsymbol{H})$. Therefore, in the rest of this section, we focus on the following optimization problem:

$$P_{\mathcal{H}}(\boldsymbol{M}) := \arg\min_{\boldsymbol{H} \in \mathcal{H}} \langle \boldsymbol{M}, \boldsymbol{H}^{-1} \rangle + \mathrm{Tr}(\boldsymbol{H}). \tag{11}$$

The results proved below applies to the original problem (3) after simple rescaling, and Proposition 3.2 follows from Proposition A.3 below.

For notational convenience, given any $\boldsymbol{M} \succ 0$, we define

$$f_{\boldsymbol{M}}(\boldsymbol{H}) := \langle \boldsymbol{M}, \boldsymbol{H}^{-1} \rangle + \mathrm{Tr}(\boldsymbol{H}). \tag{12}$$

Proposition A.3. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. For any $\boldsymbol{M} \succ 0$, there exists a unique solution $P_{\mathcal{H}}(\boldsymbol{M}) \succ 0$ to the optimization problem in (11). Furthermore, for any $\boldsymbol{M} \succ 0$, $P_{\mathcal{H}}(\boldsymbol{M})$ satisfies the following properties:*

(a) $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1} \rangle = \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$.

(b) $\overline{P_{\mathcal{H}}(\boldsymbol{M})} = \arg\min_{\boldsymbol{H} \in \mathcal{H}, \mathrm{Tr}(\boldsymbol{H}) \leq 1} \langle \boldsymbol{M}, \boldsymbol{H}^{-1} \rangle$ *where we recall that* $\overline{P_{\mathcal{H}}(\boldsymbol{M})} = \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))^{-1} P_{\mathcal{H}}(\boldsymbol{M})$.

(c) For any $\boldsymbol{H} \in \mathcal{H}$, $\langle -P_{\mathcal{H}}(\boldsymbol{M})^{-1} \boldsymbol{M} P_{\mathcal{H}}(\boldsymbol{M})^{-1} + \boldsymbol{I}_d, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M}) \rangle = 0$.

Moreover, for any $\boldsymbol{M}_1 \succeq \boldsymbol{M}_2 \succ 0$, it holds that $P_{\mathcal{H}}(\boldsymbol{M}_1) - P_{\mathcal{H}}(\boldsymbol{M}_2) \in \mathcal{H}$, and in particular, $P_{\mathcal{H}}(\boldsymbol{M}_1) \succeq P_{\mathcal{H}}(\boldsymbol{M}_2)$.

Proof of Proposition A.3. We first show that $P_{\mathcal{H}}(\boldsymbol{M})$ exists and $P_{\mathcal{H}}(\boldsymbol{M}) \succ 0$. Note that for $\mathcal{H} = \mathcal{K} \cap \mathcal{S}_+^d$, since $\mathcal{K}$ is a linear subspace of $\mathcal{M}^d$ and $\mathcal{S}_+^d$ is a closed subset of $\mathcal{M}^d$, we know that $\mathcal{H}$ is also a closed subset of $\mathcal{M}^d$. Moreover, for any sequence $\{\boldsymbol{H}_n\}_{n\geq 1}$ such that $\boldsymbol{H}_n \succ 0$ and either the smallest eigenvalue of $\boldsymbol{H}_n$ converges to 0 or the largest eigenvalue of $\boldsymbol{H}_n$ converges to ∞, the objective value $f_{\boldsymbol{M}}(\boldsymbol{H}_n)$ goes to ∞ because $\boldsymbol{M} \succ 0$. Therefore, there exists $P_{\mathcal{H}}(\boldsymbol{M}) \succ 0$ that attains the minimum objective value. For the uniqueness of $P_{\mathcal{H}}(\boldsymbol{M})$, it suffices to note that the objective function $f_M(\boldsymbol{H})$ is strictly convex in $\boldsymbol{H} \succ 0$ for $\boldsymbol{M} \succ 0$.

Proof for property (a). Suppose otherwise that $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle \neq \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$, and consider the following matrix:

$$\boldsymbol{H} = P_{\mathcal{H}}(\boldsymbol{M}) \cdot \sqrt{\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle / \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))} \in \mathcal{H}.$$

For this $\boldsymbol{H}$, we have $\langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \operatorname{Tr}(\boldsymbol{H}) = 2\sqrt{\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle \cdot \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))} < \langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle + \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$, thus contradicting the optimality of $P_{\mathcal{H}}(\boldsymbol{M})$. Therefore, it must be true that $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle = \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$.

Proof for property (b) Note that we can rewrite the original optimization problem as follows:

$$\begin{aligned} P_{\mathcal{H}}(\boldsymbol{M}) &= \arg\min_{\boldsymbol{H}\in\mathcal{H}} \langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \operatorname{Tr}(\boldsymbol{H}) \\ &= \arg\min_{\boldsymbol{H}\in\mathcal{H}} \left\langle \boldsymbol{M}, \left(\frac{\boldsymbol{H}}{\operatorname{Tr}(\boldsymbol{H})}\right)^{-1}\right\rangle \cdot \frac{1}{\operatorname{Tr}(\boldsymbol{H})} + \operatorname{Tr}(\boldsymbol{H}). \end{aligned}$$

Note that solving the above optimization problem is equivalent to first solving $\overline{P_{\mathcal{H}}(\boldsymbol{M})} = \arg\min_{\boldsymbol{H}\in\mathcal{H},\operatorname{Tr}(\boldsymbol{H})\leq 1} \langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle$ and then setting $P_{\mathcal{H}}(\boldsymbol{M}) = \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) \cdot \overline{P_{\mathcal{H}}(\boldsymbol{M})}$ where the value of $\operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$ ensures the previous property (a). Hence, we see that $\overline{P_{\mathcal{H}}(\boldsymbol{M})}$ solves the constrained version of the original optimization problem.

Proof for property (c). Since $\nabla_{\boldsymbol{H}} \operatorname{Tr}(\boldsymbol{H}) = \boldsymbol{I}_d$ and $\nabla_{\boldsymbol{H}} \operatorname{Tr}(\boldsymbol{M}^\top \boldsymbol{H}^{-1}) = -\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1}$ (see e.g. Equation (124) in Petersen et al. (2008)), we have

$$\nabla_{\boldsymbol{H}} f_{\boldsymbol{M}}(\boldsymbol{H}) = -\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1} + \boldsymbol{I}_d. \tag{13}$$

Then as $\mathcal{H}$ is a cone, by the optimality of $P_{\mathcal{H}}(\boldsymbol{M})$, it holds for any $\boldsymbol{H} \in \mathcal{H}$ that

$$0 = \langle \nabla_{\boldsymbol{H}} f_{\boldsymbol{M}}(P_{\mathcal{H}}(\boldsymbol{M})), \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M})\rangle = \langle -P_{\mathcal{H}}(\boldsymbol{M})^{-1}\boldsymbol{M}P_{\mathcal{H}}(\boldsymbol{M})^{-1} + \boldsymbol{I}_d, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M})\rangle$$

Proof for the operator monotonicity of $P_{\mathcal{H}}(\cdot)$. By property (c) of $P_{\mathcal{H}}(\cdot)$, we know that for any $\boldsymbol{M} \succ 0$, $\langle \boldsymbol{M} - P_{\mathcal{H}}(\boldsymbol{M})^2, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\boldsymbol{H}P_{\mathcal{H}}(\boldsymbol{M})^{-1} - P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle = 0$. Note that $\boldsymbol{H} \mapsto P_{\mathcal{H}}(\boldsymbol{M})^{-1}\boldsymbol{H}P_{\mathcal{H}}(\boldsymbol{M})^{-1}$ is a bijection from $\mathcal{H}$ to $\mathcal{H}$ by Lemma A.1, so we have $\langle \boldsymbol{M} - P_{\mathcal{H}}(\boldsymbol{M})^2, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle = 0$ for all $\boldsymbol{H} \in \mathcal{H}$. Applying this to both $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, it follows that for any $\boldsymbol{H} \in \mathcal{H}$

$$\langle \boldsymbol{M}_1 - P_{\mathcal{H}}(\boldsymbol{M}_1)^2, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M}_1)^{-1}\rangle = \langle \boldsymbol{M}_2 - P_{\mathcal{H}}(\boldsymbol{M}_2)^2, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M}_2)^{-1}\rangle.$$

Rearranging the above equation, we obtain

$$\begin{aligned} \langle P_{\mathcal{H}}(\boldsymbol{M}_1)^2 - P_{\mathcal{H}}(\boldsymbol{M}_2)^2, \boldsymbol{H}\rangle &= \langle \boldsymbol{M}_1 - \boldsymbol{M}_2, \boldsymbol{H}\rangle - \langle \boldsymbol{M}_1, P_{\mathcal{H}}(\boldsymbol{M}_1)^{-1}\rangle + \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}_1)) \\ &\quad + \langle \boldsymbol{M}_2, P_{\mathcal{H}}(\boldsymbol{M}_2)^{-1}\rangle - \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M}_2)) \\ &= \langle \boldsymbol{M}_1 - \boldsymbol{M}_2, \boldsymbol{H}\rangle \end{aligned}$$

where the second equality follows from the first property of $P_{\mathcal{H}}(\boldsymbol{M}_1)$ and $P_{\mathcal{H}}(\boldsymbol{M}_2)$ from Proposition A.3. Since $\boldsymbol{M}_1 \succeq \boldsymbol{M}_2$, this implies that $\langle P_{\mathcal{H}}(\boldsymbol{M}_1)^2 - P_{\mathcal{H}}(\boldsymbol{M}_2)^2, \boldsymbol{H}\rangle \geq 0$ for all $\boldsymbol{H} \in \mathcal{H}$. By Lemma A.1, we know that $P_{\mathcal{H}}(\boldsymbol{M}_1)^2 - P_{\mathcal{H}}(\boldsymbol{M}_2)^2 \in \mathcal{H}$, so it further follows from Lemma A.2 that $P_{\mathcal{H}}(\boldsymbol{M}_1)^2 \succeq P_{\mathcal{H}}(\boldsymbol{M}_2)^2$. Since matrix square root is operator monotone, it holds that $P_{\mathcal{H}}(\boldsymbol{M}_1) \succeq P_{\mathcal{H}}(\boldsymbol{M}_2)$. We also know that $P_{\mathcal{H}}(\boldsymbol{M}_1) - P_{\mathcal{H}}(\boldsymbol{M}_2) \in \mathcal{K}$ because $\mathcal{K}$ is a subalgebra. Then we conclude that $P_{\mathcal{H}}(\boldsymbol{M}_1) - P_{\mathcal{H}}(\boldsymbol{M}_2) \in \mathcal{H}$ from the definition of $\mathcal{H}$. This completes the proof. □

We can further extend the definition of $P_{\mathcal{H}}(\cdot)$ to all positive semi-definite matrices. Specifically, for any $\boldsymbol{M} \succeq 0$, we have $\boldsymbol{M} + \epsilon\boldsymbol{I}_d \succ 0$ for any $\epsilon > 0$, so $P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)$ is well-defined. Also, by the operator monotonicity of $P_{\mathcal{H}}(\cdot)$ from

Proposition A.3, $P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d) \preceq P_{\mathcal{H}}(\boldsymbol{M}+\epsilon'\boldsymbol{I}_d)$ for $\epsilon' \geq \epsilon > 0$. This implies that $P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)$ has a limit as $\epsilon \to 0$. Therefore, for any $\boldsymbol{M} \succeq 0$, we define

$$P_{\mathcal{H}}(\boldsymbol{M}) = \lim_{\epsilon \searrow 0} P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d). \tag{14}$$

Note that the above equality is also true for $\boldsymbol{M} \succ 0$. To see this, we apply the optimality of every $P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)$ to get

$$\langle \boldsymbol{M}+\epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) > \langle \boldsymbol{M}+\epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)^{-1}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)).$$

Also note that $P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d) \succ P_{\mathcal{H}}(\boldsymbol{M}-\delta\boldsymbol{I}_d) \succ 0$ for sufficiently small $\delta > 0$ such that $\boldsymbol{M}-\delta\boldsymbol{I}_d \succ 0$. Therefore, letting $\epsilon \to 0$ on both sides of the previous inequality, we can exchange the order of taking the limit and taking the inverse of $P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)$ to obtain

$$\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{-1}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) \geq \Big\langle \boldsymbol{M}, \Big(\lim_{\epsilon \searrow 0} P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)\Big)^{-1}\Big\rangle + \mathrm{Tr}\Big(\lim_{\epsilon \searrow 0} P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)\Big).$$

Then by the optimality of $P_{\mathcal{H}}(\boldsymbol{M})$ and its uniqueness, we conclude that (14) is also valid for any $\boldsymbol{M} \succ 0$.

Indeed, the definition of $P_{\mathcal{H}}(\boldsymbol{M})$ in (14) provides a continuous extension of $P_{\mathcal{H}}$ to $\mathcal{S}_+^d$, as summarized in the following proposition.

Proposition A.4. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. As a function on $\mathcal{S}_{++}^d$, $P_{\mathcal{H}}$ can be continuously extended to be a function on $\mathcal{S}_+^d$. Moreover, for any $\boldsymbol{M} \succeq 0$ such that $\boldsymbol{M} \neq 0$, $P_{\mathcal{H}}(\boldsymbol{M}) \succeq 0$ satisfies the following properties:*

(a) $\mathrm{span}(\boldsymbol{M}) \subseteq \mathrm{span}(P_{\mathcal{H}}(\boldsymbol{M}))$ *and* $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) = \inf_{\boldsymbol{H}\in\mathcal{H}} \langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \mathrm{Tr}(\boldsymbol{H})$.

(b) $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\rangle = \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))$.

(c) $\langle \boldsymbol{M}, \overline{P_{\mathcal{H}}(\boldsymbol{M})}^{\dagger}\rangle = \inf_{\boldsymbol{H}\in\mathcal{H}, \mathrm{Tr}(\boldsymbol{H})\leq 1} \langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle$ *where we recall that* $\overline{P_{\mathcal{H}}(\boldsymbol{M})} = \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))^{-1} P_{\mathcal{H}}(\boldsymbol{M})$.

(d) For any $\boldsymbol{H} \in \mathcal{H}$, $\langle P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\boldsymbol{M}P_{\mathcal{H}}(\boldsymbol{M})^{\dagger} - \boldsymbol{\Pi}_{\boldsymbol{M}}, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M})\rangle = 0$, where $\boldsymbol{\Pi}_{\boldsymbol{M}}$ is the projection matrix onto $\mathrm{span}(P_{\mathcal{H}}(\boldsymbol{M}))$.

Moreover, for any $\boldsymbol{M}_1 \succeq \boldsymbol{M}_2 \succeq 0$, it holds that $P_{\mathcal{H}}(\boldsymbol{M}_1) - P_{\mathcal{H}}(\boldsymbol{M}_2) \in \mathcal{H}$, and in particular, $P_{\mathcal{H}}(\boldsymbol{M}_1) \succeq P_{\mathcal{H}}(\boldsymbol{M}_2)$.

Proof of Proposition A.4. We divide the proof into different parts for different properties of $P_{\mathcal{H}}$.

Proof for continuous extension of $P_{\mathcal{H}}$. We first show that $P_{\mathcal{H}}$ can be continuously extended to $\mathcal{S}_+^d$, and we consider the extension of $P_{\mathcal{H}}$ as given in (14). We first show that for any $\boldsymbol{M} \succeq 0$ and any sequence $\{\boldsymbol{M}_n\}_{n=1}^{\infty}$ such that each $\boldsymbol{M}_n \succ 0$ and $\lim_{n\to\infty} \boldsymbol{M}_n = \boldsymbol{M}$, it holds that $\lim_{n\to\infty} P_{\mathcal{H}}(\boldsymbol{M}_n) = P_{\mathcal{H}}(\boldsymbol{M})$. Note that for any $\delta \in (0,1)$, there exist $\bar{\epsilon}_n \geq \underline{\epsilon}_n > 0$ such that $0 \prec (1-\delta)(\boldsymbol{M}+\underline{\epsilon}_n\boldsymbol{I}_d) \preceq \boldsymbol{M}_n \preceq \boldsymbol{M}+\bar{\epsilon}_n\boldsymbol{I}_d$ for all large enough n and moreover, $\lim_{n\to\infty}\bar{\epsilon}_n = \lim_{n\to\infty}\underline{\epsilon}_n = 0$. Then by the operator monotonicity of $P_{\mathcal{H}}(\cdot)$, we have $P_{\mathcal{H}}((1-\delta)(\boldsymbol{M}+\underline{\epsilon}_n\boldsymbol{I}_d)) \preceq P_{\mathcal{H}}(\boldsymbol{M}_n) \preceq P_{\mathcal{H}}(\boldsymbol{M}+\bar{\epsilon}_n\boldsymbol{I}_d)$. Also note that $P_{\mathcal{H}}((1-\delta)(\boldsymbol{M}+\underline{\epsilon}_n\boldsymbol{I}_d)) = \sqrt{1-\delta}P_{\mathcal{H}}(\boldsymbol{M}+\underline{\epsilon}_n\boldsymbol{I}_d)$. Letting $n\to\infty$, both $P_{\mathcal{H}}(\boldsymbol{M}+\underline{\epsilon}_n\boldsymbol{I}_d)$ and $P_{\mathcal{H}}(\boldsymbol{M}+\bar{\epsilon}_n\boldsymbol{I}_d)$ converge to $P_{\mathcal{H}}(\boldsymbol{M})$, which then implies that $\sqrt{1-\delta}P_{\mathcal{H}}(\boldsymbol{M}) \preceq \lim_{n\to\infty} P_{\mathcal{H}}(\boldsymbol{M}_n) \preceq P_{\mathcal{H}}(\boldsymbol{M})$. Since δ is arbitrary, we conclude that $\lim_{n\to\infty} P_{\mathcal{H}}(\boldsymbol{M}_n) = P_{\mathcal{H}}(\boldsymbol{M})$. Next, consider any general sequence $\{\boldsymbol{M}_n\}_{n=1}^{\infty}$ such that $\boldsymbol{M}_n \succeq 0$ and $\lim_{n\to\infty}\boldsymbol{M}_n = \boldsymbol{M}$. For any $\epsilon > 0$, there exists $\delta > 0$ such that for all $\boldsymbol{M}' \succ 0$ with $\|\boldsymbol{M}-\boldsymbol{M}'\|_{\mathrm{F}} \leq \delta$, it holds that $\|P_{\mathcal{H}}(\boldsymbol{M}) - P_{\mathcal{H}}(\boldsymbol{M}')\|_{\mathrm{F}} \leq \epsilon$. For this δ, there exists $N > 0$ such that for all $n > N$, $\|\boldsymbol{M}-\boldsymbol{M}_n\|_{\mathrm{F}} \leq \delta/2$. Then since $P_{\mathcal{H}}(\boldsymbol{M}_n) = \lim_{m\to\infty} P_{\mathcal{H}}(\boldsymbol{M}_n + \frac{\delta}{2dm}\boldsymbol{I}_d)$, where for every $m \geq 1$ we have $\|\boldsymbol{M} - (\boldsymbol{M}_n + \frac{\delta}{2m\sqrt{d}}\boldsymbol{I}_d)\|_{\mathrm{F}} \leq \|\boldsymbol{M}-\boldsymbol{M}_n\|_{\mathrm{F}} + \|\frac{\delta}{2m\sqrt{d}}\boldsymbol{I}_d\|_{\mathrm{F}} \leq \delta$, so $\|P_{\mathcal{H}}(\boldsymbol{M}) - P_{\mathcal{H}}(\boldsymbol{M}_n + \frac{\delta}{2m\sqrt{d}}\boldsymbol{I}_d)\|_{\mathrm{F}} \leq \epsilon$ for all $m \geq 1$, which implies that $\|P_{\mathcal{H}}(\boldsymbol{M}) - P_{\mathcal{H}}(\boldsymbol{M}_n)\|_{\mathrm{F}} \leq \epsilon$ for all $n > N$. Therefore, it follows that $\lim_{n\to\infty} P_{\mathcal{H}}(\boldsymbol{M}_n) = P_{\mathcal{H}}(\boldsymbol{M})$. In conclusion, $P_{\mathcal{H}}(\cdot)$ can be extended to be a continuous function on $\mathcal{S}_+^d$.

Below, we fix any $\boldsymbol{M} \succeq 0$ such that $\boldsymbol{M} \neq 0$.

Proof for $\mathrm{span}(\boldsymbol{M}) \subseteq \mathrm{span}(P_{\mathcal{H}}(\boldsymbol{M}))$**.** Let $\boldsymbol{\Pi}_{\boldsymbol{M}} := P_{\mathcal{H}}(\boldsymbol{M})P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}$ be the projection matrix onto $\mathrm{span}(P_{\mathcal{H}}(\boldsymbol{M}))$, then $\boldsymbol{\Pi}_{\boldsymbol{M}} \in \mathcal{H}$ by Lemma A.1. It suffices to show that $\langle \boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}, \boldsymbol{M}\rangle = 0$. Using the fact that $\mathrm{Tr}(\boldsymbol{AB}) \le \mathrm{Tr}(\boldsymbol{A})\,\mathrm{Tr}(\boldsymbol{B})$ for any $\boldsymbol{A}, \boldsymbol{B} \succeq 0$, we can get the following inequality for any $\epsilon > 0$:

$$\begin{aligned}\langle \boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}, \boldsymbol{M}\rangle = \mathrm{Tr}((\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}})\boldsymbol{M}) &\le \mathrm{Tr}((\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}})P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)) \cdot \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\boldsymbol{M}) \\ &= \langle \boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)\rangle \cdot \langle \boldsymbol{M}, \boldsymbol{P}_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle \end{aligned} \tag{15}$$

By Proposition A.3, we know that $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)^{-1}\rangle = \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)) - \epsilon\langle \boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d)^{-1}\rangle \le \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}+\epsilon\boldsymbol{I}_d))$, which is bounded by an absolute constant for all $\epsilon \in (0,1)$. Therefore, letting $\epsilon \to 0$ on both sided of (15), since $\langle \boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)\rangle \to \langle \boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}, P_{\mathcal{H}}(\boldsymbol{M})\rangle = 0$ by the definition of $\boldsymbol{\Pi}_{\boldsymbol{M}}$, we conclude that $\langle \boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}, \boldsymbol{M}\rangle = 0$. This implies that $\mathrm{span}(\boldsymbol{M}) \subseteq \mathrm{span}(P_{\mathcal{H}}(\boldsymbol{M}))$.

Proof for the optimality of $P_{\mathcal{H}}(\boldsymbol{M})$**.** We still use $\boldsymbol{\Pi}_{\boldsymbol{M}}$ to denote the projection matrix onto $\mathrm{span}(P_{\mathcal{H}}(\boldsymbol{M}))$. For any $\epsilon > 0$, by the optimality of $P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)$, it holds for any $\boldsymbol{H} \in \mathcal{H}$ that

$$\langle \boldsymbol{M} + \epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)) \le \langle \boldsymbol{M} + \epsilon\boldsymbol{I}_d, \boldsymbol{H}^{-1}\rangle + \mathrm{Tr}(\boldsymbol{H}). \tag{16}$$

Since $\mathrm{span}(\boldsymbol{M}) \subseteq \mathrm{span}(P_{\mathcal{H}}(\boldsymbol{M}))$, we know that $\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle = \langle \boldsymbol{M}, \boldsymbol{\Pi}_{\boldsymbol{M}}P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\boldsymbol{\Pi}_{\boldsymbol{M}}\rangle$. Then since $\lim_{\epsilon\to 0} P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d) = P_{\mathcal{H}}(\boldsymbol{M})$, it holds that $\boldsymbol{\Pi}_{\boldsymbol{M}}P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\boldsymbol{\Pi}_{\boldsymbol{M}} = (\boldsymbol{\Pi}_{\boldsymbol{M}}P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)\boldsymbol{\Pi}_{\boldsymbol{M}})^{\dagger}$ for sufficiently small $\epsilon > 0$, and also that $P_{\mathcal{H}}(\boldsymbol{M})^{\dagger} = \lim_{\epsilon\to 0}(\boldsymbol{\Pi}_{\boldsymbol{M}}P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)\boldsymbol{\Pi}_{\boldsymbol{M}})^{\dagger}$. This implies

$$\lim_{\epsilon\searrow 0}\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle = \langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\rangle. \tag{17}$$

Therefore, assuming the existence of the limit of $\langle \epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle$ as $\epsilon \to 0$, taking the limit of $\epsilon \to 0$ on both sides of (16) yields

$$\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) + \lim_{\epsilon\to 0}\langle \epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle \le \langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \mathrm{Tr}(\boldsymbol{H}). \tag{18}$$

Hence, it suffices to show that $\langle \epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle \to 0$ as $\epsilon \to 0$. To prove this, note that for any $\epsilon > 0$, $P_{\mathcal{H}}(\boldsymbol{M}) + \sqrt{\epsilon}(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}) \succ 0$, and its inverse is given by $P_{\mathcal{H}}(\boldsymbol{M})^{\dagger} + \epsilon^{-1/2}(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}})$. Also, $P_{\mathcal{H}}(\boldsymbol{M}) + \sqrt{\epsilon}(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}) \in \mathcal{H}$ by Lemma A.1. Therefore, by the optimality of $P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)$, we have

$$\begin{aligned}&\langle \boldsymbol{M} + \epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)) \\ &\quad\le \langle \boldsymbol{M} + \epsilon\boldsymbol{I}_d, (P_{\mathcal{H}}(\boldsymbol{M}) + \sqrt{\epsilon}(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}))^{-1}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}) + \sqrt{\epsilon}(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}})) \\ &\quad= \langle \boldsymbol{M} + \epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger} + \epsilon^{-1/2}(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}})\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) + \sqrt{\epsilon}(d - \mathrm{rank}(P_{\mathcal{H}}(\boldsymbol{M}))) \\ &\quad= \langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\rangle + \epsilon \cdot \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}) + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) + 2\sqrt{\epsilon}(d - \mathrm{rank}(P_{\mathcal{H}}(\boldsymbol{M})))\end{aligned}$$

where we apply $\mathrm{Tr}(\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}) = d - \mathrm{rank}(P_{\mathcal{H}}(\boldsymbol{M}))$ and the last equality follows from the fact that $\langle \boldsymbol{M}, \boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}\rangle = 0$. Rearranging the above inequality, we obtain

$$\begin{aligned}\epsilon\langle \boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle \le \langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\rangle - \langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) - \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)) \\ + \epsilon \cdot \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}) + 2\sqrt{\epsilon}(d - \mathrm{rank}(P_{\mathcal{H}}(\boldsymbol{M}))).\end{aligned}$$

Now applying (17) and noting that $\epsilon\langle \boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle \ge 0$, taking the limit $\epsilon \to 0$, we obtain

$$\lim_{\epsilon\to 0}\langle \epsilon\boldsymbol{I}_d, P_{\mathcal{H}}(\boldsymbol{M} + \epsilon\boldsymbol{I}_d)^{-1}\rangle = 0. \tag{19}$$

Then combining (18) and (19), we conclude that for any $\boldsymbol{H} \in \mathcal{H}$,

$$\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^{\dagger}\rangle + \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) \le \langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \mathrm{Tr}(\boldsymbol{H}).$$

This confirms the optimality of $P_{\mathcal{H}}(\boldsymbol{M})$. Further applying the same argument as in the proof of Proposition A.3 yields the optimality of $\overline{P_{\mathcal{H}}(\boldsymbol{M})} = \mathrm{Tr}(P_{\mathcal{H}}(\boldsymbol{M}))^{-1}P_{\mathcal{H}}(\boldsymbol{M})$.

Proof for property (d) of $P_{\mathcal{H}}(\boldsymbol{M})$. By the optimality of $P_{\mathcal{H}}(\boldsymbol{M})$, we have

$$\begin{aligned}\langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^\dagger\rangle + \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) &= \inf_{\boldsymbol{H}\in\mathcal{H}} \underbrace{\langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \operatorname{Tr}(\boldsymbol{H})}_{f_{\boldsymbol{M}}(\boldsymbol{H})} \\ &= \inf_{\boldsymbol{H}\in\mathcal{H}} \underbrace{\langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + \operatorname{Tr}(\boldsymbol{\Pi}_{\boldsymbol{M}}\boldsymbol{H}\boldsymbol{\Pi}_{\boldsymbol{M}})}_{\tilde{f}_{\boldsymbol{M}}(\boldsymbol{H})}. \end{aligned} \tag{20}$$

To see why the last equality holds, first note that $\inf_{\boldsymbol{H}\in\mathcal{H}} f_{\boldsymbol{M}}(\boldsymbol{H}) \geq \inf_{\boldsymbol{H}\in\mathcal{H}} \tilde{f}_{\boldsymbol{M}}(\boldsymbol{H})$ because $\operatorname{Tr}(\boldsymbol{H}) \geq \operatorname{Tr}(\boldsymbol{\Pi}_{\boldsymbol{M}}\boldsymbol{H}\boldsymbol{\Pi}_{\boldsymbol{M}})$. Then for the other direction, note that we can always approximate $\tilde{f}_{\boldsymbol{M}}(\boldsymbol{H})$ using $f_{\boldsymbol{M}}(\boldsymbol{H}_\delta)$ where $\boldsymbol{H}_\delta = (1-\delta)\boldsymbol{\Pi}_{\boldsymbol{M}}\boldsymbol{H}\boldsymbol{\Pi}_{\boldsymbol{M}}^\top + \delta\boldsymbol{I}_d$, for which $f_{\boldsymbol{M}}(\boldsymbol{H}_\delta) \leq \frac{1}{1-\delta}\langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle + (1-\delta)\operatorname{Tr}(\boldsymbol{\Pi}_{\boldsymbol{M}}\boldsymbol{H}\boldsymbol{\Pi}_{\boldsymbol{M}}) + \delta d \to \tilde{f}_{\boldsymbol{M}}(\boldsymbol{H})$ as $\delta \to 0$. This implies that $\inf_{\boldsymbol{H}\in\mathcal{H}} \tilde{f}_{\boldsymbol{M}}(\boldsymbol{H}) \geq \inf_{\boldsymbol{H}\in\mathcal{H}} f_{\boldsymbol{M}}(\boldsymbol{H})$, and thus the optimal objective values of the two optimization problems are the same. Now define $\widetilde{\boldsymbol{H}}_{\boldsymbol{M}} = P_{\mathcal{H}}(\boldsymbol{M}) + (\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}}) \in \mathcal{H}$ whose inverse is $\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}^{-1} = P_{\mathcal{H}}(\boldsymbol{M})^\dagger + (\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}})$, then $\widetilde{\boldsymbol{H}}_{\boldsymbol{M}} \succ 0$ is a solution to the optimization problem in (20) because $\tilde{f}_{\boldsymbol{M}}(\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}) = \langle \boldsymbol{M}, P_{\mathcal{H}}(\boldsymbol{M})^\dagger\rangle + \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{M})) = \inf_{\boldsymbol{H}\in\mathcal{H}} f_{\boldsymbol{M}}(\boldsymbol{H})$. The gradient of $\tilde{f}_{\boldsymbol{M}}$ is given by $\nabla \tilde{f}_{\boldsymbol{M}}(\boldsymbol{H}) = -\boldsymbol{H}^{-1}\boldsymbol{M}\boldsymbol{H}^{-1} + \boldsymbol{\Pi}_{\boldsymbol{M}}$. Since $\mathcal{H}$ is a cone, the optimality of $\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}$ implies that for any $\boldsymbol{H} \in \mathcal{H}$,

$$0 = \langle \nabla f(\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}), \boldsymbol{H} - \widetilde{\boldsymbol{H}}_{\boldsymbol{M}}\rangle = \langle -\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}^{-1}\boldsymbol{M}\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}^{-1} + \boldsymbol{\Pi}_{\boldsymbol{M}}, \boldsymbol{H} - \widetilde{\boldsymbol{H}}_{\boldsymbol{M}}\rangle.$$

By the definition of $\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}$, we have $\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}^{-1}\boldsymbol{M}\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}^{-1} = P_{\mathcal{H}}(\boldsymbol{M})^\dagger\boldsymbol{M}P_{\mathcal{H}}(\boldsymbol{M})^\dagger$ because $\operatorname{span}(\boldsymbol{M}) \subseteq \operatorname{span}(P_{\mathcal{H}}(\boldsymbol{M}))$. Therefore, it follows that

$$\begin{aligned} 0 &= \langle P_{\mathcal{H}}(\boldsymbol{M})^\dagger\boldsymbol{M}P_{\mathcal{H}}(\boldsymbol{M})^\dagger - \boldsymbol{\Pi}_{\boldsymbol{M}}, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M}) - (\boldsymbol{I}_d - \boldsymbol{\Pi}_{\boldsymbol{M}})\rangle \\ &= \langle P_{\mathcal{H}}(\boldsymbol{M})^\dagger\boldsymbol{M}P_{\mathcal{H}}(\boldsymbol{M})^\dagger - \boldsymbol{\Pi}_{\boldsymbol{M}}, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{M})\rangle. \end{aligned}$$

Proof for the operator monotonicity of the extended $P_{\mathcal{H}}$. For any $\boldsymbol{M}_1 \succeq \boldsymbol{M}_2 \succeq 0$, it follows from Proposition A.3 that for any $\epsilon > 0$, $P_{\mathcal{H}}(\boldsymbol{M}_1 + \epsilon\boldsymbol{I}_d) \succeq P_{\mathcal{H}}(\boldsymbol{M}_2 + \epsilon\boldsymbol{I}_d)$. Taking the limit as $\epsilon \to 0$ on both sides yields $P_{\mathcal{H}}(\boldsymbol{M}_1) \succeq P_{\mathcal{H}}(\boldsymbol{M}_2)$. Similarly, as $P_{\mathcal{H}}(\boldsymbol{M}_1 + \epsilon\boldsymbol{I}_d) - P_{\mathcal{H}}(\boldsymbol{M}_2 + \epsilon\boldsymbol{I}_d) \in \mathcal{H}$ for every $\epsilon > 0$ by Proposition A.3, we have $P_{\mathcal{H}}(\boldsymbol{M}_1) - P_{\mathcal{H}}(\boldsymbol{M}_2) = \lim_{\epsilon\to 0} P_{\mathcal{H}}(\boldsymbol{M}_1 + \epsilon\boldsymbol{I}_d) - P_{\mathcal{H}}(\boldsymbol{M}_2 + \epsilon) \in \mathcal{H}$ because $\mathcal{H}$ is a closed set. This completes the proof. □

A.3. Adaptive gradient norm corresponds to the dual norm

Proof of Lemma 3.3. We begin with the case of $t = 1$. Fix any $\boldsymbol{g} \in \mathbb{R}^d$. First for any $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\|_{\mathcal{H}} \leq 1$, by Cauchy-Schwarz inequality, it holds for all $\boldsymbol{H} \in \mathcal{H}$ with $\operatorname{Tr}(\boldsymbol{H}) \leq 1$ that

$$\boldsymbol{g}^\top\boldsymbol{w} \leq \sqrt{\boldsymbol{g}^\top\boldsymbol{H}^{-1}\boldsymbol{g}}\sqrt{\boldsymbol{w}^\top\boldsymbol{H}\boldsymbol{w}} \leq \sqrt{\langle \boldsymbol{g}\boldsymbol{g}^\top, \boldsymbol{H}^{-1}\rangle}\cdot\|\boldsymbol{w}\|_{\mathcal{H}} \leq \sqrt{\langle \boldsymbol{g}\boldsymbol{g}^\top, \boldsymbol{H}^{-1}\rangle}$$

where the second inequality follows from the definition of $\|\boldsymbol{w}\|_{\mathcal{H}}$. Now, further taking infimum over $\boldsymbol{H} \in \mathcal{H}$ with $\operatorname{Tr}(\boldsymbol{H}) \leq 1$ and then taking supremum over $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\|_{\mathcal{H}} \leq 1$, we obtain that

$$\|\boldsymbol{g}\|_{\mathcal{H}}^* \leq \sqrt{\inf_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\leq 1} \langle \boldsymbol{g}\boldsymbol{g}^\top, \boldsymbol{H}^{-1}\rangle}. \tag{21}$$

Next, we need to show that the above inequality holds also for the other direction.

By Proposition A.4, we define $\boldsymbol{H}_* = \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top))^{-1}P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top) = \inf_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\leq 1}\langle \boldsymbol{M}, \boldsymbol{H}^{-1}\rangle$. We choose correspondingly $\boldsymbol{w}_* = \frac{\boldsymbol{H}_*^\dagger\boldsymbol{g}}{\|\boldsymbol{H}_*^\dagger\boldsymbol{g}\|_{\mathcal{H}}}$, which satisfies that $\|\boldsymbol{w}_*\|_{\mathcal{H}} = 1$. Then

$$\|\boldsymbol{g}\|_{\mathcal{H}}^* \geq \boldsymbol{g}^\top\boldsymbol{w}_* = \frac{\boldsymbol{g}^\top\boldsymbol{H}_*^\dagger\boldsymbol{g}}{\|\boldsymbol{H}_*^\dagger\boldsymbol{g}\|_{\mathcal{H}}}. \tag{22}$$

Therefore, it suffices to show that $\boldsymbol{g}^\top\boldsymbol{w}_* \geq \sqrt{\boldsymbol{g}^\top\boldsymbol{H}_*^\dagger\boldsymbol{g}}$, which is equivalent to

$$\boldsymbol{g}^\top\boldsymbol{H}_*^\dagger\boldsymbol{g} \geq \|\boldsymbol{H}_*^\dagger\boldsymbol{g}\|_{\mathcal{H}}^2 = \sup_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\leq 1} \boldsymbol{g}^\top\boldsymbol{H}_*^\dagger\boldsymbol{H}\boldsymbol{H}_*^\dagger\boldsymbol{g}.$$

By the property (d) of $P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)$ from Proposition A.4, we have $\langle P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)^\dagger \boldsymbol{g}\boldsymbol{g}^\top P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)^\dagger - \boldsymbol{\Pi}_{\boldsymbol{g}\boldsymbol{g}^\top}, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)\rangle = 0$ for any $\boldsymbol{H} \in \mathcal{H}$. Rearranging this equality, we obtain

$$\begin{aligned}\boldsymbol{g}^\top P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)^\dagger \boldsymbol{H} P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)^\dagger \boldsymbol{g} &= \boldsymbol{g}^\top P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)^\dagger \boldsymbol{g} + \langle \boldsymbol{\Pi}_{\boldsymbol{g}\boldsymbol{g}^\top}, \boldsymbol{H} - P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)\rangle \\ &= \boldsymbol{g}^\top P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)^\dagger \boldsymbol{g} + \langle \boldsymbol{\Pi}_{\boldsymbol{g}\boldsymbol{g}^\top}, \boldsymbol{H}\rangle - \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top))\end{aligned}$$

where the second equality is because $\boldsymbol{\Pi}_{\boldsymbol{g}\boldsymbol{g}^\top}$ is exactly the projection matrix onto $\operatorname{span}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top))$. Applying the above equality with $\operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)) \cdot \overline{\boldsymbol{H}}$ in place of $\boldsymbol{H}$ and plugging in the definition of $\boldsymbol{H}_*$, we further have

$$\begin{aligned}\boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \overline{\boldsymbol{H}} \boldsymbol{H}_*^\dagger \boldsymbol{g} &= \boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{g} + \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top))\langle \boldsymbol{\Pi}_{\boldsymbol{g}\boldsymbol{g}^\top}, \overline{\boldsymbol{H}}\rangle - \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top)) \\ &\le \boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{g} + \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top))^2 \operatorname{Tr}(\overline{\boldsymbol{H}}) - \operatorname{Tr}(P_{\mathcal{H}}(\boldsymbol{g}\boldsymbol{g}^\top))^2 \\ &= \boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{g}.\end{aligned}$$

This implies that for any $\boldsymbol{H} \in \mathcal{H}$ with $\operatorname{Tr}(\boldsymbol{H}) \le 1$, we have $\boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{H} \boldsymbol{H}_*^\dagger \boldsymbol{g} \le \boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{g}$. Therefore, it follows from (22) that

$$\begin{aligned}\|\boldsymbol{g}\|_{\mathcal{H}}^* &\ge \sqrt{\boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{g}} \sqrt{\frac{\boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{g}}{\inf_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\le 1} \boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{H} \boldsymbol{H}_*^\dagger \boldsymbol{g}}} \\ &\ge \sqrt{\boldsymbol{g}^\top \boldsymbol{H}_*^\dagger \boldsymbol{g}} = \sqrt{\inf_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\le 1} \langle \boldsymbol{g}\boldsymbol{g}^\top, \boldsymbol{H}^{-1}\rangle}\end{aligned} \tag{23}$$

where the equality follows from the definition of $\boldsymbol{H}_*$.

Finally, combining (21) and (23), we conclude that

$$\|\boldsymbol{g}\|_{\mathcal{H}}^* = \inf_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\le 1} \sqrt{\langle \boldsymbol{g}\boldsymbol{g}^\top, \boldsymbol{H}^{-1}\rangle}. \tag{24}$$

For the general case of $\boldsymbol{g}_{1:t} = (\boldsymbol{g}_1, \dots, \boldsymbol{g}_t) \in \mathbb{R}^{d\times t}$, note that for any $\boldsymbol{H} \in \mathcal{H}$

$$\left\langle \sum_{s=1}^t \boldsymbol{g}_s \boldsymbol{g}_s^\top, \boldsymbol{H}^{-1} \right\rangle = \langle \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})\, \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})^\top, \boldsymbol{H}^{-1} \otimes \boldsymbol{I}_t\rangle = \langle \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})\, \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})^\top, (\boldsymbol{H} \otimes \boldsymbol{I}_t)^{-1}\rangle.$$

Then applying (24) with $\overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})$ in place of $\boldsymbol{g}$ and $\mathcal{H} \otimes \boldsymbol{I}_t$ in place of $\mathcal{H}$, we obtain

$$\begin{aligned}\|\overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})\|_{\mathcal{H}\otimes\boldsymbol{I}_t}^* &= \inf_{\boldsymbol{H}\otimes\boldsymbol{I}_t\in\mathcal{H}\otimes\boldsymbol{I}_t, \operatorname{Tr}(\boldsymbol{H}\otimes\boldsymbol{I}_t)\le 1} \sqrt{\langle \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})\, \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})^\top, (\boldsymbol{H} \otimes \boldsymbol{I}_t)^{-1}\rangle} \\ &= \inf_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\le 1} \sqrt{t\langle \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})\, \overline{\operatorname{vec}}(\boldsymbol{g}_{1:t})^\top, (\boldsymbol{H} \otimes \boldsymbol{I}_t)^{-1}\rangle} \\ &= \sqrt{t} \cdot \inf_{\boldsymbol{H}\in\mathcal{H}, \operatorname{Tr}(\boldsymbol{H})\le 1} \sqrt{\left\langle \sum_{s=1}^t \boldsymbol{g}_s \boldsymbol{g}_s^\top, \boldsymbol{H}^{-1} \right\rangle}\end{aligned}$$

where the second equality is because $\operatorname{Tr}(\boldsymbol{H} \otimes \boldsymbol{I}_t) = \operatorname{Tr}(\boldsymbol{H}) \operatorname{Tr}(\boldsymbol{I}_t) = t \cdot \operatorname{Tr}(\boldsymbol{H})$. This completes the proof. □

A.4. Examples of ill-structured preconditioner sets

A.4.1. Preconditioner set of two-sided Shampoo

Consider $\mathcal{H} = \{\boldsymbol{H} \in \mathcal{S}_+^d : \boldsymbol{H} = \boldsymbol{U} \otimes \boldsymbol{V} \text{ for some } \boldsymbol{U} \in \mathbb{R}^{d_L\times d_L}, \boldsymbol{V} \in \mathbb{R}^{d_R\times d_R}\}$. In particular, we consider the special case of $d_L = d_R = 2$, and the following two matrices

$$\boldsymbol{G}_1(\epsilon) = \operatorname{diag}(1, \epsilon, \epsilon, \epsilon), \qquad \boldsymbol{G}_2(\epsilon) = \operatorname{diag}(1, \epsilon, \epsilon, 1)$$

where $\epsilon > 0$ is a small constant to be determined later. Clearly $\boldsymbol{G}_1(\epsilon) \preceq \boldsymbol{G}_2(\epsilon)$ when $\epsilon \leq 1$, but we will show that $\boldsymbol{H}_{\boldsymbol{G}_1(\epsilon)} \preceq \boldsymbol{H}_{\boldsymbol{G}_2(\epsilon)}$ does not hold for sufficiently small $\epsilon > 0$. For any $\boldsymbol{H} = \boldsymbol{U} \otimes \boldsymbol{V} \in \mathcal{H}$, it holds that both $\boldsymbol{U}$ and $\boldsymbol{V}$ are PSD, and we explicitly parametrize them as

$$\boldsymbol{U} = \begin{pmatrix} u_1 & u_2 \\ u_2 & u_3 \end{pmatrix}, \qquad \boldsymbol{V} = \begin{pmatrix} v_1 & v_2 \\ v_2 & v_3 \end{pmatrix}$$

where $u_1, u_3 \geq 0$, $u_1 u_3 \geq u_2^2$, and similarly, $v_1, v_3 \geq 0$, $v_1 v_3 \geq v_3^2$. Correspondingly,

$$\boldsymbol{H} = \boldsymbol{U} \otimes \boldsymbol{V} = \begin{pmatrix} u_1v_1 & u_1v_2 & u_2v_1 & u_2v_2 \\ u_1v_2 & u_1v_3 & u_2v_2 & u_2v_3 \\ u_2v_1 & u_2v_2 & u_3v_1 & u_3v_2 \\ u_2v_2 & u_3v_2 & u_3v_2 & u_3v_3 \end{pmatrix}.$$

We first analyze $\boldsymbol{H}_{\boldsymbol{G}_2(\epsilon)}$. For convenience, we consider

$$\begin{aligned} \boldsymbol{H}^{-1}_{\boldsymbol{G}_2(\epsilon)} &= \arg\min_{\boldsymbol{H} \in \mathcal{H}} \langle \boldsymbol{G}_2(\epsilon), \boldsymbol{H} \rangle + \operatorname{Tr}(\boldsymbol{H}^{-1}) \\ &= \arg\min_{\boldsymbol{U}, \boldsymbol{V} \in \mathcal{S}_+^2} \underbrace{\langle \boldsymbol{G}_2(\epsilon), \boldsymbol{U} \otimes \boldsymbol{V} \rangle + \operatorname{Tr}((\boldsymbol{U} \otimes \boldsymbol{V})^{-1})}_{=: f_{\boldsymbol{G}_2(\epsilon)}(\boldsymbol{U}, \boldsymbol{V})} \end{aligned}$$

Recall the properties of the Kronecker product that $(\boldsymbol{U} \otimes \boldsymbol{V})^{-1} = \boldsymbol{U}^{-1} \otimes \boldsymbol{V}^{-1}$ and that $\operatorname{Tr}(\boldsymbol{U}^{-1} \otimes \boldsymbol{V}^{-1}) = \operatorname{Tr}(\boldsymbol{U}^{-1}) \cdot \operatorname{Tr}(\boldsymbol{V}^{-1})$. Then further using the definition of $\boldsymbol{G}_2(\epsilon)$, we have

$$\begin{aligned} f_{\boldsymbol{G}_2(\epsilon)}(\boldsymbol{U}, \boldsymbol{V}) &= u_1v_1 + u_3v_3 + \operatorname{Tr}(\boldsymbol{U}^{-1}) \cdot \operatorname{Tr}(\boldsymbol{V}^{-1}) + \epsilon(u_1v_3 + u_3v_1) \\ &= u_1v_1 + u_3v_3 + \frac{u_1 + u_3}{u_1u_3 - u_2^2} \cdot \frac{v_1 + v_3}{v_1v_3 - v_2^2} + \epsilon(u_1v_3 + u_3v_1) \end{aligned}$$

where we apply the explicit expression of $\boldsymbol{U}^{-1}$ and $\boldsymbol{V}^{-1}$. First observe that to minimize $f_{\boldsymbol{G}_2(\epsilon)}(\boldsymbol{U}, \boldsymbol{V})$, we must have $u_2 = v_2 = 0$ because otherwise $\frac{u_1+u_3}{u_1u_3-u_2^2} > \frac{u_1+u_3}{u_1u_3}$ and similarly $\frac{v_1+v_3}{v_1v_3-v_2^2} > \frac{v_1+v_3}{v_1v_3}$. Therefore, it suffices to further minimize the following:

$$f_{\boldsymbol{G}_2(\epsilon)}(\boldsymbol{U}, \boldsymbol{V}) = \underbrace{u_1v_1 + u_3v_3 + \frac{u_1 + u_3}{u_1u_3} \cdot \frac{v_1 + v_3}{v_1v_3}}_{\tilde{f}_{\boldsymbol{G}_2}(\boldsymbol{U}, \boldsymbol{V})} + \epsilon(u_1v_3 + u_3v_1).$$

For the first term $\tilde{f}_{\boldsymbol{G}_2}(\boldsymbol{U}, \boldsymbol{V})$, we can apply the AM-GM inequality to get $\tilde{f}_{\boldsymbol{G}_2}(\boldsymbol{U}, \boldsymbol{V}) \geq (4\sqrt{u_1u_3v_1v_3})^{1/3}$, where the equality is achieved when $u_1 = u_3$, $v_1 = v_3$, and $u_1v_1 = \sqrt{2}$. Moreover, for sufficiently small $\epsilon > 0$, the contribution of the second term $\epsilon(u_1v_3 + u_3v_1)$ is negligible, and thus we conclude that $\lim_{\epsilon \to 0} \boldsymbol{H}_{\boldsymbol{G}_2(\epsilon)} = \frac{\sqrt{2}}{2}\boldsymbol{I}_4$.

Similarly, for $\boldsymbol{G}_1(\epsilon)$, we have

$$\begin{aligned} f_{\boldsymbol{G}_1(\epsilon)}(\boldsymbol{U}, \boldsymbol{V}) &= u_1v_1 + \frac{u_1 + u_3}{u_1u_3 - u_2^2} \cdot \frac{v_1 + v_3}{v_1v_3 - v_2^2} + \epsilon(u_1v_3 + u_3v_1 + u_3v_3) \\ &\geq u_1v_1 + \frac{u_1 + u_3}{u_1u_3} \cdot \frac{v_1 + v_3}{v_1v_3} + \epsilon(u_1v_3 + u_3v_1 + u_3v_3) \\ &= \underbrace{u_1v_1 + \frac{1}{u_1v_1}}_{\tilde{f}_{\boldsymbol{G}_1}(\boldsymbol{U}, \boldsymbol{V})} + \frac{1}{u_1v_3} + \frac{1}{u_3v_1} + \frac{1}{u_3v_3} + \epsilon(u_1v_3 + u_3v_1 + u_3v_3) \end{aligned}$$

where the equality holds when $u_2 = v_2 = 0$. The first term $\tilde{f}_{\boldsymbol{G}_1}(\boldsymbol{U}, \boldsymbol{V})$ attains the minimum value when $u_1v_1 = 1$, and the remainder can be made small by choosing large u_3, v_3 when ϵ is sufficiently small. Therefore, we conclude that $\lim_{\epsilon \to 0} \boldsymbol{H}_{\boldsymbol{G}_1(\epsilon)} = \operatorname{diag}(1, 0, 0, 0)$.

Now comparing the limits of $\boldsymbol{H}_{\boldsymbol{G}_1}(\epsilon)$ and $\boldsymbol{H}_{\boldsymbol{G}_2}(\epsilon)$ as $\epsilon \to 0$, we can see that for sufficiently small $\epsilon > 0$, it holds that $\boldsymbol{H}_{\boldsymbol{G}_1(\epsilon)}[1, 1] > \boldsymbol{H}_{\boldsymbol{G}_2(\epsilon)}[1, 1]$. Hence, for sufficiently small $\epsilon > 0$, $\boldsymbol{H}_{\boldsymbol{G}_1}(\epsilon) \preceq \boldsymbol{H}_{\boldsymbol{G}_2(\epsilon)}$ does not hold.

A.4.2. TRIDIAGONAL MATRICES

Consider $\mathcal{H} = \{\boldsymbol{H} \in \mathcal{S}_+^d : \boldsymbol{H} \text{ is tridiagonal}\}$. Here, for $\boldsymbol{H}$ to be tridiagonal, it has nonzero elements on the main diagonal, the first diagonal above the main diagonal, and the first diagonal below the main diagonal. We consider the specific example of 3-by-3 tridiagonal matrices, and examine $P_{\mathcal{H}}(\cdot)$ for the following two matrices

$$\boldsymbol{M} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \quad \boldsymbol{M}' = \begin{pmatrix} 10000 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

It is clear that $0 \prec \boldsymbol{M} \preceq \boldsymbol{M}'$. We numerically solve for $P_{\mathcal{H}}(\boldsymbol{M})$ and $P_{\mathcal{H}}(\boldsymbol{M}')$ to get

$$P_{\mathcal{H}}(\boldsymbol{M}) \approx \begin{pmatrix} 1.382548 & 0.297594 & 0 \\ 0.297594 & 1.318491 & 0.297594 \\ 0 & 0.297594 & 1.382548 \end{pmatrix}, \qquad P_{\mathcal{H}}(\boldsymbol{M}') \approx \begin{pmatrix} 100.000004 & 0.007229 & 0 \\ 0.007229 & 1.365999 & 0.366002 \\ 0 & 0.366002 & 1.366032 \end{pmatrix}.$$

Note that $P_{\mathcal{H}}(\boldsymbol{M}) \preceq P_{\mathcal{H}}(\boldsymbol{M}')$ does not hold because the last diagonal entry of $P_{\mathcal{H}}(\boldsymbol{M})$ is larger than that of $P_{\mathcal{H}}(\boldsymbol{M}')$.

B. Calculations for Examples of Well-Structured Preconditioner Sets

As mentioned in Section 3.3, we provide calculations for how to derive each specific algorithm from Algorithm 1 with specific choice of $\mathcal{H}$ and explain each entry in Table 1. Recall from Algorithm 1 that

$$\boldsymbol{M}_t = \epsilon \boldsymbol{I}_d + \sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\top.$$

B.1. AdaGrad-Norm

For AdaGrad-Norm, we have $\mathcal{H} = \{c \cdot \boldsymbol{I}_d \mid c \geq 0\}$.

Calculation for $\boldsymbol{H}_t$. Then for any $\boldsymbol{H} = c \cdot \boldsymbol{I}_d \in \mathcal{H}$,

$$\left\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1} \right\rangle + \eta^2 \operatorname{Tr}(\boldsymbol{H}) = \frac{1}{c} \operatorname{Tr}(\boldsymbol{M}_t) + \eta^2 c d \geq 2\eta \sqrt{d \cdot \operatorname{Tr}(\boldsymbol{M}_t)}$$

where the equality is achieved by choosing

$$c = \frac{1}{\eta} \sqrt{\frac{1}{d} \operatorname{Tr}(\boldsymbol{M}_t)} = \frac{1}{\eta} \sqrt{\epsilon + \frac{1}{d} \sum_{s=1}^{t} \|\boldsymbol{g}_s\|_2^2}.$$

This corresponds exactly to the update rule of AdaGrad-Norm, which adjusts the global learning rate based on the accumulated ℓ_2 norm of past gradients.

Calculation for $\|\cdot\|_{\mathcal{H}}$. For the associated $\|\cdot\|_{\mathcal{H}}$, we have

$$\|\boldsymbol{x}\|_{\mathcal{H}} = \sup_{0 \leq c \leq 1/d} \sqrt{c \cdot \boldsymbol{x}^\top \boldsymbol{I}_d \boldsymbol{x}} = \frac{\|\boldsymbol{x}\|_2}{\sqrt{d}}.$$

Calculation for $\|\|\boldsymbol{g}_{1:t}\|\|_{\mathcal{H}}$. We can calculate the adaptive gradient norm similarly:

$$\|\|\boldsymbol{g}_{1:t}\|\|_{\mathcal{H}} = \inf_{0 \leq c \leq 1/d} \sqrt{\left\langle \sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\top, (c\boldsymbol{I}_d)^{-1} \right\rangle} = \sqrt{d} \sqrt{\operatorname{Tr}\left(\sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\top \right)} = \sqrt{d} \sqrt{\sum_{s=1}^{t} \|\boldsymbol{g}_s\|_2^2}.$$

B.2. Diagonal AdaGrad

For diagonal AdaGrad, $\mathcal{H} = \mathcal{D}_+^d = \{\operatorname{diag}(c_1, \ldots, c_d) \mid c_1, \ldots, c_d \geq 0\}$.

Calculation for $\boldsymbol{H}_t$. For any $\boldsymbol{H} = \operatorname{diag}(c_1, \ldots, c_d) \in \mathcal{H}$, we have

$$\left\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1} \right\rangle + \eta^2 \operatorname{Tr}(\boldsymbol{H}) = \sum_{i=1}^{d} \left(\frac{\boldsymbol{M}_{t,ii}}{c_i} + \eta^2 c_i \right) \geq \sum_{i=1}^{d} 2\eta \sqrt{\boldsymbol{M}_{t,ii}}$$

where the equality is achieved by choosing

$$c_i = \frac{1}{\eta} \sqrt{\boldsymbol{M}_{t,ii}} = \frac{1}{\eta} \sqrt{\epsilon + \sum_{s=1}^{t} g_{s,i}^2}, \quad \text{for } i = 1, \ldots, d.$$

This corresponds to the update rule diagonal AdaGrad, which computes the historical sum of squared gradients for each individual coordinate.

Calculation for $\|\cdot\|_{\mathcal{H}}$. For the associated $\|\cdot\|_{\mathcal{H}}$, we have

$$\|\boldsymbol{x}\|_{\mathcal{H}} = \sup_{\sum_{i=1}^{d} c_i \leq 1} \sqrt{\boldsymbol{x}^\top \operatorname{diag}(c_1, \ldots, c_d) \boldsymbol{x}} = \sup_{\sum_{i=1}^{d} c_i \leq 1} \sqrt{\sum_{i=1}^{d} c_i x_i^2} = \max_{i \in [d]} |x_i| = \|\boldsymbol{x}\|_\infty$$

Calculation for $\|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}}$. We can calculate the adaptive gradient norm similarly:

$$\begin{aligned}
\|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}} &= \inf_{\sum_{i=1}^{d} c_i \leq 1} \sqrt{\left\langle \sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\top, \operatorname{diag}(c_1 \ldots, c_d)^{-1} \right\rangle} \\
&= \inf_{\sum_{i=1}^{d} c_i \leq 1} \sqrt{\sum_{i=1}^{d} \frac{1}{c_i} \sum_{s=1}^{t} g_{s,i}^2} \\
&= \sum_{i=1}^{d} \sqrt{\sum_{s=1}^{t} g_{s,i}^2}
\end{aligned}$$

where the last equality is because of the Cauchy inequality: for $c_1, \ldots, c_d \geq 0$ such that $\sum_{i=1}^{d} c_i \leq 1$,

$$\sqrt{\sum_{i=1}^{d} \frac{1}{c_i} \sum_{s=1}^{t} g_{s,i}^2} \geq \sqrt{\sum_{i=1}^{d} \frac{1}{c_i} \sum_{s=1}^{t} g_{s,i}^2} \sqrt{\sum_{i=1}^{d} c_i} \geq \sum_{i=1}^{d} \sqrt{\sum_{s=1}^{t} g_{s,i}^2}.$$

B.3. Full-matrix AdaGrad

For full-matrix AdaGrad, $\mathcal{H} = \mathcal{S}_+^d$.

Calculation for $\boldsymbol{H}_t$. Note that $\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1} \rangle + \eta^2 \operatorname{Tr}(\boldsymbol{H})$ is a convex function of $\boldsymbol{H} \succ 0$, so we can get $\boldsymbol{H}_t$ by first calculating the gradient of the objective function and then setting it to zero. Specifically, for any $\boldsymbol{H} \succ 0$, we have

$$\nabla_{\boldsymbol{H}} \left(\left\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1} \right\rangle + \eta^2 \operatorname{Tr}(\boldsymbol{H}) \right) = -\boldsymbol{H}^{-1} \boldsymbol{M}_t \boldsymbol{H}^{-1} + \eta^2 \boldsymbol{I}_d.$$

Setting it to zero yields

$$\boldsymbol{H}_t = \frac{1}{\eta} \boldsymbol{M}_t^{\frac{1}{2}} = \frac{1}{\eta} \left(\epsilon \boldsymbol{I}_d + \sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\top \right)^{\frac{1}{2}}.$$

This corresponds to the update rule of full-matrix AdaGrad.

Calculation for $\|\cdot\|_{\mathcal{H}}$. For the associated $\|\cdot\|_{\mathcal{H}}$, we have

$$\begin{aligned}\|\boldsymbol{x}\|_{\mathcal{H}} &= \sup_{\boldsymbol{H}\succeq 0,\operatorname{Tr}(\boldsymbol{H})\le 1} \sqrt{\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x}} \\ &= \sup_{\boldsymbol{H}\succeq 0,\operatorname{Tr}(\boldsymbol{H})\le 1} \sqrt{\langle \boldsymbol{x}\boldsymbol{x}^\top, \boldsymbol{H}\rangle} = \sqrt{\lambda_1(\boldsymbol{x}\boldsymbol{x}^\top)} = \|\boldsymbol{x}\|_2 .\end{aligned}$$

Calculation for $|||\boldsymbol{g}_{1:t}|||_{\mathcal{H}}$. We can adapt the calculation for $\boldsymbol{H}_t$ to the constrained optimization problem in the definition of $|||\boldsymbol{g}_{1:t}|||_{\mathcal{H}}$ to get

$$\begin{aligned}|||\boldsymbol{g}_{1:t}|||_{\mathcal{H}} &= \inf_{\boldsymbol{H}\in\mathcal{H},\operatorname{Tr}(\boldsymbol{H})\le 1} \sqrt{\Big\langle \sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top, \boldsymbol{H}^{-1}\Big\rangle} \\ &= \sqrt{\Big\langle \sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top, \Big(\frac{1}{\operatorname{Tr}[(\sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top)^{\frac12}]}\Big(\sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top\Big)^{\frac12}\Big)^{-1}\Big\rangle} \\ &= \operatorname{Tr}\Big[\Big(\sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top\Big)^{\frac12}\Big].\end{aligned}$$

B.4. One-sided Shampoo

For one-sided Shampoo, $\mathcal{H} = \mathcal{S}_+^{d_L}\otimes \boldsymbol{I}_{d_R}$.

Calculation for $\boldsymbol{H}_t$. For any $\boldsymbol{H} = \boldsymbol{H}_L\otimes\boldsymbol{I}_{d_R}\in\mathcal{H}$, $\boldsymbol{H}^{-1} = \boldsymbol{H}_L^{-1}\otimes\boldsymbol{I}_{d_R}$, so we have $\operatorname{Tr}(\boldsymbol{H}) = d_R\operatorname{Tr}(\boldsymbol{H}_L)$ and

$$\begin{aligned}\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1}\rangle &= \langle \boldsymbol{M}_t, (\boldsymbol{H}_L\otimes\boldsymbol{I}_{d_R})^{-1}\rangle \\ &= \langle \epsilon\boldsymbol{I}_d, \boldsymbol{H}_L^{-1}\otimes\boldsymbol{I}_{d_R}\rangle + \sum_{s=1}^t \langle \overline{\operatorname{vec}}(\boldsymbol{G}_s)\,\overline{\operatorname{vec}}(\boldsymbol{G}_s)^\top, \boldsymbol{H}_L^{-1}\otimes\boldsymbol{I}_{d_R}\rangle \\ &= \epsilon d_R\operatorname{Tr}(\boldsymbol{H}_L^{-1}) + \sum_{s=1}^t \operatorname{Tr}(\boldsymbol{G}_s^\top\boldsymbol{H}_L^{-1}\boldsymbol{G}_s) \\ &= \operatorname{Tr}\Big[\Big(\epsilon d_R\boldsymbol{I}_{d_L} + \sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\Big)\boldsymbol{H}_L^{-1}\Big]\end{aligned}$$

where we use the fact that $\langle\overline{\operatorname{vec}}(\boldsymbol{X})\,\overline{\operatorname{vec}}(\boldsymbol{X})^\top, \boldsymbol{H}_L\otimes\boldsymbol{I}_{d_R}\rangle = \langle \boldsymbol{X}\boldsymbol{X}^\top, \boldsymbol{H}_L\rangle$ for any $\boldsymbol{X}\in\mathbb{R}^{d_L\times d_R}$. Again, since the objective function is convex in $\boldsymbol{H}_L\succ 0$, we can derive $\boldsymbol{H}_L$ by first calculating the gradient and then setting it to zero. Taking derivative with respect to $\boldsymbol{H}_L$, we obtain

$$\nabla_{\boldsymbol{H}_L}\left(\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1}\rangle + \eta^2\operatorname{Tr}(\boldsymbol{H})\right) = -\boldsymbol{H}_L^{-1}\Big(\epsilon d_R\boldsymbol{I}_{d_L} + \sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\Big)\boldsymbol{H}_L^{-1} + \eta^2 d_R\boldsymbol{I}_{d_L}.$$

Setting it to 0, we obtain

$$\begin{aligned}\arg\min_{\boldsymbol{H}_L\in\mathcal{S}_+^{d_L}} \langle \boldsymbol{M}_t, (\boldsymbol{H}_L\otimes\boldsymbol{I}_{d_R})^{-1}\rangle + \eta^2 d_R\operatorname{Tr}(\boldsymbol{H}_L\otimes\boldsymbol{I}_{d_R}) &= \frac{1}{\eta\sqrt{d_R}}\Big(\epsilon d_R\boldsymbol{I}_{d_L} + \sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\Big)^{\frac12} \\ &= \frac{1}{\eta}\Big(\epsilon\boldsymbol{I}_{d_L} + \frac{1}{d_R}\sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\Big)^{\frac12}.\end{aligned}$$

This is exactly the $\boldsymbol{L}_t$ in Algorithm 2. Therefore, the preconditioner $\boldsymbol{H}_t$ in one-sided Shampoo is given by

$$\boldsymbol{H}_t = \frac{1}{\eta}\Big(\epsilon\boldsymbol{I}_{d_L} + \frac{1}{d_R}\sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\Big)^{\frac12}\otimes\boldsymbol{I}_{d_R}.$$

As a result, Algorithm 1 with $\mathcal{H} = \mathcal{S}_+^{d_L}\otimes\boldsymbol{I}_{d_R}$ recovers Algorithm 2.

Calculation for $\|\cdot\|_{\mathcal{H}}$. For the associated $\|\cdot\|_{\mathcal{H}}$, we have

$$\begin{aligned}\|\boldsymbol{x}\|_{\mathcal{H}} &= \sup_{\boldsymbol{H}=\boldsymbol{H}_L\otimes\boldsymbol{I}_{d_R},\boldsymbol{H}_L\succeq 0,\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{x}} \\ &= \sup_{\boldsymbol{H}_L\succeq 0,\mathrm{Tr}(\boldsymbol{H}_L)\le 1/d_R} \sqrt{\langle \overline{\mathrm{vec}}\,(\boldsymbol{X})\,\overline{\mathrm{vec}}\,(\boldsymbol{X})^\top, \boldsymbol{H}_L\otimes\boldsymbol{I}_{d_R}\rangle} \\ &= \sup_{\boldsymbol{H}_L\succeq 0,\mathrm{Tr}(\boldsymbol{H}_L)\le 1/d_R} \sqrt{\langle \boldsymbol{X}\boldsymbol{X}^\top, \boldsymbol{H}_L\rangle} \\ &= \frac{1}{\sqrt{d_R}}\|\boldsymbol{X}\|_{\mathrm{op}}\end{aligned}$$

where the last equality is achieved at $\boldsymbol{H}_L = \boldsymbol{u}\boldsymbol{u}^\top/d_R$ for $\boldsymbol{u}$ being the leading eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$. The derivation above provides a proof for Lemma 4.1.

Calculation for $\|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}}$. Again, for the adaptive gradient norm, we can adapt the calculation for $\boldsymbol{H}_t$ to the constrained optimization problem in the definition of $\|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}}$ to get

$$\begin{aligned}\|\!|\boldsymbol{g}_{1:t}|\!\|_{\mathcal{H}} &= \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\left\langle \sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top, \boldsymbol{H}^{-1}\right\rangle} \\ &= \inf_{\boldsymbol{H}_L\in\mathcal{S}_+^{d_L},\mathrm{Tr}(\boldsymbol{H}_L)\le\frac{1}{d_R}} \sqrt{\left\langle \sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top, \boldsymbol{H}_L^{-1}\right\rangle} \\ &= \sqrt{\left\langle \sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top, \left[\frac{1}{d_R\,\mathrm{Tr}[(\sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top)^{\frac12}]}\left(\sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\right)^{\frac12}\right]^{-1}\right\rangle} \\ &= \sqrt{d_R}\,\mathrm{Tr}\left[\left(\sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\right)^{\frac12}\right] = \mathrm{Tr}\left[\left(d_R\sum_{s=1}^t \boldsymbol{G}_s\boldsymbol{G}_s^\top\right)^{\frac12}\right].\end{aligned}$$

C. Analysis for EMA adaptive optimization

As mentioned in Section 3, our analysis can be generalized to Algorithm 4, which is an EMA variant of Algorithm 1. To simplify the theoretical analysis, we instead focus on Algorithm 5 in which the current gradient is not scaled by $1-\beta_2$ when computing $\boldsymbol{M}_t$. It serves as a bridge between Algorithms 1 and 4. When $\beta_2 = 1$, Algorithm 5 exactly recovers Algorithm 1. When the learning rate $\tilde{\eta}$ is rescaled by $\sqrt{1-\beta_2}$ and ϵ is rescaled by $1/(1-\beta_2)$, Algorithm 5 is equivalent to Algorithm 4. Theorem C.1 extends Theorem 3.4 to get online regret bound for Algorithm 5. The proof is in Appendix D.3.

Theorem C.1. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Then for any convex loss functions $\tilde{L}_1,\ldots,\tilde{L}_T$, the regret of Algorithm 5 compared to any $\boldsymbol{x}^*\in\mathcal{X}$ can be bounded as*

$$\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}\left[\tilde{L}_t(\boldsymbol{x}_t)-\tilde{L}_t(\boldsymbol{x}^*)\right] \le \left(\frac{D^2}{2\tilde{\eta}}+\tilde{\eta}\right)\inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1}\sqrt{\left\langle \tilde{\boldsymbol{M}}_T, \boldsymbol{H}^{-1}\right\rangle}$$

where $D = \max_{t\in[T]}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}$.

Theorem C.2 provides the convergence rate for Algorithm 5, which generalizes Theorem 3.8. The proof is in Appendix D.3.

Theorem C.2. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Consider any independent stochastic convex loss functions $\tilde{L}_1,\ldots,\tilde{L}_T$ satisfying Assumption 3.7, and let $H(\tilde{L},\mathcal{H})$ be the $\mathcal{H}$-smoothness of their expectation $\tilde{L}$. Suppose the global minimizer of L, denoted by $\boldsymbol{x}^*$, is in $\mathcal{X}$. Then for the iterates $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T$ of Algorithm 5, denoting $\bar{\boldsymbol{x}}_{1:T} = (\sum_{t=1}^T \beta_2^{\frac{T-t}{2}})^{-1}\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}\boldsymbol{x}_t$, it holds that*

$$\mathbb{E}\tilde{L}(\bar{\boldsymbol{x}}_{1:T})-\tilde{L}_t(\boldsymbol{x}^*) \le \frac{16}{\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}}\|\mathcal{X}\|_{\mathcal{H}}^2 H(\tilde{L},\mathcal{H}) + \frac{4\sqrt{2}}{\sqrt{\sum_{t=1}^T \beta_2^{T-t}}}\|\mathcal{X}\|_{\mathcal{H}}\sigma + \frac{4\sqrt{2}}{\sum_{t=1}^T \beta_2^{\frac{-t}{2}}}\|\mathcal{X}\|_{\mathcal{H}}\, d\sqrt{\epsilon}$$

Algorithm 4 EMA Adaptive Regularization Meta-Algorithm

Hyperparam: $\epsilon > 0$, convex set $\mathcal{X} \subseteq \mathbb{R}^d$, learning rate η, preconditioners $\mathcal{H} \subset \mathcal{S}_+^d$, $\beta_2 \in (0,1)$
Input: initialization $\boldsymbol{x}_1$, loss functions $\{L_t\}_{t=1}^T : \mathbb{R}^d \to \mathbb{R}$
$\boldsymbol{M}_0 \leftarrow \epsilon \boldsymbol{I}_d$
for $t = 1, 2, \ldots, T$ **do**
  $\boldsymbol{g}_t \leftarrow \nabla L_t(\boldsymbol{x}_t)$
  $\boldsymbol{M}_t \leftarrow \beta_2 \boldsymbol{M}_{t-1} + (1-\beta_2)\boldsymbol{g}_t \boldsymbol{g}_t^\top$
  $\boldsymbol{H}_t \leftarrow \arg\min_{\boldsymbol{H} \in \mathcal{H}} \left\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1} \right\rangle + \eta^2 \operatorname{Tr}(\boldsymbol{H})$
  $\boldsymbol{x}_{t+1} \leftarrow \Pi_{\mathcal{X}}^{\boldsymbol{H}_t} \left(\boldsymbol{x}_t - \boldsymbol{H}_t^{-1} \boldsymbol{g}_t \right)$
Return $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$

Algorithm 5 Weighted Adaptive Regularization Meta-Algorithm

Hyperparam: $\epsilon > 0$, convex set $\mathcal{X} \subseteq \mathbb{R}^d$, learning rate $\tilde{\eta}$, preconditioners $\mathcal{H} \subset \mathcal{S}_+^d$, $\beta_2 \in (0,1)$
Input: initialization $\boldsymbol{x}_1$, loss functions $\{\tilde{L}_t\}_{t=1}^T : \mathbb{R}^d \to \mathbb{R}$
$\tilde{\boldsymbol{M}}_0 \leftarrow \epsilon \boldsymbol{I}_d$
for $t = 1, 2, \ldots, T$ **do**
  $\tilde{\boldsymbol{g}}_t \leftarrow \nabla \tilde{L}_t(\tilde{\boldsymbol{x}}_t)$
  $\tilde{\boldsymbol{M}}_t \leftarrow \beta_2 \tilde{\boldsymbol{M}}_{t-1} + \tilde{\boldsymbol{g}}_t \tilde{\boldsymbol{g}}_t^\top$
  $\tilde{\boldsymbol{H}}_t \leftarrow \arg\min_{\tilde{\boldsymbol{H}} \in \mathcal{H}} \left\langle \boldsymbol{M}_t, \tilde{\boldsymbol{H}}^{-1} \right\rangle + \tilde{\eta}^2 \operatorname{Tr}(\tilde{\boldsymbol{H}})$
  $\tilde{\boldsymbol{x}}_{t+1} \leftarrow \Pi_{\mathcal{X}}^{\tilde{\boldsymbol{H}}_t} \left(\tilde{\boldsymbol{x}}_t - \tilde{\boldsymbol{H}}_t^{-1} \tilde{\boldsymbol{g}}_t \right)$
Return $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_T$

where $\sigma = \inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \sqrt{\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1} \rangle}$.

When $\beta_2 = 1$, Theorem C.2 recovers the result in Theorem 3.8 and provides a $O(T^{-\frac{1}{2}})$ convergence rate. When $\beta_2 < 1$, the optimality gap is upper bounded by $O\big((1-\beta_2) \left\|\mathcal{X}\right\|_{\mathcal{H}}^2 H(\tilde{L}, \mathcal{H}) + \sqrt{1-\beta_2} \left\|\mathcal{X}\right\|_{\mathcal{H}} \sigma\big)$ when $T = \Theta\left(\frac{1}{1-\beta_2}\right)$.

D. Proof for the Unified Analysis

D.1. Regret bound

We first present the proof for the main result on the regret bound for Algorithm 1 with a well-structured preconditioner set $\mathcal{H}$.

Theorem 3.4. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Then for any convex loss functions $L_1, \ldots, L_T$, the regret of Algorithm 1 compared to any $\boldsymbol{x}^* \in \mathcal{X}$ can be bounded as*

$$\sum_{t=1}^T L_t(\boldsymbol{x}_t) - \sum_{t=1}^T L_t(\boldsymbol{x}^*) \le \left(\frac{D^2}{2\eta} + \eta \right) \left(G + d\sqrt{\epsilon} \right)$$

where $G = \left\|\left\|\boldsymbol{g}_{1:T}\right\|\right\|_{\mathcal{H}}$, $D = \max_{t \in [T]} \left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|_{\mathcal{H}}$.

Proof of Theorem 3.4. First we will analyze the property of each $\boldsymbol{H}_t$. Recall from Proposition 3.2 that $\boldsymbol{H}_t$ satisfies $\left\langle \boldsymbol{M}_t, \boldsymbol{H}_t^{-1} \right\rangle = \eta^2 \operatorname{Tr}(\boldsymbol{H}_t)$ and that $\overline{\boldsymbol{H}}_t := \operatorname{Tr}(\boldsymbol{H}_t)^{-1} \boldsymbol{H}_t = \arg\min_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \left\langle \boldsymbol{M}_t, \boldsymbol{H}^{-1} \right\rangle$. Therefore,

$$\left\langle \boldsymbol{M}_t, \boldsymbol{H}_t^{-1} \right\rangle = \eta^2 \operatorname{Tr}(\boldsymbol{H}_t) = \eta \sqrt{\left\langle \boldsymbol{M}_t, \boldsymbol{H}_t^{-1} \right\rangle \operatorname{Tr}(\boldsymbol{H}_t)} = \eta \sqrt{\langle \boldsymbol{M}_t, \overline{\boldsymbol{H}}_t^{-1} \rangle}. \tag{25}$$

Now recall the regret bound from Theorem 2.1:

$$\begin{aligned}
\sum_{t=1}^T L_t(\boldsymbol{x}_t) - \sum_{t=1}^T L_t(\boldsymbol{x}^*) &\le \frac{1}{2} \left(\left\langle \boldsymbol{M}_T, \boldsymbol{H}_T^{-1} \right\rangle + \eta^2 \operatorname{Tr}(\boldsymbol{H}_T) - \eta^2 \operatorname{Tr}(\boldsymbol{H}_0) \right) \\
&\quad + \frac{1}{2} \sum_{t=1}^T \left(\left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|_{\boldsymbol{H}_t}^2 - \left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right\|_{\boldsymbol{H}_t}^2 \right) \\
&= \frac{1}{2} \left(2\eta \sqrt{\langle \boldsymbol{M}_T, \overline{\boldsymbol{H}}_T^{-1} \rangle} - \eta^2 \operatorname{Tr}(\boldsymbol{H}_0) \right) \\
&\quad + \frac{1}{2} \sum_{t=1}^T \left(\left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|_{\boldsymbol{H}_t}^2 - \left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right\|_{\boldsymbol{H}_t}^2 \right)
\end{aligned} \tag{26}$$

where the equality follows from the facts in (25). Next, for the second term on the right-hand side of (26), we rearrange the summation to obtain

$$\sum_{t=1}^{T}\left(\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\boldsymbol{H}_t}^2-\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\|_{\boldsymbol{H}_t}^2\right)\le\|\boldsymbol{x}_1-\boldsymbol{x}^*\|_{\boldsymbol{H}_1}^2+\sum_{t=2}^{T}\left(\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\boldsymbol{H}_t}^2-\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\boldsymbol{H}_{t-1}}^2\right)$$
$$=\|\boldsymbol{x}_1-\boldsymbol{x}^*\|_{\boldsymbol{H}_1}^2+\sum_{t=2}^{T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\boldsymbol{H}_t-\boldsymbol{H}_{t-1}}^2.$$

Notice that $\boldsymbol{M}_t-\boldsymbol{M}_{t-1}=\boldsymbol{g}_t\boldsymbol{g}_t^\top\succeq 0$, and thus $\boldsymbol{H}_t-\boldsymbol{H}_{t-1}\in\mathcal{H}$ by Proposition 3.2. This implies

$$\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\boldsymbol{H}_t-\boldsymbol{H}_{t-1}}^2=(\boldsymbol{x}_t-\boldsymbol{x}^*)^\top(\boldsymbol{H}_t-\boldsymbol{H}_{t-1})(\boldsymbol{x}_t-\boldsymbol{x}^*)\le\operatorname{Tr}(\boldsymbol{H}_t-\boldsymbol{H}_{t-1})\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2$$

where the inequality follows from the definition of $\|\cdot\|_{\mathcal{H}}$ in (1). It then follows that

$$\sum_{t=1}^{T}\left(\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\boldsymbol{H}_t}^2-\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\|_{\boldsymbol{H}_t}^2\right)\le\operatorname{Tr}(\boldsymbol{H}_1)\|\boldsymbol{x}_1-\boldsymbol{x}^*\|_{\mathcal{H}}^2+\sum_{t=2}^{T}\operatorname{Tr}(\boldsymbol{H}_t-\boldsymbol{H}_{t-1})\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2$$
$$\le\operatorname{Tr}(\boldsymbol{H}_T)\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2$$
$$=\frac{1}{\eta}\sqrt{\langle\boldsymbol{M}_T,\overline{\boldsymbol{H}}_T^{-1}\rangle}\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2$$

where the equality again follows from (25). Plugging this back into (26), we obtain

$$\sum_{t=1}^{T}L_t(\boldsymbol{x}_t)-\sum_{t=1}^{T}L_t(\boldsymbol{x}^*)\le\eta\sqrt{\langle\boldsymbol{M}_T,\overline{\boldsymbol{H}}_T^{-1}\rangle}+\frac{1}{2\eta}\sqrt{\langle\boldsymbol{M}_T,\overline{\boldsymbol{H}}_T^{-1}\rangle}\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2$$
$$=\left(\frac{\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\eta}+\eta\right)\inf_{\boldsymbol{H}\in\mathcal{H},\operatorname{Tr}(\boldsymbol{H})\le 1}\sqrt{\langle\boldsymbol{M}_T,\boldsymbol{H}^{-1}\rangle}$$
$$\le\left(\frac{\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\eta}+\eta\right)\sqrt{\langle\boldsymbol{M}_T-\boldsymbol{M}_0,\boldsymbol{H}^{-1}\rangle+\langle\boldsymbol{M}_0,\boldsymbol{H}^{-1}\rangle}$$

for any $\boldsymbol{H}\in\mathcal{H}$ with $\operatorname{Tr}(\boldsymbol{H})=1$. In particular, we choose $\boldsymbol{H}=\alpha\boldsymbol{H}_T^*+\frac{1-\alpha}{d}\boldsymbol{I}_d$ for some $\alpha\in(0,1)$ where $\boldsymbol{H}_T^*=\arg\min_{\boldsymbol{H}\in\mathcal{H},\operatorname{Tr}(\boldsymbol{H})\le 1}\sqrt{\langle\boldsymbol{M}_T-\boldsymbol{M}_0,\boldsymbol{H}^{-1}\rangle}$, so $\langle\boldsymbol{M}_T-\boldsymbol{M}_0,\boldsymbol{H}^{-1}\rangle=\|\boldsymbol{g}_{1:T}\|_{\mathcal{H}}^2$ according to the definition of the adaptive gradient norm in (7). Since $\boldsymbol{H}^{-1}\preceq\frac{1}{\alpha}(\boldsymbol{H}_T^*)^{-1}$ and $\boldsymbol{H}^{-1}\preceq\frac{d}{1-\alpha}\boldsymbol{I}_d$, we further have

$$\sum_{t=1}^{T}L_t(\boldsymbol{x}_t)-\sum_{t=1}^{T}L_t(\boldsymbol{x}^*)\le\left(\frac{\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\eta}+\eta\right)\sqrt{\frac{1}{\alpha}\langle\boldsymbol{M}_T-\boldsymbol{M}_0,(\boldsymbol{H}_T^*)^{-1}\rangle+\frac{d}{1-\alpha}\langle\boldsymbol{M}_0,\boldsymbol{I}_d\rangle}$$
$$=\left(\frac{\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\eta}+\eta\right)\sqrt{\frac{1}{\alpha}\|\boldsymbol{g}_{1:T}\|_{\mathcal{H}}^2+\frac{d^2\epsilon}{1-\alpha}}.$$

Finally choosing $\alpha=\frac{\|\boldsymbol{g}_{1:T}\|_{\mathcal{H}}}{\|\boldsymbol{g}_{1:T}\|_{\mathcal{H}}+d\sqrt{\epsilon}}$, we obtain

$$\sum_{t=1}^{T}L_t(\boldsymbol{x}_t)-\sum_{t=1}^{T}L_t(\boldsymbol{x}^*)=\left(\frac{\max_{1\le t\le T}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\eta}+\eta\right)\left(\|\boldsymbol{g}_{1:T}\|_{\mathcal{H}}+d\sqrt{\epsilon}\right)$$

This completes the proof. □

The proof for Corollary 3.5 is straightforward.

Corollary D.1. *Under the setting of Theorem 2.1, further suppose that $\mathcal{X}$ is a bounded set in $\mathbb{R}^d$. Then choosing $\eta=\sqrt{2}\|\mathcal{X}\|_{\mathcal{H}}$, the regret bound for Algorithm 1 becomes*

$$\sum_{t=1}^{T}L_t(\boldsymbol{x}_t)-\sum_{t=1}^{T}L_t(\boldsymbol{x}^*)\le 2\sqrt{2}\|\mathcal{X}\|_{\mathcal{H}}\left(G+d\sqrt{\epsilon}\right).$$

Proof of Corollary 3.5. Note that $D = \max_{t\in[T]} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\mathcal{H}} \le \max_{t\in[T]} \|\boldsymbol{x}_t\|_{\mathcal{H}} + \|\boldsymbol{x}^*\|_{\mathcal{H}} \le 2\|\mathcal{X}\|_{\mathcal{H}}$ because $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ and $\boldsymbol{x}^*$ are all in $\mathcal{X}$. The proof is completed by setting $\eta = \sqrt{2}\|\mathcal{X}\|_{\mathcal{H}}$ to minimize $2\frac{\|\mathcal{X}\|_{\mathcal{H}}^2}{\eta} + \eta$. □

D.2. Convergence rate

Next, we present the proof for the convergence rate of Algorithm 1 with a well-structured preconditioner set $\mathcal{H}$.

Theorem 3.8. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Consider any independent stochastic convex loss functions $L_1, \dots, L_T$ satisfying Assumption 3.7, and let $H(L, \mathcal{H})$ be the $\mathcal{H}$-smoothness of their expectation L. Suppose the global minimizer of L, denoted by $\boldsymbol{x}^*$, is in $\mathcal{X}$. Then for the iterates $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ of Algorithm 1, denoting $\bar{\boldsymbol{x}}_{1:T} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t$, it holds that*

$$\begin{aligned}&\mathbb{E}\left[L(\bar{\boldsymbol{x}}_{1:T}) - L(\boldsymbol{x}^*)\right]\\ \le&\frac{16}{T}\|\mathcal{X}\|_{\mathcal{H}}^2 H(L,\mathcal{H}) + \frac{4\sqrt{2}}{\sqrt{T}}\|\mathcal{X}\|_{\mathcal{H}}\sigma + \frac{4\sqrt{2}d\sqrt{\epsilon}}{T}\|\mathcal{X}\|_{\mathcal{H}}\end{aligned}$$

where $\sigma = \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1}\sqrt{\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1}\rangle}$.

Proof of Theorem 3.8. Let $\boldsymbol{H}^* \in \mathcal{H}$ be the matrix given in Definition 3.6 for the loss function L, such that $\nabla^2 L(\boldsymbol{x}) \preceq \boldsymbol{H}^*$. Therefore, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$\begin{aligned}L(\boldsymbol{x}') &\le L(\boldsymbol{x}) + \nabla L(\boldsymbol{x})^\top(\boldsymbol{x}' - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{x}' - \boldsymbol{x})^\top \boldsymbol{H}^*(\boldsymbol{x}' - \boldsymbol{x})\\ &= L(\boldsymbol{x}) + \frac{1}{2}\left(\boldsymbol{x}' - \boldsymbol{x} + (\boldsymbol{H}^*)^{-1}\nabla L(\boldsymbol{x})\right)^\top \boldsymbol{H}^*\left(\boldsymbol{x}' - \boldsymbol{x} + (\boldsymbol{H}^*)^{-1}\nabla L(\boldsymbol{x})\right) - \frac{1}{2}\nabla L(\boldsymbol{x})^\top(\boldsymbol{H}^*)^{-1}\nabla L(\boldsymbol{x}).\end{aligned}$$

Then we have that

$$\begin{aligned}L(\boldsymbol{x}^*) &= \min_{\boldsymbol{x}'} L(\boldsymbol{x}')\\ &\le \min_{\boldsymbol{x}'} L(\boldsymbol{x}) + \frac{1}{2}\left(\boldsymbol{x}' - \boldsymbol{x} + (\boldsymbol{H}^*)^{-1}\nabla L(\boldsymbol{x})\right)^\top \boldsymbol{H}^*\left(\boldsymbol{x}' - \boldsymbol{x} + (\boldsymbol{H}^*)^{-1}\nabla L(\boldsymbol{x})\right) - \frac{1}{2}\nabla L(\boldsymbol{x})^\top(\boldsymbol{H}^*)^{-1}\nabla L(\boldsymbol{x})\\ &= L(\boldsymbol{x}) - \frac{1}{2}\nabla L(\boldsymbol{x})^\top(\boldsymbol{H}^*)^{-1}\nabla L(\boldsymbol{x}).\end{aligned}$$

Applying the above inequality to $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ and denoting $\bar{\boldsymbol{g}}_t = \nabla L(\boldsymbol{x}_t)$, we then have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T (L(\boldsymbol{x}_t) - L(\boldsymbol{x}^*))\right] &\ge \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \bar{\boldsymbol{g}}_t^\top (\boldsymbol{H}^*)^{-1}\bar{\boldsymbol{g}}_t\right]\\ &= \frac{1}{2}\left\langle \mathbb{E}\left[\sum_{t=1}^T \bar{\boldsymbol{g}}_t\bar{\boldsymbol{g}}_t^\top\right], (\boldsymbol{H}^*)^{-1}\right\rangle\\ &= \frac{1}{2\,\mathrm{Tr}(\boldsymbol{H}^*)}\left\langle \mathbb{E}\left[\sum_{t=1}^T \bar{\boldsymbol{g}}_t\bar{\boldsymbol{g}}_t^\top\right], (\boldsymbol{H}^*/\mathrm{Tr}(\boldsymbol{H}^*))^{-1}\right\rangle\\ &\ge \frac{1}{2H(L,\mathcal{H})}\inf_{\mathrm{Tr}(\boldsymbol{H})\le 1, \boldsymbol{H}\in\mathcal{H}}\left\langle \mathbb{E}\left[\sum_{t=1}^T \bar{\boldsymbol{g}}_t\bar{\boldsymbol{g}}_t^\top\right], \boldsymbol{H}^{-1}\right\rangle. \qquad (27)\end{aligned}$$

For any $\delta > 0$, we choose $\boldsymbol{H}_{\boldsymbol{g}} \in \mathcal{H}$ such that $\langle \mathbb{E}[\sum_{t=1}^T \bar{\boldsymbol{g}}_t\bar{\boldsymbol{g}}_t^\top], \boldsymbol{H}_{\boldsymbol{g}}^{-1}\rangle \le \delta + \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1}\langle \mathbb{E}[\sum_{t=1}^T \bar{\boldsymbol{g}}_t\bar{\boldsymbol{g}}_t^\top], \boldsymbol{H}^{-1}\rangle$ and $\mathrm{Tr}(\boldsymbol{H}_{\boldsymbol{g}}) \le 1$. Similarly, we choose $\boldsymbol{H}_{\boldsymbol{\Sigma}} \in \mathcal{H}$ such that $\langle \boldsymbol{\Sigma}, \boldsymbol{H}_{\boldsymbol{\Sigma}}^{-1}\rangle \le \delta + \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1}\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1}\rangle$ and $\mathrm{Tr}(\boldsymbol{H}_{\boldsymbol{\Sigma}}) \le 1$. Correspondingly, we define $\boldsymbol{H}' = \alpha\boldsymbol{H}_{\boldsymbol{g}} + (1-\alpha)\boldsymbol{H}_{\boldsymbol{\Sigma}}$, which satisfies that $\boldsymbol{H}' \in \mathcal{H}$ and $\mathrm{Tr}(\boldsymbol{H}') \le 1$. Then by Jensen's inequality, we have

$$\mathbb{E}\inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1}\sqrt{\langle \boldsymbol{M}_T - \boldsymbol{M}_0, \boldsymbol{H}^{-1}\rangle} \le \mathbb{E}\sqrt{\langle \boldsymbol{M}_T - \boldsymbol{M}_0, (\boldsymbol{H}')^{-1}\rangle} \le \sqrt{\mathbb{E}\langle \boldsymbol{M}_T - \boldsymbol{M}_0, (\boldsymbol{H}')^{-1}\rangle}.$$

Plugging in $\boldsymbol{M}_T - \boldsymbol{M}_0 = \sum_{t=1}^T \boldsymbol{g}_t \boldsymbol{g}_t^\top$, since $\mathbb{E}[\sum_{t=1}^t \boldsymbol{g}_t \boldsymbol{g}_t^\top] \preceq \mathbb{E}[\sum_{t=1}^T \bar{\boldsymbol{g}}_t \bar{\boldsymbol{g}}_t^\top] + T\boldsymbol{\Sigma}$, we further have

$$\begin{aligned}
\mathbb{E} \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\langle \boldsymbol{M}_T - \boldsymbol{M}_0, \boldsymbol{H}^{-1}\rangle} &\le \sqrt{\left\langle \mathbb{E}\left[\sum_{t=1}^T \bar{\boldsymbol{g}}_t \bar{\boldsymbol{g}}_t^\top\right], (\boldsymbol{H}')^{-1}\right\rangle + T\left\langle \boldsymbol{\Sigma}, (\boldsymbol{H}')^{-1}\right\rangle} \\
&\le \sqrt{\frac{1}{\alpha}\left\langle \mathbb{E}\left[\sum_{t=1}^T \bar{\boldsymbol{g}}_t \bar{\boldsymbol{g}}_t^\top\right], \boldsymbol{H}_{\boldsymbol{g}}^{-1}\right\rangle + \frac{1}{1-\alpha} T\left\langle \boldsymbol{\Sigma}, \boldsymbol{H}_{\boldsymbol{\Sigma}}^{-1}\right\rangle} \\
&= \sqrt{\left\langle \mathbb{E}\left[\sum_{t=1}^T \bar{\boldsymbol{g}}_t \bar{\boldsymbol{g}}_t^\top\right], \boldsymbol{H}_{\boldsymbol{g}}^{-1}\right\rangle} + \sqrt{T\left\langle \boldsymbol{\Sigma}, \boldsymbol{H}_{\boldsymbol{\Sigma}}^{-1}\right\rangle}
\end{aligned}$$

where in the last step we choose α to be

$$\alpha = \frac{\sqrt{\langle \mathbb{E}[\sum_{t=1}^T \bar{\boldsymbol{g}}_t \bar{\boldsymbol{g}}_t^\top], \boldsymbol{H}_{\boldsymbol{g}}^{-1}\rangle}}{\sqrt{\langle \mathbb{E}[\sum_{t=1}^T \bar{\boldsymbol{g}}_t \bar{\boldsymbol{g}}_t^\top], \boldsymbol{H}_{\boldsymbol{g}}^{-1}\rangle} + \sqrt{T\langle \boldsymbol{\Sigma}, \boldsymbol{H}_{\boldsymbol{\Sigma}}^{-1}\rangle}}.$$

Recall that $\sigma = \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1}\rangle}$. Then it follows from (27) and the definitions of $\boldsymbol{H}_{\boldsymbol{g}}, \boldsymbol{H}_{\boldsymbol{\Sigma}}$ that

$$\mathbb{E}\Big[\inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\langle \boldsymbol{M}_T - \boldsymbol{M}_0, \boldsymbol{H}^{-1}\rangle}\Big] \le \sqrt{2H(L,\mathcal{H})\cdot \mathbb{E}\left[\sum_{t=1}^T (L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*))\right] + \delta} + \sqrt{T\sigma^2 + T\delta}$$

Since $\delta > 0$ is arbitrary, this implies that

$$\mathbb{E}\Big[\inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\langle \boldsymbol{M}_T - \boldsymbol{M}_0, \boldsymbol{H}^{-1}\rangle}\Big] \le \sqrt{2H(L,\mathcal{H})\cdot \mathbb{E}\left[\sum_{t=1}^T (L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*))\right]} + \sqrt{T\sigma^2}$$

Now recall the regret bound from Corollary 3.5, and it follows from the above inequality that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T (L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*))\right] &\le 2\sqrt{2}\|\mathcal{X}\|_{\mathcal{H}} \cdot \mathbb{E}\left[\inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\langle \boldsymbol{M}_t - \boldsymbol{M}_0, \boldsymbol{H}^{-1}\rangle} + d\sqrt{\epsilon}\right] \\
&\le 2\sqrt{2}\,\|\mathcal{X}\|_{\mathcal{H}} \left(\sqrt{2H(L,\mathcal{H})\cdot \mathbb{E}\left[\sum_{t=1}^T (L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*))\right]} + \sqrt{T\sigma^2} + d\sqrt{\epsilon}\right).
\end{aligned}$$

Solving the above inequality yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*))\right] \le \frac{16}{T}\|\mathcal{X}\|_{\mathcal{H}}^2 H(L,\mathcal{H}) + \frac{4\sqrt{2}}{\sqrt{T}}\|\mathcal{X}\|_{\mathcal{H}}\sigma + \frac{4\sqrt{2}}{T}\|\mathcal{X}\|_{\mathcal{H}} d\sqrt{\epsilon}.$$

This completes the proof. □

D.3. Analysis for EMA style optimizers

Here we present the proof for the theorems in Appendix C.

Theorem C.1. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Then for any convex loss functions $\tilde{L}_1, \ldots, \tilde{L}_T$, the regret of Algorithm 5 compared to any $\boldsymbol{x}^* \in \mathcal{X}$ can be bounded as*

$$\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}\left[\tilde{L}_t(\boldsymbol{x}_t) - \tilde{L}_t(\boldsymbol{x}^*)\right] \le \left(\frac{D^2}{2\tilde{\eta}} + \tilde{\eta}\right) \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\le 1} \sqrt{\left\langle \tilde{\boldsymbol{M}}_T, \boldsymbol{H}^{-1}\right\rangle}$$

where $D = \max_{t\in[T]} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\mathcal{H}}$.

Proof of Theorem C.1. We can choose $L_t = \beta_2^{-\frac{t}{2}} \tilde{L}_t$, $\epsilon = \tilde{\epsilon}$ and $\eta = \tilde{\eta}$. Then we claim Algorithm 5 is equivalent to Algorithm 1 with hyperparameter ϵ, η and loss functions $\{L_t\}_{t=1}^T$, which will be shown by induction.

Assume $\tilde{\boldsymbol{x}}_s = \boldsymbol{x}_s$ for $s \le t$. Then we know $\boldsymbol{g}_s = \nabla L_s(\boldsymbol{x}_s) = \beta_2^{-\frac{s}{2}} \nabla \tilde{L}_s(\tilde{\boldsymbol{x}}_s)$ for $s \le t$.

We consider the update in the step t of Algorithm 5.

$$\tilde{\boldsymbol{M}}_t = \sum_{i=1}^t \beta_2^{t-i} \tilde{\boldsymbol{g}}_i \tilde{\boldsymbol{g}}_i^\top + \beta_2^t \tilde{\boldsymbol{M}}_0 = \beta_2^t \left[\sum_{i=1}^t \boldsymbol{g}_t \boldsymbol{g}_t^\top + \epsilon \boldsymbol{I}_d \right] = \beta_2^t \boldsymbol{M}_t,$$

$$\tilde{\boldsymbol{H}}_t = \arg\min_{\tilde{\boldsymbol{H}} \in \mathcal{H}} \left\langle \tilde{\boldsymbol{M}}_t, \tilde{\boldsymbol{H}}^{-1} \right\rangle + \tilde{\eta}^2 \operatorname{Tr}(\tilde{\boldsymbol{H}}) = \arg\min_{\boldsymbol{H} \in \mathcal{H}} \left\langle \beta_2^t \boldsymbol{M}_t, \boldsymbol{H}^{-1} \right\rangle + \eta^2 \operatorname{Tr}(\boldsymbol{H}) = \beta_2^{\frac{t}{2}} \boldsymbol{H}_t,$$

$$\tilde{\boldsymbol{H}}_t^{-1} \tilde{\boldsymbol{g}}_t = \beta_2^{-\frac{t}{2}} \boldsymbol{H}_t^{-1} \beta_2^{\frac{t}{2}} \boldsymbol{g}_t = \boldsymbol{H}_t^{-1} \boldsymbol{g}_t,$$

$$\tilde{\boldsymbol{x}}_{t+1} = \Pi_{\mathcal{X}}^{\tilde{\boldsymbol{H}}_t} \left(\tilde{\boldsymbol{x}}_t - \tilde{\boldsymbol{H}}_t^{-1} \tilde{\boldsymbol{g}}_t \right) = \Pi_{\mathcal{X}}^{\tilde{\boldsymbol{H}}_t} \left(\boldsymbol{x}_t - \boldsymbol{H}_t^{-1} \boldsymbol{g}_t \right) = \Pi_{\mathcal{X}}^{\boldsymbol{H}_t} \left(\boldsymbol{x}_t - \boldsymbol{H}_t^{-1} \boldsymbol{g}_t \right) = \boldsymbol{x}_{t+1}.$$

Therefore, we can obtain the regret bound for Algorithm 4 with Theorem 3.4.

$$\begin{aligned}
\sum_{t=1}^T \beta_2^{-\frac{t}{2}} \left[\tilde{L}_t(\boldsymbol{x}_t) - \tilde{L}_t(\boldsymbol{x}^*) \right] &= \sum_{t=1}^T \left[L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*) \right] \\
&\le \left(\frac{\max_{1 \le t \le T} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\eta} + \eta \right) \inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \sqrt{\langle \boldsymbol{M}_T, \boldsymbol{H}^{-1} \rangle} \\
&= \beta_2^{-\frac{T}{2}} \left(\frac{\max_{1 \le t \le T} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\tilde{\eta}} + \tilde{\eta} \right) \inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \sqrt{\left\langle \tilde{\boldsymbol{M}}_T, \boldsymbol{H}^{-1} \right\rangle}
\end{aligned}$$

and

$$\sum_{t=1}^T \beta_2^{-\frac{t-1}{2}} \frac{\beta_2^{-\frac{1}{2}} - 1}{\beta_2^{-\frac{T}{2}} - 1} \left[\tilde{L}_t(\boldsymbol{x}_t) - \tilde{L}_t(\boldsymbol{x}^*) \right] \le \frac{1 - \sqrt{\beta_2}}{1 - \beta_2^{\frac{T}{2}}} \left(\frac{\max_{1 \le t \le T} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\mathcal{H}}^2}{2\tilde{\eta}} + \tilde{\eta} \right) \inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \sqrt{\left\langle \tilde{\boldsymbol{M}}_T, \boldsymbol{H}^{-1} \right\rangle}$$

□

Theorem C.2. *Let $\mathcal{H}$ be a well-structured preconditioner set under Definition 3.1. Consider any independent stochastic convex loss functions $\tilde{L}_1, \ldots, \tilde{L}_T$ satisfying Assumption 3.7, and let $H(\tilde{L}, \mathcal{H})$ be the $\mathcal{H}$-smoothness of their expectation $\tilde{L}$. Suppose the global minimizer of L, denoted by $\boldsymbol{x}^*$, is in $\mathcal{X}$. Then for the iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ of Algorithm 5, denoting $\bar{\boldsymbol{x}}_{1:T} = (\sum_{t=1}^T \beta_2^{\frac{T-t}{2}})^{-1} \sum_{t=1}^T \beta_2^{\frac{T-t}{2}} \boldsymbol{x}_t$, it holds that*

$$\mathbb{E}\tilde{L}(\bar{\boldsymbol{x}}_{1:T}) - \tilde{L}_t(\boldsymbol{x}^*) \le \frac{16}{\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}} \|\mathcal{X}\|_{\mathcal{H}}^2 H(\tilde{L}, \mathcal{H}) + \frac{4\sqrt{2}}{\sqrt{\sum_{t=1}^T \beta_2^{T-t}}} \|\mathcal{X}\|_{\mathcal{H}} \sigma + \frac{4\sqrt{2}}{\sum_{t=1}^T \beta_2^{\frac{-t}{2}}} \|\mathcal{X}\|_{\mathcal{H}} d\sqrt{\epsilon}$$

where $\sigma = \inf_{\boldsymbol{H} \in \mathcal{H}, \operatorname{Tr}(\boldsymbol{H}) \le 1} \sqrt{\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1} \rangle}$.

Proof of Theorem C.2. Similar to the proof of Theorem 3.8, we know

$$\tilde{L}(\tilde{\boldsymbol{x}}_t) - \tilde{L}(\tilde{\boldsymbol{x}}^*) \ge \left\langle \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\tilde{\boldsymbol{H}}^*)^{-1} \right\rangle$$

and

$$\begin{aligned}
\sum_{t=1}^T \beta_2^{\frac{T-t}{2}} \tilde{L}(\tilde{\boldsymbol{x}}_t) - \tilde{L}(\tilde{\boldsymbol{x}}^*) &\ge \left\langle \sum_{t=1}^T \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\tilde{\boldsymbol{H}}^*)^{-1} \right\rangle \\
&\ge \frac{1}{2H(\tilde{L}, \mathcal{H})} \inf_{\operatorname{Tr}(\boldsymbol{H}) \le 1, \boldsymbol{H} \in \mathcal{H}} \left\langle \sum_{t=1}^T \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, \boldsymbol{H}^{-1} \right\rangle
\end{aligned}$$

For any $\delta > 0$, we choose $\boldsymbol{H}_{\boldsymbol{g}} \in \mathcal{H}$ such that $\mathrm{Tr}(\boldsymbol{H}_{\boldsymbol{g}}) \leq 1$ and

$$\left\langle \sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, \boldsymbol{H}_{\boldsymbol{g}}^{-1} \right\rangle \leq \delta + \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\leq 1} \left\langle \sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, \boldsymbol{H}^{-1} \right\rangle.$$

We also choose $\boldsymbol{H}_{\boldsymbol{\Sigma}} \in \mathcal{H}$ such that $\left\langle \boldsymbol{\Sigma}, \boldsymbol{H}_{\boldsymbol{\Sigma}}^{-1} \right\rangle \leq \delta + \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\leq 1} \left\langle \boldsymbol{\Sigma}, \boldsymbol{H}^{-1} \right\rangle$ and $\mathrm{Tr}(\boldsymbol{H}_{\boldsymbol{\Sigma}}) \leq 1$. We define $\boldsymbol{H}' = \alpha\boldsymbol{H}_{\boldsymbol{g}} + (1-\alpha)\boldsymbol{H}_{\boldsymbol{\Sigma}}$. Then $\boldsymbol{H}' \in \mathcal{H}$ and $\mathrm{Tr}(\boldsymbol{H}') \leq 1$. We have that

$$\begin{aligned}
\mathbb{E} \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\leq 1} \sqrt{\left\langle \tilde{\boldsymbol{M}}_T - \beta_2^T \tilde{\boldsymbol{M}}_0, \boldsymbol{H}^{-1} \right\rangle} &\leq \mathbb{E}\sqrt{\left\langle \tilde{\boldsymbol{M}}_T - \beta_2^T \tilde{\boldsymbol{M}}_0, (\boldsymbol{H}')^{-1} \right\rangle} \\
&\leq \sqrt{\mathbb{E}\left\langle \tilde{\boldsymbol{M}}_T - \beta_2^T \tilde{\boldsymbol{M}}_0, (\boldsymbol{H}')^{-1} \right\rangle} \\
&\leq \sqrt{\left\langle \sum_{t=1}^{T} \beta_2^{T-t} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\boldsymbol{H}')^{-1} \right\rangle + \sum_{t=1}^{T} \beta_2^{T-t} \left\langle \boldsymbol{\Sigma}, (\boldsymbol{H}')^{-1} \right\rangle}.
\end{aligned}$$

Now plugging in the definition of $\boldsymbol{H}'$, we obtain

$$\begin{aligned}
\mathbb{E} \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\leq 1} \sqrt{\left\langle \tilde{\boldsymbol{M}}_T - \beta_2^T \tilde{\boldsymbol{M}}_0, \boldsymbol{H}^{-1} \right\rangle} &\leq \sqrt{\left\langle \sum_{t=1}^{T} \beta_2^{T-t} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\alpha\boldsymbol{H}_{\boldsymbol{g}})^{-1} \right\rangle + \sum_{t=1}^{T} \beta_2^{T-t} \left\langle \boldsymbol{\Sigma}, ((1-\alpha)\boldsymbol{H}_{\boldsymbol{\Sigma}})^{-1} \right\rangle} \\
&= \sqrt{\frac{1}{\alpha}\left\langle \sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\boldsymbol{H}_{\boldsymbol{g}})^{-1} \right\rangle + \frac{1}{1-\alpha}\sum_{t=1}^{T} \beta_2^{T-t} \left\langle \boldsymbol{\Sigma}, (\boldsymbol{H}_{\boldsymbol{\Sigma}})^{-1} \right\rangle} \\
&\leq \sqrt{\left\langle \sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\boldsymbol{H}_{\boldsymbol{g}})^{-1} \right\rangle} + \sqrt{\sum_{t=1}^{T} \beta_2^{T-t} \left\langle \boldsymbol{\Sigma}, (\boldsymbol{H}_{\boldsymbol{\Sigma}})^{-1} \right\rangle}.
\end{aligned}$$

where in the last step we choose α to be

$$\alpha = \frac{\sqrt{\left\langle \sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\boldsymbol{H}_{\boldsymbol{g}})^{-1} \right\rangle}}{\sqrt{\left\langle \sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \mathbb{E}\tilde{\boldsymbol{g}}_t \mathbb{E}\tilde{\boldsymbol{g}}_t^\top, (\boldsymbol{H}_{\boldsymbol{g}})^{-1} \right\rangle} + \sqrt{\sum_{t=1}^{T} \beta_2^{T-t} \left\langle \boldsymbol{\Sigma}, (\boldsymbol{H}_{\boldsymbol{\Sigma}})^{-1} \right\rangle}}.$$

Then we have that

$$\mathbb{E} \inf_{\boldsymbol{H}\in\mathcal{H},\mathrm{Tr}(\boldsymbol{H})\leq 1} \sqrt{\left\langle \tilde{\boldsymbol{M}}_T - \beta_2^T \tilde{\boldsymbol{M}}_0, \boldsymbol{H}^{-1} \right\rangle} \leq \sqrt{2H(\tilde{L},\mathcal{H})\left(\mathbb{E}\sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} [L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*)]\right) + \delta} + \sqrt{\sum_{t=1}^{T} \beta_2^{T-t}}\sqrt{\sigma^2 + \delta}.$$

When taking δ to 0, we have that

$$\mathbb{E}\sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \left[\tilde{L}_t(\boldsymbol{x}_t) - \tilde{L}_t(\boldsymbol{x}^*)\right] \leq 2\sqrt{2}D\left[\sqrt{2H(\tilde{L},\mathcal{H})\left(\mathbb{E}\sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} [L_t(\boldsymbol{x}_t) - L_t(\boldsymbol{x}^*)]\right)} + \sqrt{\sum_{t=1}^{T} \beta_2^{T-t}}\sigma + \beta_2^{\frac{T}{2}} d\sqrt{\epsilon}\right]$$

and

$$\mathbb{E}\sum_{t=1}^{T} \beta_2^{\frac{T-t}{2}} \left[\tilde{L}_t(\boldsymbol{x}_t) - \tilde{L}_t(\boldsymbol{x}^*)\right] \leq 16D^2 H(\tilde{L},\mathcal{H}) + 4\sqrt{2\sum_{t=1}^{T} \beta_2^{T-t}} D\sigma + 4\sqrt{2}D\beta_2^{\frac{T}{2}} d\sqrt{\epsilon}.$$

If we choose $\bar{\boldsymbol{x}}_{1:T} = \frac{1}{\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}} \sum_{t=1}^T \beta_2^{\frac{T-t}{2}} \boldsymbol{x}_t$, then from the convexity of $\tilde{L}$ we know that

$$\begin{aligned}
\mathbb{E}[\tilde{L}(\bar{\boldsymbol{x}}_{1:T}) - \tilde{L}(\boldsymbol{x}^*)] &\le \mathbb{E}\frac{1}{\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}} \sum_{t=1}^T \beta_2^{\frac{T-t}{2}} \left[\tilde{L}_t(\boldsymbol{x}_t) - \tilde{L}_t(\boldsymbol{x}^*)\right] \\
&\le \frac{1}{\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}} \left(16D^2 H(\tilde{L}, \mathcal{H}) + 4\sqrt{2\sum_{t=1}^T \beta_2^{T-t}} D\sigma + 4\sqrt{2} D\beta_2^{\frac{T}{2}} d\sqrt{\epsilon}\right) \\
&\le \frac{16}{\sum_{t=1}^T \beta_2^{\frac{T-t}{2}}} D^2 H(\tilde{L}, \mathcal{H}) + \frac{4\sqrt{2}}{\sqrt{\sum_{t=1}^T \beta_2^{T-t}}} D\sigma + \frac{4\sqrt{2}}{\sum_{t=1}^T \beta_2^{\frac{-t}{2}}} Dd\sqrt{\epsilon}
\end{aligned}$$

This completes the proof. □

E. Proof for One-Sided Shampoo

E.1. Proof for left smoothness

Lemma E.1 (Left smoothness for one-sided Shampoo)**.** *Let $\mathcal{H} = \mathcal{S}_+^{d_L} \otimes \boldsymbol{I}_{d_R}$ be the well-structured preconditioner set for one-sided Shampoo. Then the $\mathcal{H}$-smoothness $H(L, \mathcal{H})$ defined in Definition 3.6 is equal to the smallest number $H \ge 0$ such that there exists $\boldsymbol{H}_{d_L}^* \in \mathbb{R}^{d_L \times d_L}$ satisfying that $H = d_R \operatorname{Tr}(\boldsymbol{H}_{d_L}^*)$ and that for any $\boldsymbol{X}, \boldsymbol{\Delta} \in \mathbb{R}^{d_L \times d_R}$,*

$$\left|\nabla^2 L(\boldsymbol{X})[\boldsymbol{\Delta}, \boldsymbol{\Delta}]\right| \le \left\langle \boldsymbol{H}_{d_L}^*, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \right\rangle.$$

In this case, the $\mathcal{H}$-smoothness is denoted by $H_{\text{left}}(L)$.

Proof of Lemma 4.3. First, for any $\boldsymbol{X}, \boldsymbol{\Delta} \in \mathbb{R}^{d_L \times d_R}$ and $\boldsymbol{x} = \overline{\text{vec}}(\boldsymbol{X})$, we have

$$\overline{\text{vec}}(\boldsymbol{\Delta})^\top \nabla^2 L(\boldsymbol{x}) \overline{\text{vec}}(\boldsymbol{\Delta}) = \nabla^2 L(\boldsymbol{X})[\boldsymbol{\Delta}, \boldsymbol{\Delta}].$$

Therefore, given the Kronecker product form of all $\boldsymbol{H} \in \mathcal{H}$, to find $\boldsymbol{H} \in \mathcal{H}$ with the smallest trace such that $-\boldsymbol{H} \preceq \nabla^2 L(\boldsymbol{x}) \preceq \boldsymbol{H}$, it is equivalent to find $\boldsymbol{H} = \boldsymbol{H}_{d_L} \otimes \boldsymbol{I}_{d_R} \in \mathcal{H}$ with the smallest trace such that for any $\boldsymbol{\Delta} \in \mathbb{R}^{d_L \times d_R}$,

$$\begin{aligned}
|\nabla^2 L(\boldsymbol{X})[\boldsymbol{\Delta}, \boldsymbol{\Delta}]| &\le \overline{\text{vec}}(\boldsymbol{\Delta})^\top (\boldsymbol{H}_{d_L} \otimes \boldsymbol{I}_{d_R}) \overline{\text{vec}}(\boldsymbol{\Delta}) \\
&= \operatorname{Tr}(\boldsymbol{\Delta}^\top \boldsymbol{H}_{d_L} \boldsymbol{\Delta}) \\
&= \langle \boldsymbol{H}_{d_L}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \rangle.
\end{aligned}$$

Further note that $\operatorname{Tr}(\boldsymbol{H}) = d_R \operatorname{Tr}(\boldsymbol{H}_{d_L})$, and thus we conclude that it is equivalent to find $\boldsymbol{H}_{d_L} \in \mathcal{S}_+^{d_L}$ with the smallest trace such that the above inequality holds for all $\boldsymbol{\Delta} \in \mathbb{R}^{d_L \times d_R}$. This completes the proof. □

E.2. Proof for regret bound

Theorem 4.2 (Regret bound for one-sided Shampoo)**.** *For convex functions $L_1, \ldots, L_T$, the regret of one-sided Shampoo (Algorithm 2) compared to any $\boldsymbol{X}^* \in \mathbb{R}^{d_L \times d_R}$ satisfies*

$$\sum_{t=1}^T L_t(\boldsymbol{X}_t) - \sum_{t=1}^T L_t(\boldsymbol{X}^*) \le \left(\frac{D_{\text{op}}^2}{2d_R\eta} + \eta\right)\left(G + d\sqrt{\epsilon}\right),$$

where $D_{\text{op}} = \max_{t\in[T]} \|\boldsymbol{X}_t - \boldsymbol{X}^*\|_{\text{op}}$ and $G = \sqrt{d_R} \operatorname{Tr}\left[\left(\sum_{t=1}^T \boldsymbol{G}_t \boldsymbol{G}_t^\top\right)^{\frac{1}{2}}\right]$. When the domain $\mathcal{X}$ is bounded in operator norm, i.e., $\|\mathcal{X}\|_{\text{op}} < \infty$, further choosing $\eta = \sqrt{2/d_R}\|\mathcal{X}\|_{\text{op}}$, it holds

$$\begin{aligned}
&\sum_{t=1}^T L_t(\boldsymbol{X}_t) - \sum_{t=1}^T L_t(\boldsymbol{X}^*) \\
\le& 2\sqrt{2}\|\mathcal{X}\|_{\text{op}} \left(\operatorname{Tr}\left[\left(\sum_{t=1}^T \boldsymbol{G}_t \boldsymbol{G}_t^\top\right)^{\frac{1}{2}}\right] + \frac{d}{\sqrt{d_R}}\sqrt{\epsilon}\right).
\end{aligned}$$

Proof of Theorem 4.2. We will apply Theorem 3.4 to one-sided Shampoo. According to the analysis in Appendix B.4, Algorithm 1 with $\mathcal{H} = \left(\mathbb{R}^{d_L \times d_L} \otimes \boldsymbol{I}_{d_R}\right) \cap \mathcal{S}_+^d$ recovers one-sided Shampoo. We can plug in $\|\boldsymbol{x}\|_{\mathcal{H}} = \frac{\|\boldsymbol{X}\|_{\text{op}}}{\sqrt{d_R}}$ and $\|\|\boldsymbol{g}_{1:T}\|\|_{\mathcal{H}} = \sqrt{d_R}\,\text{Tr}\left[\left(\sum_{t=1}^T \boldsymbol{G}_t\boldsymbol{G}_t^\top\right)^{\frac{1}{2}}\right]$ into Theorem 3.4 and get that

$$\sum_{t=1}^T L_t(\boldsymbol{X}_t) - \sum_{t=1}^T L_t(\boldsymbol{X}^*) \le \sqrt{2}\left(\frac{D_{\text{op}}^2}{2d_R\eta} + \eta\right)\left(G + \min\left(d\sqrt{\epsilon}, \frac{d^2\epsilon}{2G}\right)\right)$$

with $D_{\text{op}} = \max_{t\in[T]} \|\boldsymbol{X}_t - \boldsymbol{X}^*\|_{\text{op}}$ and $G = \sqrt{d_R}\,\text{Tr}\left[\left(\sum_{t=1}^T \boldsymbol{G}_t\boldsymbol{G}_t^\top\right)^{\frac{1}{2}}\right]$. □

F. Additional Results for Experiments

F.1. Efficient implementation of full-matrix AdaGrad.

Directly applying full-matrix AdaGrad to this 10^6-dimensional problem is impractical. Instead, we consider the eigendecomposition of $\boldsymbol{H}$ to be $\boldsymbol{U}^\top\boldsymbol{\Sigma}\boldsymbol{U}$ and define the transformation $\mathcal{T}(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{X}$. We further define d orthogonal matrices $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_d \in \mathbb{R}^{d\times d}$ such that the first row of $\boldsymbol{V}_i$ is in the same direction of $\mathcal{T}(\boldsymbol{X}^* - \boldsymbol{X}_0)_{i,:}$ and define $\boldsymbol{V} = \text{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_d)$. We can know that $\boldsymbol{V}\,\overline{\text{vec}}(\mathcal{T}(\boldsymbol{X}^* - \boldsymbol{X}_0) = \boldsymbol{v} \otimes \boldsymbol{e}_1$ where $\boldsymbol{v}_i = \|\mathcal{T}(\boldsymbol{X}^* - \boldsymbol{X}_0)_{i,:}\|_2$ for $i \in [d]$.

Then it holds that

$$\begin{aligned}
f(\boldsymbol{X}) = \left\langle \boldsymbol{H}, (\boldsymbol{X} - \boldsymbol{X}^*)(\boldsymbol{X} - \boldsymbol{X}^*)^\top \right\rangle &= \left\langle \boldsymbol{\Sigma}, (\boldsymbol{U}\boldsymbol{X} - \boldsymbol{U}\boldsymbol{X}^*)(\boldsymbol{U}\boldsymbol{X} - \boldsymbol{U}\boldsymbol{X}^*)^\top \right\rangle \\
&= \left\langle \boldsymbol{\Sigma}, (\mathcal{T}(\boldsymbol{X}) - \mathcal{T}(\boldsymbol{X}^*))(\mathcal{T}(\boldsymbol{X}) - \mathcal{T}(\boldsymbol{X}^*))^\top \right\rangle \\
&= \left\langle \boldsymbol{\Sigma} \otimes \boldsymbol{I}_d, \overline{\text{vec}}(\mathcal{T}(\boldsymbol{X} - \boldsymbol{X}^*)^\top)\,\overline{\text{vec}}(\mathcal{T}(\boldsymbol{X} - \boldsymbol{X}^*)^\top)^\top \right\rangle \\
&= \left\langle \boldsymbol{\Sigma} \otimes \boldsymbol{I}_d, \boldsymbol{V}\,\overline{\text{vec}}(\mathcal{T}(\boldsymbol{X} - \boldsymbol{X}^*)^\top)\,\overline{\text{vec}}(\mathcal{T}(\boldsymbol{X} - \boldsymbol{X}^*)^\top)^\top\boldsymbol{V}^\top \right\rangle
\end{aligned}$$

Further denoting $\tilde{\boldsymbol{x}} = \boldsymbol{V}\,\overline{\text{vec}}(\mathcal{T}(\boldsymbol{X})^\top)$ and $\tilde{\boldsymbol{y}} = \boldsymbol{V}\,\overline{\text{vec}}(\mathcal{T}(\boldsymbol{X}_0)^\top)$, then we obtain

$$\begin{aligned}
f(\boldsymbol{X}) &= \sum_{i=1}^d \sigma_i \left\|\tilde{\boldsymbol{x}}_{(i-1)d+1:id} - (\boldsymbol{v}\otimes\boldsymbol{e}_1)_{(i-1)d+1:id}\right\|_2^2 \\
&= \sum_{i=1}^d \sigma_i \left[(\tilde{\boldsymbol{x}}_{(i-1)d+1} - \tilde{\boldsymbol{y}}_{(i-1)d+1} - \boldsymbol{v}_i)^2 + \sum_{j=2}^d (\tilde{\boldsymbol{x}}_{(i-1)d+j} - \tilde{\boldsymbol{y}}_{(i-1)d+j})^2\right]
\end{aligned}$$

Running full-matrix AdaGrad on $f(\boldsymbol{X})$ starting from $\boldsymbol{X}_0$ can be implemented equivalently by using $\tilde{\boldsymbol{x}}$ as variable starting from $\tilde{\boldsymbol{y}}$. Only $\tilde{\boldsymbol{x}}_{(i-1)d+1}$ will receive non-zero gradient so full-matrix AdaGrad actually only cares these d coordinates, which reduces the original problem to a problem only with d variables.

F.2. Results for EMA algorithms

As mentioned in Section 5, we compare AdaSGD, Adam, one-sided EMA Shampoo and full-matrix AdaSGD, which are EMA version of AdaGrad-Norm, diagonal AdaGrad, one-sided Shampoo and full-matrix Adagrad. The results are plotted in Figure 2. We set $\beta_2 = 0.95$ and disable first-order momentum, i.e., $\beta_1 = 0$ in Adam.

We tried 60 learning rates between 1×10^{-4} and 1×10^2. The relationship between loss and learning rate is shown in Figure 3.

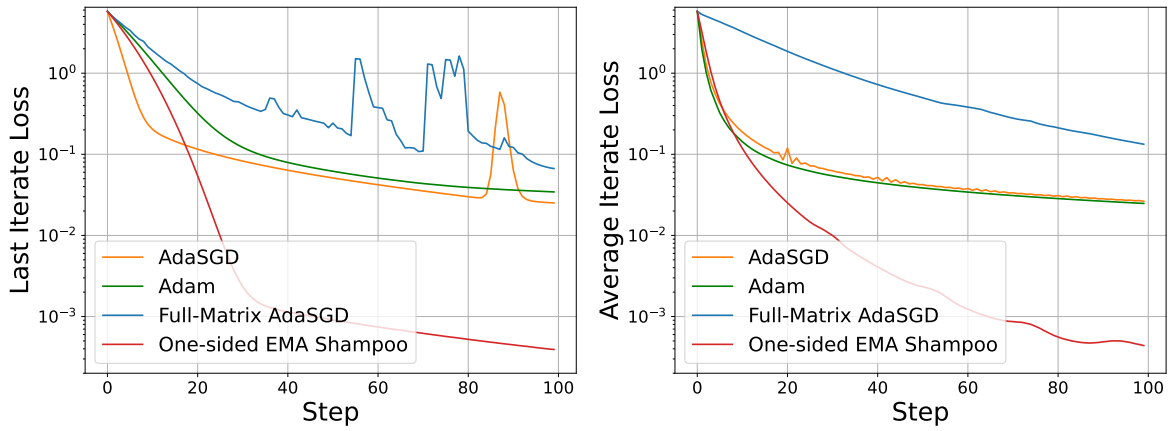

Figure 2. We plot the last iterate training loss $f(\boldsymbol{X}_t) = \left\langle \boldsymbol{H}, (\boldsymbol{X}_t - \boldsymbol{X}^*)(\boldsymbol{X}_t - \boldsymbol{X}^*)^\top \right\rangle$ and the average iterate training loss $f(\frac{1-\beta_2}{1-\beta_2^t}\sum_{s=1}^t \beta_2^{t-s}\boldsymbol{X}_s)$ over steps for optimizers obtained from Algorithm 5.

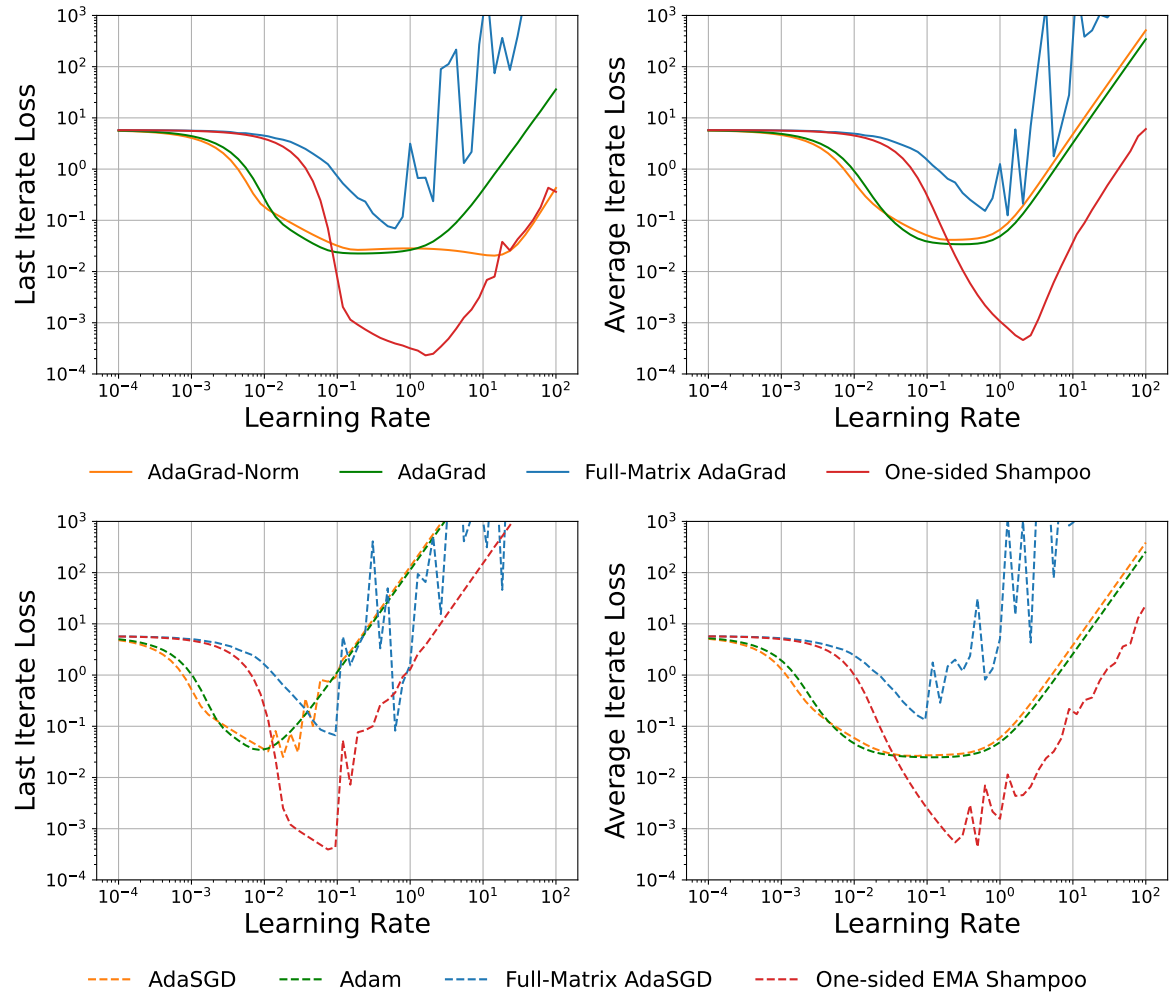

Figure 3. We plot the last iterate loss and average iterate loss versus learning rate. For each learning rate the plotted value is the average of last iterate loss and average of average iterate loss across five random seeds.