CONVERGENCE OF ROTATION-BASED MATRIX OPTIMIZERS: A UNIFIED ANALYSIS OF SOAP, CONDA, AND SPLUS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we develop a unified theoretical framework for analyzing the convergence of rotation-based matrix optimizers, which apply orthogonal transformations to map the momentum into a rotated space, perform coordinate-wise or normalized updates there, and then rotate the updates back. Our framework encompasses prominent matrix optimizers, including SOAP, Conda, and truncated SPlus, as well as matrix-parameterized Adam as a special case. As our main result, we establish, for the first time, the convergence rate of SOAP with sharp dimension dependence, as well as the convergence rate of Adam measured by the nuclear norm of the gradient. Our framework also covers a more general variant of rotation-based optimizers that permits arbitrary orthogonal rotation matrices, which may depend on the current stochastic gradient at each iteration. Technically, our framework leverages a row-column trace control argument that converts elementwise bounds into bounds on two diagonal control matrices, thereby improving the dimension dependence of the resulting bound.

1 INTRODUCTION

Pre-training large language models incurs substantial computational costs. Over the past decade, adaptive gradient methods, including AdaGrad (Duchi et al., 2011; McMahan & Streeter, 2010), RMSProp (Tieleman & Hinton, 2012), Adam (Kingma & Ba, 2015), and AdamW (Loshchilov & Hutter, 2019), have served as the de facto standard for optimizing deep neural networks. These methods treat the network's parameters as a single high-dimensional vector and do not explicitly exploit row/column correlations within matrices. More recently, matrix-based optimizers, which exploit the inherent matrix structure of deep neural network parameters, have garnered increasing attention and demonstrated the potential to outperform their vector-based counterparts. Prominent examples include Shampoo (Gupta et al., 2018), Muon (Jordan et al., 2024), SOAP (Vyas et al., 2025), SPlus (Frans et al., 2025), and Conda (Wang et al., 2025). The latter three approaches can be collectively classified as rotation-based optimizers, which rotate the gradient or momentum into a new space via an orthogonal transformation, produce an update in that space, and then rotate the update back to the original space. These methods differ mainly in how the rotation matrices and the transformed update are constructed: SOAP combines the eigenbasis of Shampoo's preconditioners, used as the rotation matrix, with an Adam-style update; SPlus replaces the Adam-style update with a sign-type update; and Conda employs the singular-vector basis of the momentum as the rotation matrix. A general algorithmic formulation encompassing most popular algorithms in this family is provided in (Gong et al., 2026). While Gong et al. (2026) provide a unified framework at the algorithmic level, our work provides a unified framework at the level of theoretical analysis.

Although the convergence behavior of vector-based optimizers has been thoroughly studied in the literature (Défossez et al., 2022; Shi et al., 2020; Li et al., 2025b; Zhang et al., 2022; Hong & Lin, 2024; Li et al., 2023; 2025a), the theoretical understanding of matrix-based optimizers remains comparatively underdeveloped. To date, convergence and convergence rates have been established for only a few matrix optimizers, such as Muon (Li & Hong, 2025; Kim & Oh, 2026; Shen et al., 2026; Chen et al., 2026; Sato et al., 2025) and Shampoo (Li et al., 2026; Xie et al., 2026). For rotation-based matrix optimizers, most notably SOAP, rigorous convergence analysis is still lacking. Relative to Muon and Shampoo, the main technical hurdles stem from two intrinsic properties of rotation-based

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

updates. First, it is non-trivial to combine the properties in the two spaces; moreover, the rotation matrices may depend on the current stochastic gradient. Second, coordinate-wise updates in the rotated space introduce an undesirable dependence on the number of matrix entries, mn.

1.1 CONTRIBUTIONS

To address the above issues, we treat SOAP, Conda, and SPlus in a unified framework with the following contributions:

1. We establish the following convergence rate
$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq \mathcal{O}\left(\sqrt{m+n}\max\left\{\sqrt[4]{\frac{\sigma^2 L\left(f(\mathbf{X}_1)-f^*\right)}{K}}, \sqrt{\frac{L\left(f(\mathbf{X}_1)-f^*\right)}{K}}\right\}\right)$$
for generalized SOAP, Conda and the truncated modification of SPlus, matching the state-of-the-art rate of Shampoo (Li et al., 2026) for matrix optimizers.
2. Our analysis extends to a considerably more general variant that imposes no structural restrictions on the rotation matrices beyond orthogonality. This flexibility allows the rotation matrices to be constructed adaptively from the available optimization history up to the current step (and thus may depend on the current stochastic gradient), covering those used in SOAP, Conda, and SPlus. Consequently, the theoretical guarantees we establish not only apply directly to these methods but also shed light on a broader class of rotation-based matrix optimizers, providing a unified analytical framework for understanding their convergence behavior.
3. In particular, our analysis also applies when the rotation matrices reduce to the identity matrix, in which case SOAP reduces to Adam and thus the above convergence rate also applies to Adam. By contrast, the existing literature measures the convergence rate of Adam only in the ℓ_1 or ℓ_2 norm; we provide the first nuclear-norm convergence rate for Adam with matrix-valued parameters, of the same order as that of matrix optimizers such as AdamW-style Shampoo (Li et al., 2026).
4. Technically, we propose a row-column trace control argument in which the control matrices are carefully designed, reducing the problem to rigorously bounding their traces and lowering the dimension dependence from $\sqrt{mn}$ to $\sqrt{m+n}$. To handle the difficulty posed by orthogonal transformations between the two spaces, we exploit unitary invariance, which holds for any unitarily invariant norm such as the nuclear norm and the Frobenius norm.

1.2 PROBLEM SETUP, NOTATION, AND ASSUMPTIONS

We study the following nonconvex problem with matrix-valued parameters in this paper

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X}), \tag{1}$$

where $f(\mathbf{X}) = \mathbb{E}_{\zeta\sim\mathcal{P}}[f(\mathbf{X};\zeta)]$ and ζ is a sample drawn from the data distribution $\mathcal{P}$. We assume that f is bounded below and denote $f^* = \inf_{\mathbf{X}} f(\mathbf{X})$.

We denote vectors by lowercase bold letters and matrices by uppercase bold letters. For vectors, denote $\|\cdot\|_2$ as the ℓ_2 Euclidean norm. For matrices, denote $\|\cdot\|_F$, $\|\cdot\|_{op}$, and $\|\cdot\|_*$ as the Frobenius norm, spectral norm (largest singular value), and nuclear norm (sum of singular values), respectively. Denote $\mathbf{X}_{k,i,j}$, $\mathbf{X}_{k,i,:}$, and $\mathbf{X}_{k,:,j}$ to be the (i,j)-th element, i-th row, and j-th column of matrix $\mathbf{X}$ at iteration k, respectively. For matrices $\mathbf{X}$ and $\mathbf{Y}$, denote $\frac{\mathbf{X}}{\mathbf{Y}}$, $\mathbf{X}^2$, $|\mathbf{X}|$, and $\mathbf{X}^{\frac{1}{p}}$ as the elementwise division, square, absolute value, and the $\frac{1}{p}$-th power (different from the matrix power, with a slight abuse of terminology), respectively. Denote $\mathcal{F}_k = \sigma(\zeta_1, \zeta_2, \cdots, \zeta_k)$ to be the sigma-algebra generated by the samples observed up to iteration k, denote $\mathbb{E}_{\mathcal{F}_k}[\cdot]$ as the expectation with respect to $\mathcal{F}_k$ and $\mathbb{E}_k[\cdot|\mathcal{F}_{k-1}]$ the conditional expectation with respect to ζ_k given $\mathcal{F}_{k-1}$. For the sake of brevity, $\mathbb{E}_{\mathcal{F}_K}[\cdot]$ will be denoted as $\mathbb{E}[\cdot]$.

We make the following assumptions throughout this paper:

Assumption 1 *Smoothness:* $\|\nabla f(\mathbf{Y}) - \nabla f(\mathbf{X})\|_F \leq L\|\mathbf{Y}-\mathbf{X}\|_F, \forall \mathbf{X}, \mathbf{Y}$.

Algorithm 1 General Rotation-Based Matrix Optimizer

Hyperparameters: η, θ.
Initialize $\mathbf{X}_1, \mathbf{M}_0 = \mathbf{0}$.
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{G}_k = \text{GradOracle}(\mathbf{X}_k)$
  $\mathbf{M}_k = \theta \mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k$
  Generate arbitrary orthogonal matrices $\mathbf{P}_k \in \mathbb{R}^{m\times m}$ and $\mathbf{Q}_k \in \mathbb{R}^{n\times n}$
  $\mathbf{X}_{k+1} = \mathbf{X}_k - \eta \mathbf{P}_k \mathbf{\Phi}_k \mathbf{Q}_k^\top$, where $\mathbf{\Phi}_k = \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k}$ for some $\mathbf{D}_k$
end for

Table 1: Representative Rotation-Based Optimizers with Their Rotation Matrices and Updates

Optimizers	Rotation Matrices	Update in the rotated space
SOAP	Eigenbasis of Shampoo's preconditioner	Adam-style update
SPlus	Eigenbasis of Shampoo's preconditioner	Sign-type update
Conda	Singular-vector basis of the momentum	Adam-style update
Our generalizations	Arbitrary orthogonal matrices	Adam-style or truncated sign-type update

Assumption 2 *Unbiased estimator:* $\mathbb{E}_k \left[\mathbf{G}_k \middle| \mathcal{F}_{k-1}\right] = \nabla f(\mathbf{X}_k)$.

Assumption 3 *Bounded noise variance:* $\mathbb{E}_k \left[\|\mathbf{G}_k - \nabla f(\mathbf{X}_k)\|_F^2 \middle| \mathcal{F}_{k-1}\right] \le \sigma^2$.

1.3 Related Work

Among matrix-based methods, the convergence properties of Muon have been extensively investigated in several recent works, owing to its relatively simple structure (Li & Hong, 2025; Shen et al., 2026; Chen et al., 2026; Sato et al., 2025). For more intricate optimizers, recent progress includes the work of Li et al. (2026), who established rigorous convergence guarantees for a practical AdamW-style implementation of Shampoo. Nevertheless, with the exception of these two cases, the convergence analysis of other matrix-based optimizers remains an open problem to the best of our knowledge, especially for the representative SOAP optimizer. Perhaps the most closely related study is that of He et al. (2025), which analyzes a variant of the GaLore algorithm (Zhao et al., 2024). However, their approach replaces the orthogonal rotation employed by SOAP and semi-orthogonal low-rank projection by GaLore with a random projection matrix and uses momentum stochastic gradient descent (MSGD) within the projected space, whereas SOAP and GaLore leverage the adaptive mechanisms of Adam. Consequently, the analytical techniques developed in (He et al., 2025) are far from sufficient to establish the convergence of the full SOAP algorithm, as well as of Conda, and SPlus. He et al. (2025) also constructed a counterexample demonstrating the non-convergence of GaLore due to its low-rank projection. Other works have studied SOAP from different perspectives (Lin et al., 2026; Lu et al., 2025), but none has provided a convergence guarantee.

2 Rotation-based Matrix Optimizers

We begin by examining the general framework of rotation-based matrix optimizers presented in Algorithm 1. Here $\mathbf{\Phi}_k$ denotes the update matrix in the rotated space, such as the Adam-style update in SOAP and Conda, and the sign-type update in SPlus. We require the update $\mathbf{\Phi}_k$ to admit the following representation for an elementwise positive matrix $\mathbf{D}_k$:

$$\mathbf{\Phi}_k = \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k},$$

where $\mathbf{D}_k$ can be regarded as a normalizer. We describe the choice of $\mathbf{D}_k$ for each algorithm individually in the following subsections. The orthogonal rotation matrices $\mathbf{P}_k$ and $\mathbf{Q}_k$ may be generated by any procedure, including periodic updates, and need not be independent of the current

Algorithm 2 SOAP (Vyas et al., 2025)

Hyperparameters: $\eta, \theta, \beta, \varepsilon, T$
Initialize $\mathbf{X}_1$, $\mathbf{M}_0$, $\mathbf{V}_0$, $\mathbf{L}_0$, $\mathbf{R}_0$, $\mathbf{P}_1$, $\mathbf{Q}_1$.
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{G}_k = \text{GradOracle}(\mathbf{X}_k)$
  $\mathbf{M}_k = \theta \mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k$
  $\mathbf{G}'_k = \mathbf{P}_k^\top \mathbf{G}_k \mathbf{Q}_k$
  $\mathbf{M}'_k = \mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k$
  $\mathbf{V}_k = \beta \mathbf{V}_{k-1} + (1-\beta)(\mathbf{G}'_k)^2$
  $\mathbf{X}_{k+1} = \mathbf{X}_k - \eta \mathbf{P}_k \frac{\mathbf{M}'_k}{\sqrt{\mathbf{V}_k + \varepsilon}} \mathbf{Q}_k^\top$
  $\mathbf{L}_k = \beta \mathbf{L}_{k-1} + (1-\beta)\mathbf{G}_k \mathbf{G}_k^\top$
  $\mathbf{R}_k = \beta \mathbf{R}_{k-1} + (1-\beta)\mathbf{G}_k^\top \mathbf{G}_k$
  if $k \mod T = 0$ **then**
    $\mathbf{P}_{k+1} = \text{Eigenvector}(\mathbf{L}_k)$
    $\mathbf{Q}_{k+1} = \text{Eigenvector}(\mathbf{R}_k)$
  else
    $\mathbf{P}_{k+1} = \mathbf{P}_k$, $\mathbf{Q}_{k+1} = \mathbf{Q}_k$
  end if
end for

Algorithm 3 SPlus (Frans et al., 2025)

Hyperparameters: $\eta, \theta, \beta, \kappa, T$
Initialize $\mathbf{X}_1$, $\mathbf{X}'_1$, $\mathbf{M}_0$, $\mathbf{L}_0$, $\mathbf{R}_0$, $\mathbf{P}_0$, $\mathbf{Q}_0$.
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{G}_k = \text{GradOracle}(\mathbf{X}'_k)$
  $\mathbf{M}_k = \theta \mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k$
  $\mathbf{L}_k = (1-\beta)\mathbf{L}_{k-1} + \beta \mathbf{G}_k \mathbf{G}_k^\top$
  $\mathbf{R}_k = (1-\beta)\mathbf{R}_{k-1} + \beta \mathbf{G}_k^\top \mathbf{G}_k$
  if $k \mod T = 0$ **then**
    $\mathbf{P}_k = \text{Eigenvector}(\mathbf{L}_k)$
    $\mathbf{Q}_k = \text{Eigenvector}(\mathbf{R}_k)$
  else
    $\mathbf{P}_k = \mathbf{P}_{k-1}$, $\mathbf{Q}_k = \mathbf{Q}_{k-1}$
  end if
  $\mathbf{X}'_{k+1} = \mathbf{X}'_k - \frac{2\eta}{m+n} \mathbf{P}_k \text{Sign}(\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k) \mathbf{Q}_k^\top$
  $\mathbf{X}_{k+1} = (1-\kappa)\mathbf{X}_k + \kappa \mathbf{X}'_{k+1}$
end for

Algorithm 4 Generalized SOAP

Hyperparameters: $\eta, \theta, \beta, \varepsilon$
Initialize $\mathbf{X}_1$, $\mathbf{M}_0 = \mathbf{0}$, $\mathbf{V}_0 = \mathbf{0}$.
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{G}_k = \text{GradOracle}(\mathbf{X}_k)$
  $\mathbf{M}_k = \theta \mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k$
  Generate arbitrary orthogonal matrices
  $\mathbf{P}_k \in \mathbb{R}^{m \times m}$ and $\mathbf{Q}_k \in \mathbb{R}^{n \times n}$
  $\mathbf{V}_k = \beta \mathbf{V}_{k-1} + (1-\beta)(\mathbf{P}_k^\top \mathbf{G}_k \mathbf{Q}_k)^2$
  $\mathbf{X}_{k+1} = \mathbf{X}_k - \eta \mathbf{P}_k \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\sqrt{\mathbf{V}_k + \varepsilon}} \mathbf{Q}_k^\top$
end for

Algorithm 5 Conda (Wang et al., 2025)

Hyperparameters: $\eta, \theta, \beta, \varepsilon$
Initialize $\mathbf{X}_1$, $\mathbf{M}_0 = \mathbf{0}$, $\mathbf{V}_0 = \mathbf{0}$.
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{G}_k = \text{GradOracle}(\mathbf{X}_k)$
  $\mathbf{M}_k = \theta \mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k$
  $\mathbf{P}_k, \Sigma_k, \mathbf{R}_k^\top = \text{SVD}(\mathbf{M}_k)$
  $\mathbf{V}_k = \beta \mathbf{V}_{k-1} + (1-\beta)(\mathbf{P}_k^\top \mathbf{G}_k)^2$
  $\mathbf{X}_{k+1} = \mathbf{X}_k - \eta \mathbf{P}_k \frac{\mathbf{P}_k^\top \mathbf{M}_k}{\sqrt{\mathbf{V}_k + \varepsilon}}$
end for

Algorithm 6 Generalized truncated SPlus

Hyperparameters: η, θ, δ
Initialize $\mathbf{X}_1$, $\mathbf{M}_0 = \mathbf{0}$.
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{G}_k = \text{GradOracle}(\mathbf{X}_k)$
  $\mathbf{M}_k = \theta \mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k$
  Generate orthogonal matrices
  $\mathbf{P}_k \in \mathbb{R}^{m \times m}$ and $\mathbf{Q}_k \in \mathbb{R}^{n \times n}$
  $\mathbf{X}_{k+1} = \mathbf{X}_k - \eta \mathbf{P}_k \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\max\{|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k|, \delta\}} \mathbf{Q}_k^\top$
end for

stochastic gradient at each iteration. We also provide examples of the generation process for each algorithm in the following subsections. The rotation matrices and the transformed update capture the main differences among these algorithms, thus allowing the framework to represent and include many popular algorithms. Table 1 lists the representative examples studied in this paper.

2.1 SOAP

We present the generalized SOAP algorithm in Algorithm 4, which includes the original SOAP (Algorithm 2) as a special case. In this case, the matrices $\mathbf{\Phi}_k$ and $\mathbf{D}_k$ in Algorithm 1 can be written as

$$\mathbf{\Phi}_k = \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\sqrt{\mathbf{V}_k + \varepsilon}}, \quad \mathbf{D}_k = \sqrt{\mathbf{V}_k + \varepsilon}. \tag{2}$$

The original SOAP periodically constructs orthogonal rotation matrices from the eigenbases of Shampoo's preconditioners $\mathbf{L}$ and $\mathbf{R}$, whereas the generalized form accommodates essentially arbitrary orthogonal matrices, including those used in SOAP as special cases. This enables new classes of rotation matrices that may offer superior practical performance.

2.2 CONDA

Conda is a special case of the generalized SOAP in Algorithm 4, obtained by choosing $\mathbf{P}_k \in \mathbb{R}^{m\times m}$ as the matrix of left singular vectors of $\mathbf{M}_k$ (suppose that $m \le n$), setting $\mathbf{Q}_k$ to the identity matrix, and using the same $\mathbf{\Phi}_k$ matrix as in SOAP. The relaxation of the constraints on $\mathbf{P}_k$ and $\mathbf{Q}_k$ in generalized SOAP, which allows them to depend on $\mathbf{G}_k$, is especially important here, since Conda constructs $\mathbf{P}_k$ from the current momentum, which depends on $\mathbf{G}_k$. This illustrates more explicitly why general rotation matrices are needed, enabling Conda to fit naturally within the same framework as SOAP.

2.3 SPLUS

SPlus adopts the same rotation matrices as SOAP, together with the sign-type update in the rotated space, as presented in Algorithm 3. In contrast to SOAP, however, its rotation matrices $\mathbf{P}_k$ and $\mathbf{Q}_k$ depend on the current stochastic gradient, whereas SOAP uses one-step-delayed counterparts. We further generalize SPlus to accommodate arbitrary orthogonal matrices in Algorithm 6. When the sign-type update is adopted, the matrices $\mathbf{\Phi}_k$ and $\mathbf{D}_k$ in Algorithm 1 are given by

$$\mathbf{\Phi}_k = \mathrm{Sign}\left(\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k\right) = \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k|}, \quad \mathbf{D}_k = |\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k|.$$

In our analysis, we assume $\mathbf{D}_{k,i,j} \ge \delta$. This assumption, however, does not always hold for the original sign update in SPlus. SOAP avoids this issue by introducing the stabilization parameter ε. We therefore focus on a truncated modification of the sign function in Algorithm 6:

$$\mathbf{\Phi}_k = \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k}, \qquad \mathbf{D}_{k,i,j} = \max\left\{\left|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k\right|_{i,j}, \delta\right\}. \tag{3}$$

We consider only the internal iterates and omit the output averaging step used in the original SPlus algorithm. A convergence guarantee for the averaged output would require an additional argument.

3 UNIFIED CONVERGENCE ANALYSIS FRAMEWORK

3.1 BOUNDING THE NORMALIZED ROTATED GRADIENT

We first establish unified properties for the general framework presented in Algorithm 1; the argument does not rely on any specific properties of $\mathbf{D}_k$, except that it is elementwise positive with a lower bound. Following the proofs in (Li et al., 2025a), we obtain the following lemma for Algorithm 1, which bounds the normalized rotated gradient in expectation.

Lemma 1 *Suppose that Assumptions 1-3 and condition $\mathbf{D}_{k,i,j} \ge \delta$ hold. Let $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ with any $\gamma \in (0,1]$, $1-\theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$, and $\eta = \sqrt{\frac{\delta^2(f(\mathbf{X}_1)-f^*)}{4LK\hat{\sigma}^2}}$. Then for the general Algorithm 1, we have*

$$\mathcal{S}_K := \sum_{k=1}^{K} \mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \le \frac{7}{\delta}\sqrt{K\hat{\sigma}^2 L\left(f(\mathbf{X}_1)-f^*\right)}.$$

In contrast to (Li et al., 2025a), we handle the difficulty posed by orthogonal transformations between the original space and rotated space by exploiting the property $\|\mathbf{P}^\top \mathbf{X}\mathbf{Q}\| = \|\mathbf{X}\|$ whenever $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal, for any unitarily invariant norm $\|\cdot\|$, as is the case for the nuclear norm and the Frobenius norm. In other words, the Frobenius and nuclear norms of $\mathbf{X}$ are preserved under orthogonal rotation from the original space to the rotated space.

3.2 ROW-COLUMN TRACE CONTROL

If we strictly follow the argument in (Li et al., 2025a), we obtain a convergence rate of order $\mathcal{O}(\frac{\sqrt{mn}}{K^{1/4}})$ measured by the nuclear norm of the gradient:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \le \mathcal{O}\left(\sqrt{mn}\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1)-f^*\right)}{K}}\right).$$

The dimension factor is larger by a factor of $\sqrt{\frac{mn}{m+n}}$ than the convergence rate of Shampoo (Li et al., 2026). Upon closer inspection, the problematic $\sqrt{mn}$ term stems from the use of an elementwise upper bound. This is not an issue in the analysis of Adam in (Li et al., 2025a), as that work focuses on vector-based optimizers. Matrix-based optimizers, by contrast, require a different analysis that exploits the matrix structure, especially SOAP, which employs an Adam-type update in the rotated space. To address this issue, we propose a row-column trace control argument that separately controls the row-wise and column-wise maxima. Namely, denote

$$\begin{aligned} b_{k,i} &\geq \max_{1\leq j\leq n} \mathbf{D}_{k,i,j}, \quad \mathbf{B}_k = \operatorname{diag}\left(b_{k,1}, b_{k,2}, \dots b_{k,m}\right), \\ c_{k,j} &\geq \max_{1\leq i\leq m} \mathbf{D}_{k,i,j}, \quad \mathbf{C}_k = \operatorname{diag}\left(c_{k,1}, c_{k,2} \cdots c_{k,n}\right). \end{aligned} \tag{4}$$

With the control matrices defined above, we obtain the following lemma. The idea is that introducing control matrices allows the desired convergence rate for the nuclear norm of the gradient to decompose into (i) a term $\mathcal{S}_K$, already bounded in Lemma 1, and (ii) the traces of the two control matrices, $\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]$ and $\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]$. Thus, we reduce the problem to bounding these traces for different optimizers.

Lemma 2 *For the general Algorithm 1 and $\mathbf{B}_k$ and $\mathbf{C}_k$ defined in (4), we have*

$$\left(\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2 \leq \frac{1}{2}\left(\sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right] + \sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]\right) \cdot \sum_{k=1}^{K} \mathbb{E}\left[\left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2\right]. \tag{5}$$

We explain the rationale for choosing the control matrices to be of the row-wise and column-wise maxima diagonal form. As proved in Appendix A, by employing unitary invariance of nuclear norm, Schatten-Hölder inequality, and Hölder's inequality, we obtain

$$\left(\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2 \leq \sqrt{\sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right] \cdot \sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]} \cdot \sum_{k=1}^{K} \mathbb{E}\left[\left\|\mathbf{B}_k^{-\frac{1}{4}} \mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k \mathbf{C}_k^{-\frac{1}{4}}\right\|_F^2\right].$$

Then, for the last term, we have

$$\begin{aligned} \left\|\mathbf{B}_k^{-\frac{1}{4}} \mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k \mathbf{C}_k^{-\frac{1}{4}}\right\|_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\frac{\left(\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right)_{i,j}}{\sqrt[4]{b_{k,i}c_{k,j}}}\right)^2 \\ &\leq \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right)_{i,j}^2 = \left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2, \end{aligned}$$

where we use the diagonal forms of $\mathbf{B}_k$ and $\mathbf{C}_k$, together with $b_{k,i} \geq \max_{1\leq j\leq n} \mathbf{D}_{k,i,j}$ and $c_{k,j} \geq \max_{1\leq i\leq m} \mathbf{D}_{k,i,j}$. We finally relate the nuclear norm of the gradient to the Frobenius norm of the normalized rotated gradient.

The reason introducing control matrices yields a sharper bound lies in the number of terms that must be controlled. In other words, the bound improves because row/column trace aggregation replaces mn coordinate-wise controls with $m + n$ trace controls, providing more refined control than elementwise bounding. From this point on, we examine several algorithms through the lens of this framework.

3.3 Algorithms with EMA-based Normalizers

We say that an algorithm has EMA-based normalizers if its associated matrix $\mathbf{D}_k$ necessarily incorporates an exponential moving average (EMA) of squared rotated gradients.

3.3.1 SOAP

SOAP is a canonical example of an EMA-based algorithm: By (2), $\mathbf{D}_k = \sqrt{\mathbf{V}_k + \varepsilon}$, which involves the EMA term $\mathbf{V}_k$. By exploiting the recursive update of $\mathbf{V}_k$, the following lemma bounds the traces of the control matrices designed for SOAP.

Lemma 3 *Suppose that Assumption 3 holds. Let* $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ *with any* $\gamma \in (0,1]$, $0 < \beta < 1$, *and* $\mathbf{V}_0 = \mathbf{0}$. *Then for generalized SOAP (Algorithm 4) with specific control matrices* $\mathbf{B}_k$ *and* $\mathbf{C}_k$, *we have*

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right] &\le K\left(\sqrt{2m}\hat{\sigma}+m\sqrt{\varepsilon}\right)+4\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\sqrt[4]{\mathbf{V}_k+\varepsilon}}\right\|_F^2\right],\\ \sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right] &\le K\left(\sqrt{2n}\hat{\sigma}+n\sqrt{\varepsilon}\right)+4\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\sqrt[4]{\mathbf{V}_k+\varepsilon}}\right\|_F^2\right].\end{aligned} \tag{6}$$

Design of the Control Matrices: We focus on constructing the row control matrix; the column control matrix is constructed analogously. Design the row controller as follows:

$$\phi_{k,i}=\max_{1\le j\le n}\left|\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j},\quad \varphi_{k,i}=\max_{1\le j\le n}\left|\mathbf{P}_k^\top\left(\mathbf{G}_k-\nabla f\left(\mathbf{X}_k\right)\right)\mathbf{Q}_k\right|_{i,j},$$
$$\bar{\alpha}_{k,i}=\beta\bar{\alpha}_{k-1,i}+2(1-\beta)\left(\phi_{k,i}^2+\varphi_{k,i}^2\right)\text{ with }\bar{\alpha}_{0,i}=0,\quad b_{k,i}=\sqrt{\bar{\alpha}_{k,i}+\varepsilon}.$$

Then from the recursion of $\mathbf{V}_k$ in Algorithm 4 and by induction, we have

$$\bar{\alpha}_{k,i}\ge\mathbf{V}_{k,i,j}\quad\text{and}\quad b_{k,i}\ge\sqrt{\mathbf{V}_{k,i,j}+\varepsilon}=\mathbf{D}_{k,i,j}\text{ for all }j=1,2,\cdots,n.$$

Thus, the coefficient $b_{k,i}$ satisfies (4) and it is a valid choice for the row controller. Next, we upper bound it. To separate the stochastic error contribution of $b_{k,i}$, we further define

$$\bar{\varphi}_{k,i}=\beta\bar{\varphi}_{k-1,i}+2(1-\beta)\varphi_{k,i}^2\text{ with }\bar{\varphi}_{0,i}=0,\quad n_{k,i}=\sqrt{\bar{\varphi}_{k,i}+\varepsilon}.$$

Then, it is possible to construct a recursive inequality

$$b_{k,i}-n_{k,i}\le\sqrt{\beta}(b_{k-1,i}-n_{k-1,i})+2(1-\beta)\frac{\phi_{k,i}^2}{b_{k,i}}.$$

This further gives an upper bound of $b_{k,i}$:

$$b_{k,i}\le n_{k,i}+2(1-\beta)\sum_{t=1}^{k}\beta^{\frac{k-t}{2}}\frac{\phi_{t,i}^2}{b_{t,i}}.$$

Summing over k and i and taking expectation, the $\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}[n_{k,i}]$ term is bounded above by $\sqrt{2Km\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}[\varphi_{k,i}^2]}+Km\sqrt{\varepsilon}$, while the latter $2(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k}\beta^{\frac{k-t}{2}}\mathbb{E}\left[\frac{\phi_{t,i}^2}{b_{t,i}}\right]$ is at most $4\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\frac{\phi_{k,i}^2}{b_{k,i}}\right]$. By employing

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\varphi_{k,i}^2\right] &\le\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}\left[\left(\mathbf{P}_k^\top\left(\mathbf{G}_k-\nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right)_{i,j}^2\right]\\ &=\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{P}_k^\top\left(\mathbf{G}_k-\nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right\|_F^2\right]=\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{G}_k-\nabla f(\mathbf{X}_k)\right\|_F^2\right]\le K\sigma^2\end{aligned}$$

and

$$\sum_{k=1}^{K}\sum_{i=1}^{m}\frac{\phi_{k,i}^2}{b_{k,i}}\le\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right)_{i,j}^2}{b_{k,i}}\le\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\sqrt[4]{\mathbf{V}_k+\varepsilon}}\right)_{i,j}^2,$$

we have the conclusion. ■

Combining the preceding lemmas yields the following theorem for the generalized SOAP.

Theorem 1 *Suppose that Assumptions 1-3 and condition* $\sqrt{\mathbf{V}_{k,i,j}+\varepsilon}\ge\delta$ *hold. Let* $\hat{\sigma}^2=\max\left\{\sigma^2,\frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ *with any* $\gamma\in(0,1]$, $1-\theta=\sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$, $0<\beta<1$, $\eta=\sqrt{\frac{\delta^2(f(\mathbf{X}_1)-f^*)}{4LK\hat{\sigma}^2}}$, *and* $\varepsilon=\frac{\tau\hat{\sigma}^2}{m+n}$, *where* τ *is a tuning factor with* $\tau\in(0,1]$. *Then for generalized SOAP (Algorithm 4), we have*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f(\mathbf{X}_k)\right\|_*\right]\le\left(2\sqrt{m+n}+16\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2L\left(f\left(\mathbf{X}_1\right)-f^*\right)}{K}}.$$

When $\delta = \sqrt{\varepsilon} = \sqrt{\frac{\tau}{m+n}}\hat{\sigma}$, the right-hand side of the above convergence rate becomes

$$\left(2 + \frac{16}{\sqrt{\tau}}\right)\sqrt{m+n} \cdot \sqrt[4]{\frac{\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}{K}}$$
$$= \left(2 + \frac{16}{\sqrt{\tau}}\right)\sqrt{m+n} \cdot \max\left\{\sqrt[4]{\frac{\sigma^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}{K}}, \sqrt{\frac{L\left(f\left(\mathbf{X}_1\right) - f^*\right)}{K\gamma}}\right\},$$

which is exactly the same order as Shampoo's convergence rate (Li et al., 2026). If we further set $\tau = \Theta(1)$, then, since $\|\nabla f(\mathbf{X}_k)\|_F \leq \|\nabla f(\mathbf{X}_k)\|_* \leq \sqrt{\min\{m,n\}}\|\nabla f(\mathbf{X}_k)\|_F$, this rate is comparable to the optimal convergence rate of SGD and conditionally matches the lower bound of stochastic nonconvex optimization (Arjevani et al., 2023) under the ideal condition $\|\nabla f(\mathbf{X}_k)\|_* = \Theta(\sqrt{\min\{m,n\}})\|\nabla f(\mathbf{X}_k)\|_F$ and balanced m and n.

Nevertheless, even for modern industrial large language models, the dimensions m and n are not particularly large. For example, in GPT-3 with 175 billion parameters, the QKV projection matrices have dimensions $m = n = 12288$, while the weight matrices in the feed-forward layer have dimensions $(m, n) = (12288, 49152)$ or $(49152, 12288)$. So $\frac{\hat{\sigma}^2}{m+n}$ may not be a very small constant. Consequently, our parameter $\varepsilon = \frac{\tau\hat{\sigma}^2}{m+n}$ with $\tau = \Theta(1)$ is set larger than what is typically used in practice. However, choosing τ small would slow the theoretical convergence rate.

On the other hand, whenever $\mathbf{V}_{k,i,j} = \Theta(\frac{\hat{\sigma}^2}{m+n})$, we have $\delta^2 = \Theta(\frac{\hat{\sigma}^2}{m+n})$ regardless of the value chosen for ε, and the convergence rate remains unaffected. This permits ε to be chosen much smaller than δ^2, consistent with practical settings.

Remark 1 *We can replace the row-column trace control argument with a one-sided trace control argument (row or column) along the smaller of the two dimensions, m and n, thereby reducing the dimension dependence of our convergence rate from $\sqrt{m+n}$ to $\sqrt{\min\{m,n\}}$, at the cost of larger parameter setting, $\varepsilon = \frac{\tau\hat{\sigma}^2}{\min\{m,n\}}$, particularly when $\min\{m,n\} = 1$. See Appendix D for the details.*

3.3.2 ADAM

Theorem 1 holds for arbitrary orthogonal matrices $\mathbf{P}_k$ and $\mathbf{Q}_k$. In particular, setting $\mathbf{P}_k = \mathbf{I}_m$ and $\mathbf{Q}_k = \mathbf{I}_n$ reduces SOAP to Adam when applied to problems whose parameters are matrices, so Theorem 1 also applies to Adam. This observation requires no separate convergence argument, as it follows directly from the general framework by choosing the identity matrices as the orthogonal bases. Thus, we establish a nuclear-norm convergence guarantee for Adam, whereas existing analyses of Adam under comparable assumptions typically report bounds in vector ℓ_1 norm (Li et al., 2025a) or ℓ_2 norm (Li et al., 2023). For example, when applied to matrix-valued parameters, the analysis in (Li et al., 2025a) establishes the following convergence rate for Adam

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_1\right] \leq \mathcal{O}\left(\sqrt{mn}\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)}{K}}\right).$$

In contrast, our new convergence rate is measured by the nuclear norm on the left-hand side and reduces $\sqrt{mn}$ to $\sqrt{m+n}$ on the right-hand side, at the cost of requiring a larger value of ε; Li et al. (2025a) set only $\varepsilon = \frac{\hat{\sigma}^2}{mn}$. This leads to a striking finding: Adam with the coordinate-wise update attains a nuclear-norm convergence rate of the same order as that of AdamW-style Shampoo (Li et al., 2026).

3.3.3 CONDA

Conda provides another useful illustration of our unified analysis framework: Theorem 1 applies directly to Conda with the choice of rotation matrices specified in Section 2.2. It shows that our framework has a broad range of applicability, spanning simple rotation matrices such as those in Conda as well as more complicated ones such as in SOAP.

3.4 Algorithms with Instantaneous normalizers

In contrast to the types of algorithms discussed above, there exist algorithms that do not require EMA in the normalizer $\mathbf{D}_k$. We refer to these as algorithms with instantaneous normalizers. They have the favorable property that helps alleviate the memory burden.

3.4.1 SPlus

The formal definition of $\mathbf{\Phi}_k$ and $\mathbf{D}_k$ for the truncated SPlus studied in this paper is shown in (3). Since no EMA values are involved for its normalizer, directly following the proofs of SOAP to form a recursion to design the control matrices becomes difficult, which forces us to take a different route (see Appendix C for more details). Nevertheless, we establish the convergence rate of the truncated SPlus in the following theorem.

Theorem 2 *Suppose that Assumptions 1-3 hold. Let* $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ *with any* $\gamma \in (0,1]$, $1-\theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$, $\eta = \sqrt{\frac{\delta^2(f(\mathbf{X}_1)-f^*)}{4LK\hat{\sigma}^2}}$, *and* $\delta = \frac{\tau\hat{\sigma}}{\sqrt{m+n}}$, *where* τ *is a tuning factor with* $\tau \in (0,1]$. *Then for generalized truncated SPlus (Algorithm 6), we have*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \le \left(4+\frac{11}{\tau}\right)\sqrt{m+n}\cdot\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right)-f^*\right)}{K}}.$$

Design of the Control Matrices: We only focus on the row control. Define

$$j_i^* = \underset{1\le j\le n}{\operatorname{argmax}} \left|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k\right|_{i,j}, \quad b_{k,i} = \left|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k\right|_{i,j_i^*} + \delta \ge \max_{1\le j\le n} \mathbf{D}_{k,i,j}$$

for (3). Therefore,

$$\begin{aligned} b_{k,i} &\le \left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j_i^*} + \left|\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right|_{i,j_i^*} + \delta \\ &\le \frac{\left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j_i^*}^2}{2b_{k,i}} + \frac{b_{k,i}}{2} + \left|\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right|_{i,j_i^*} + \delta \\ &\le \sum_{j=1}^{n} \frac{\left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j}^2}{2\mathbf{D}_{k,i,j}} + \frac{b_{k,i}}{2} + \left\|\left(\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right)_{i,:}\right\|_2 + \delta, \end{aligned}$$

which gives an upper bound of $b_{k,i}$. Similar to the discussion below Lemma 3, using

$$\sum_{k=1}^{K}\mathbb{E}\left[\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\|_F\right] \le \sqrt{13}K\hat{\sigma}$$

established in Lemma 8, we can finally obtain an upper bound on $\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]$, as shown in Lemma 9. ■

4 Conclusion

This paper provides a unified convergence analysis framework for rotation-based matrix optimizers, including SOAP, Conda, truncated SPlus, as well as matrix-parameterized Adam as a special case. This framework permits arbitrary orthogonal rotations that may depend on the current stochastic gradient and yield convergence bounds in terms of the nuclear norm of the gradient. Technically, our main approach leverages a row-column trace control argument that decomposes the control term into row-wise and column-wise maxima and converts elementwise bounds into bounds on two diagonal control matrices. The framework can be applied to other rotation-based optimizers whenever their normalizers admit a uniform positive lower bound and suitable row-column trace control bounds.

Limitations. Our theorems rely on parameter choices that depend on problem-dependent quantities, such as the initial optimality gap and the noise variance bound. Our analysis requires a uniform lower bound on $\mathbf{D}_k$. Although SOAP and the truncated SPlus satisfy this condition, it imposes an additional restriction on other rotation-based optimizers, such as the untruncated SPlus. The SPlus result pertains to its internal iterates; extending the analysis to averaged outputs remains open.

AI USE STATEMENT

In this work, we used generative AI tools for 'edit a research paper to improve readability' with recommended disclosure. We also used generative AI to review our paper before submitting and revised it accordingly. We have not used generative AI tools for the tasks with required disclosure. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Rajendra Bhatia (ed.). *Matrix Analysis*. Springer, New York, 1997.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and AdaGrad. *Transactions on Machine Learning Research*, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

Kevin Frans, Sergey Levine, and Pieter Abbeel. A stable whitening optimizer for efficient neural network training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wenbo Gong, Javier Zazo, Qijun Luo, Puqian Wang, James Hensman, and Chao Ma. ARO: A new lens on matrix optimization for large models. *arXiv:2602.09006*, 2026.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, 2018.

Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. In *International Conference on Machine Learning (ICML)*, 2025.

Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton-Schulz. In *International Conference on Learning Representations (ICLR)*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\frac{\sqrt{d}}{K^{1/4}})$ convergence rate of AdamW measured by ℓ_1 norm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\frac{\sqrt{d}}{T^{1/4}})$ convergence rate of RMSProp and its momentum extension measured by ℓ_1 norm. *Journal of Machine Learning Research*, 26(131): 1–25, 2025b.

Huan Li, Yiming Dong, and Zhouchen Lin. Convergence rate analysis of the AdamW-style Shampoo: Unifying one-sided and two-sided preconditioning. In *International Conference on Machine Learning (ICML)*, 2026.

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

Wu Lin, Scott C. Lowe, Felix Dangel, Runa Eschenhagen, Zikun Xu, and Roger Baker Grosse. Understanding and improving Shampoo and SOAP via Kullback-Leibler minimization. In *International Conference on Learning Representations (ICLR)*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Yanqing Lu, Letao Wang, and Jinbo Liu. Understanding SOAP from the perspective of gradient whitening. *arXiv:2509.22938*, 2025.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, 2010.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of Muon optimizer. *arXiv:2507.01598*, 2025.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSProp converges with proper hyperparameter. In *International Conference on Learning Representations (ICLR)*, 2020.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. In *COURSERA: Neural Networks for Machine Learning*, 2012.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam. In *International Conference on Learning Representations (ICLR)*, 2025.

Junjie Wang, Pan Zhou, Yiming Dong, Huan Li, Jia Li, Xun Zhou, Qicheng Lao, Cong Fang, and Zhouchen Lin. Conda: Column-normalized adam for training large language models faster. *arXiv:2509.24218*, 2025.

Shuo Xie, Tianhao Wang, Beining Wu, and Zhiyuan Li. A tale of two geometries: Adaptive optimizers and non-Euclidean descent. In *International Conference on Learning Representations (ICLR)*, 2026.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhiquan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning (ICML)*, 2024.

A PROOFS FOR THE GENERAL FRAMEWORK

We first state Hölder's inequality and Schatten-Hölder inequality, as well as two standard identities for Schatten norms.

Lemma 4 (Hölder's Inequality) *For all $x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n \geq 0$, if $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} y_i^q \right)^{\frac{1}{q}}.$$

Lemma 5 (Schatten-Hölder inequality) *(Bhatia, 1997) For $\mathbf{X}_i \in \mathbb{R}^{m_i \times m_{i+1}}, \forall i = 1, \ldots, n$, if $\sum_{i=1}^{n} \frac{1}{p_i} = 1$, we have*

$$\left\| \prod_{i=1}^{n} \mathbf{X}_i \right\|_{S_1} \leq \prod_{i=1}^{n} \|\mathbf{X}_i\|_{S_{p_i}},$$

where the Schatten-p norm is defined as

$$\|\mathbf{X}\|_{S_p} = \begin{cases} \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{X})^p \right)^{1/p}, & 1 \leq p < \infty, \\ \sigma_1(\mathbf{X}), & p = \infty. \end{cases}$$

Lemma 6 *For any diagonal matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ with positive entries, $\mathbf{Y} \in \mathbb{R}^{m \times n}$, and $p > 1$, we have*

$$\left\| \mathbf{X}^{\frac{1}{p}} \right\|_{S_p}^{p} = \operatorname{tr} \mathbf{X}, \qquad \|\mathbf{Y}\|_{S_2}^2 = \|\mathbf{Y}\|_F^2.$$

Proof 1 *Denote σ_i and λ_i as the i-th singular value and eigenvalue, respectively. Therefore $\sigma_i \left(\mathbf{X}^{\frac{1}{p}} \right) = \sigma_i \left(\mathbf{X} \right)^{\frac{1}{p}}$ and $\sigma_i(\mathbf{X}) = \lambda_i(\mathbf{X})$. So we have*

$$\left\| \mathbf{X}^{\frac{1}{p}} \right\|_{S_p}^{p} = \left(\sum_{i=1}^{n} \lambda_i \left(\mathbf{X} \right)^{\frac{1}{p} \cdot p} \right)^{\frac{1}{p} \cdot p} = \sum_{i=1}^{n} \lambda_i \left(\mathbf{X} \right) = \operatorname{tr} \mathbf{X},$$

and

$$\|\mathbf{Y}\|_{S_2}^2 = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i \left(\mathbf{Y} \right)^2 \right)^{\frac{1}{2} \cdot 2} = \operatorname{tr} \mathbf{Y}\mathbf{Y}^\top = \|\mathbf{Y}\|_F^2.$$

■

Remark 2 *Throughout this paper, we define $\mathbf{X}^{\frac{1}{p}}$ as the element-wise $\frac{1}{p}$-th power. To avoid having to define the matrix power, we prove the first equality only for diagonal $\mathbf{X}$, although it holds for any symmetric positive semi-definite matrix when the fractional power is interpreted as a matrix power.*

Definition 1 (Unitarily invariant norm) *(Bhatia, 1997) A norm is unitarily invariant if it satisfies $\|\mathbf{P}\mathbf{A}\mathbf{Q}^\top\| = \|\mathbf{A}\|$ for orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$. Both the Frobenius norm and nuclear norm are unitarily invariant norms.*

The proof of Lemma 1 presented below follows the technical framework of AdamW established in (Li et al., 2025a). The difference is that we extend the argument to accommodate orthogonal rotations, which is the main difficulty in analyzing rotation-based optimizers such as SOAP.

Proof 2 (Proof of Lemma 1) *As the gradient is L-Lipschitz, we have*

$$\begin{aligned}
& f(\mathbf{X}_{k+1}) - f(\mathbf{X}_k) \\
\leq & \langle \nabla f(\mathbf{X}_k), \mathbf{X}_{k+1} - \mathbf{X}_k \rangle + \frac{L}{2} \|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F^2 \\
= & -\eta \left\langle \nabla f(\mathbf{X}_k), \mathbf{P}_k \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k} \mathbf{Q}_k^\top \right\rangle + \frac{L\eta^2}{2} \left\| \mathbf{P}_k \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k} \mathbf{Q}_k^\top \right\|_F^2 \\
= & -\eta \left\langle \mathbf{P}_k^\top \nabla f(\mathbf{X}_k) \mathbf{Q}_k, \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k} \right\rangle + \frac{L\eta^2}{2} \left\| \mathbf{P}_k \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k} \mathbf{Q}_k^\top \right\|_F^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{a}{=} -\eta \left\langle \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}, \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\rangle + \frac{L\eta^2}{2} \left\| \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k} \right\|_F^2 \\
&\overset{b}{\leq} -\eta \left\langle \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}, \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\rangle + \frac{L\eta^2}{2\delta} \left\| \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 \\
&= -\frac{\eta}{2} \left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 - \frac{\eta}{2} \left\| \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 \\
&\quad + \frac{\eta}{2} \left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} - \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 + \frac{L\eta^2}{2\delta} \left\| \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 \\
&\overset{c}{\leq} -\frac{\eta}{2} \left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 - \frac{\eta}{4} \left\| \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 + \frac{\eta}{2\delta} \left\| \mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k - \mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k \right\|_F^2 \\
&\overset{d}{=} -\frac{\eta}{2} \left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 - \frac{\eta}{4} \left\| \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 + \frac{\eta}{2\delta} \left\| \nabla f(\mathbf{X}_k) - \mathbf{M}_k \right\|_F^2,
\end{aligned} \tag{7}
$$

where we use the notation that $\frac{\mathbf{X}}{\mathbf{Y}}$ and $\mathbf{X}^{\frac{1}{2}}$ are the elementwise division and square-root in $\overset{a}{=}$, $\overset{b}{\leq}$, and $\overset{c}{\leq}$, the fact that $\|\cdot\|_F$ is a unitarily invariant norm such that $\|\mathbf{PAQ}^\top\|_F = \|\mathbf{A}\|_F$ for orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$ in $\overset{a}{=}$ and $\overset{d}{=}$, the condition $\mathbf{D}_{k,i,j} \geq \delta$ in $\overset{b}{\leq}$ and $\overset{c}{\leq}$, and let $\eta \leq \frac{\delta}{2L}$ in $\overset{c}{\leq}$. Taking expectation on both sides of (7) with respect to ζ_k conditioned on $\mathcal{F}_{k-1}$, multiplying both sides of (11) by $\frac{\eta}{2\delta(1-\theta)}$, and adding to (7), we have

$$
\begin{aligned}
&\mathbb{E}_k \left[f(\mathbf{X}_{k+1}) - f^* + \frac{\eta\theta}{2\delta(1-\theta)} \left\| \nabla f(\mathbf{X}_k) - \mathbf{M}_k \right\|_F^2 + \frac{\eta}{4} \left\| \frac{\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 \middle| \mathcal{F}_{k-1} \right] \\
\leq & f(\mathbf{X}_k) - f^* - \frac{\eta}{2} \mathbb{E}_k \left[\left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 \middle| \mathcal{F}_{k-1} \right] \\
&+ \frac{\eta\theta}{2\delta(1-\theta)} \left\| \mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1}) \right\|_F^2 + \frac{L^2\eta^3}{2\delta^2(1-\theta)^2} \left\| \frac{\mathbf{P}_{k-1}^\top \mathbf{M}_{k-1} \mathbf{Q}_{k-1}}{\mathbf{D}_{k-1}^{\frac{1}{2}}} \right\|_F^2 + \frac{\eta(1-\theta)\sigma^2}{2\delta} \\
\leq & f(\mathbf{X}_k) - f^* - \frac{\eta}{2} \mathbb{E}_k \left[\left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 \middle| \mathcal{F}_{k-1} \right] \\
&+ \frac{\eta\theta}{2\delta(1-\theta)} \left\| \mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1}) \right\|_F^2 + \frac{\eta}{4} \left\| \frac{\mathbf{P}_{k-1}^\top \mathbf{M}_{k-1} \mathbf{Q}_{k-1}}{\mathbf{D}_{k-1}^{\frac{1}{2}}} \right\|_F^2 + \frac{\eta(1-\theta)\sigma^2}{2\delta},
\end{aligned} \tag{8}
$$

where we let $\eta^2 \leq \frac{\delta^2(1-\theta)^2}{2L^2}$ in the last inequality. Taking expectation with respect to $\mathcal{F}_{k-1}$ and summing (7) with $k = 1$ and (8) over $k = 2, 3, \cdots, K$, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{F}_K} \left[f(\mathbf{X}_{K+1}) - f^* + \frac{\eta\theta}{2\delta(1-\theta)} \left\| \nabla f(\mathbf{X}_K) - \mathbf{M}_K \right\|_F^2 + \frac{\eta}{4} \left\| \frac{\mathbf{P}_K^\top \mathbf{M}_K \mathbf{Q}_K}{\mathbf{D}_K^{\frac{1}{2}}} \right\|_F^2 \right] \\
\leq & f(\mathbf{X}_1) - f^* - \frac{\eta}{2} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k} \left[\left\| \frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}} \right\|_F^2 \right] \\
&+ \left(\frac{\eta}{2\delta} + \frac{\eta\theta}{2\delta(1-\theta)} \right) \mathbb{E}_{\mathcal{F}_1} \left[\left\| \mathbf{M}_1 - \nabla f(\mathbf{X}_1) \right\|_F^2 \right] + \frac{(K-1)\eta(1-\theta)\sigma^2}{2\delta}
\end{aligned}
$$

$$
\begin{aligned}
=&f(\mathbf{X}_1) - f^* - \frac{\eta}{2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \\
&+ \frac{\eta}{2\delta(1-\theta)}\mathbb{E}_{\mathcal{F}_1}\left[\left\|\mathbf{M}_1 - \nabla f(\mathbf{X}_1)\right\|_F^2\right] + \frac{(K-1)\eta(1-\theta)\sigma^2}{2\delta}.
\end{aligned} \tag{9}
$$

As the gradient is L-Lipschitz, we have

$$
\begin{aligned}
f^* &\le f\left(\mathbf{X} - \frac{1}{L}\nabla f(\mathbf{X})\right) \le f(\mathbf{X}) - \frac{1}{L}\left\langle \nabla f(\mathbf{X}), \nabla f(\mathbf{X})\right\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(\mathbf{X})\right\|_F^2 \\
&= f(\mathbf{X}) - \frac{1}{2L}\left\|\nabla f(\mathbf{X})\right\|_F^2.
\end{aligned}
$$

Using the recursion of $\mathbf{M}_1$ *and* $\mathbf{M}_0 = \mathbf{0}$*, as well as Assumptions 2-3, we have*

$$
\begin{aligned}
\mathbb{E}_{\mathcal{F}_1}\left[\left\|\nabla f(\mathbf{X}_1) - \mathbf{M}_1\right\|_F^2\right] =&\mathbb{E}_{\mathcal{F}_1}\left[\left\|\theta\nabla f(\mathbf{X}_1) + (1-\theta)\left(\nabla f(\mathbf{X}_1) - \mathbf{G}_1\right)\right\|_F^2\right] \\
=&\theta^2\left\|\nabla f(\mathbf{X}_1)\right\|_F^2 + (1-\theta)^2\mathbb{E}_{\mathcal{F}_1}\left[\left\|\nabla f(\mathbf{X}_1) - \mathbf{G}_1\right\|_F^2\right] \\
\le&2L\left(f(\mathbf{X}_1) - f^*\right) + (1-\theta)^2\sigma^2.
\end{aligned} \tag{10}
$$

Plugging into (9), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{F}_K}\left[f(\mathbf{X}_{K+1}) - f^* + \frac{\eta\theta}{2\delta(1-\theta)}\left\|\nabla f(\mathbf{X}_K) - \mathbf{M}_K\right\|_F^2 + \frac{\eta}{4}\left\|\frac{\mathbf{P}_K^\top \mathbf{M}_K \mathbf{Q}_K}{\mathbf{D}_K^{\frac{1}{2}}}\right\|_F^2\right] \\
\le& f(\mathbf{X}_1) - f^* - \frac{\eta}{2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + \frac{L\eta}{\delta(1-\theta)}\left(f(\mathbf{X}_1) - f^*\right) + \frac{K\eta(1-\theta)\sigma^2}{2\delta} \\
\le& f(\mathbf{X}_1) - f^* - \frac{\eta}{2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + \frac{L\eta}{\delta(1-\theta)}\left(f(\mathbf{X}_1) - f^*\right) + \frac{K\eta(1-\theta)\hat{\sigma}^2}{2\delta},
\end{aligned}
$$

where we denote $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ *with any* $\gamma \in (0,1]$*. So we have*

$$
\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \le&\frac{2}{\eta}\left(f(\mathbf{X}_1) - f^*\right) + \frac{2L}{\delta(1-\theta)}\left(f(\mathbf{X}_1) - f^*\right) + \frac{K(1-\theta)\hat{\sigma}^2}{\delta} \\
=&\frac{7}{\delta}\sqrt{K\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)}
\end{aligned}
$$

by letting $1-\theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$ *and* $\eta = \sqrt{\frac{\delta^2(f(\mathbf{X}_1)-f^*)}{4LK\hat{\sigma}^2}}$*, which also satisfies* $\eta^2 = \frac{\delta^2(1-\theta)^2}{4L^2} \le \frac{\delta^2(1-\theta)^2}{2L^2}$ *and* $\eta \le \frac{\delta}{2L}$*.* ■

We present the proof of Lemma 2 below, which follows the technical framework of AdamW-style Shampoo established in (Li et al., 2026). The difference is that we introduce the control matrices $\mathbf{B}_k$ and $\mathbf{C}_k$ and convert $\nabla f(\mathbf{X}_k)$ into $\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}$.

Proof 3 (Proof of Lemma 2) *Recall that we define*

$$
\begin{aligned}
b_{k,i} \ge \max_{1\le j\le n}\mathbf{D}_{k,i,j}, \quad \mathbf{B}_k = \operatorname{diag}\left(b_{k,1}, b_{k,2}, \dots b_{k,m}\right), \\
c_{k,j} \ge \max_{1\le i\le m}\mathbf{D}_{k,i,j}, \quad \mathbf{C}_k = \operatorname{diag}\left(c_{k,1}, c_{k,2}\cdots c_{k,n}\right).
\end{aligned}
$$

So we have

$$\begin{aligned}
&\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2 \overset{a}{=} \left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right\|_*\right]\right)^2\\
\overset{b}{\leq}&\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{\frac{1}{4}}\right\|_{S_4}\left\|\mathbf{B}_k^{-\frac{1}{4}}\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\mathbf{C}_k^{-\frac{1}{4}}\right\|_{S_2}\left\|\mathbf{C}_k^{\frac{1}{4}}\right\|_{S_4}\right]\right)^2\\
\overset{c}{\leq}&\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{\frac{1}{4}}\right\|_{S_4}^4\right]\right)^{\frac{1}{2}}\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{-\frac{1}{4}}\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\mathbf{C}_k^{-\frac{1}{4}}\right\|_{S_2}^2\right]\cdot\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{C}_k^{\frac{1}{4}}\right\|_{S_4}^4\right]\right)^{\frac{1}{2}}\\
\overset{d}{=}&\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]}\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{-\frac{1}{4}}\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\mathbf{C}_k^{-\frac{1}{4}}\right\|_F^2\right]\\
=&\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]}\cdot\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\left(\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right)_{i,j}}{\sqrt[4]{b_{k,i}c_{k,j}}}\right)^2\right]\\
\overset{e}{\leq}&\frac{1}{2}\left(\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]+\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]\right)\cdot\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right)_{i,j}^2\right]\\
=&\frac{1}{2}\left(\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]+\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]\right)\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right],
\end{aligned}$$

where we use the fact that $\|\cdot\|_*$ is a unitarily invariant norm such that $\|\mathbf{P}^\top\mathbf{A}\mathbf{Q}\|_* = \|\mathbf{A}\|_*$ for orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$ in $\overset{a}{=}$, the Schatten-Hölder inequality and $\|\mathbf{X}\|_* = \|\mathbf{X}\|_{S_1}$ in $\overset{b}{\leq}$, Hölder's inequality in $\overset{c}{\leq}$, Lemma 6 in $\overset{d}{=}$, and $b_{k,i}c_{k,j} \geq \mathbf{D}_{k,i,j}^2$ in $\overset{e}{\leq}$. ■

The following lemma is similar to the corresponding lemma in (Li et al., 2025a) and we list the proof here only for the sake of completeness.

Lemma 7 *Suppose that Assumptions 1-3 and condition $\mathbf{D}_{k,i,j} \geq \delta$ hold. Then for Algorithm 1, for $k \geq 2$, we have*

$$\begin{aligned}
&\mathbb{E}_k\left[\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\|_F^2\,\middle|\mathcal{F}_{k-1}\right]\\
\leq&\theta\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\|_F^2 + \frac{L^2\eta^2}{\delta(1-\theta)}\left\|\frac{\mathbf{P}_{k-1}^\top\mathbf{M}_{k-1}\mathbf{Q}_{k-1}}{\mathbf{D}_{k-1}^{\frac{1}{2}}}\right\|_F^2 + (1-\theta)^2\sigma^2.
\end{aligned}\tag{11}$$

Proof 4 *Denoting $\Gamma_k = \mathbf{G}_k - \nabla f(\mathbf{X}_k)$, we have $\mathbb{E}_k\left[\Gamma_k\middle|\mathcal{F}_{k-1}\right] = 0$ and $\mathbb{E}_k\left[\|\Gamma_k\|_F^2\,\middle|\mathcal{F}_{k-1}\right] \leq \sigma^2$. From the update of $\mathbf{M}_k$, we have*

$$\begin{aligned}
\mathbf{M}_k - \nabla f(\mathbf{X}_k) =&\theta\mathbf{M}_{k-1} + (1-\theta)\mathbf{G}_k - \nabla f(\mathbf{X}_k)\\
=&\theta\left(\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right) + (1-\theta)\left(\nabla f(\mathbf{X}_k) + \Gamma_k\right) - \nabla f(\mathbf{X}_k) + \theta\nabla f(\mathbf{X}_{k-1})\\
=&\theta\left(\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right) + (1-\theta)\Gamma_k - \theta\left(\nabla f(\mathbf{X}_k) - \nabla f(\mathbf{X}_{k-1})\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_k\left[\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\|_F^2\,\middle|\mathcal{F}_{k-1}\right]\\
=&\|\theta\left(\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right) - \theta\left(\nabla f(\mathbf{X}_k) - \nabla f(\mathbf{X}_{k-1})\right)\|_F^2 + (1-\theta)^2\mathbb{E}_k\left[\|\Gamma_k\|_F^2\,\middle|\mathcal{F}_{k-1}\right]\\
\leq&\theta^2\left(1 + \frac{1-\theta}{\theta}\right)\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\|_F^2 + \theta^2\left(1 + \frac{\theta}{1-\theta}\right)\|\nabla f(\mathbf{X}_k) - \nabla f(\mathbf{X}_{k-1})\|_F^2\\
&+ (1-\theta)^2\mathbb{E}_k\left[\|\Gamma_k\|_F^2\,\middle|\mathcal{F}_{k-1}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \theta \left\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right\|_F^2 + \frac{1}{1-\theta}\left\|\nabla f(\mathbf{X}_k) - \nabla f(\mathbf{X}_{k-1})\right\|_F^2 + (1-\theta)^2\sigma^2 \\
&\leq \theta \left\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right\|_F^2 + \frac{L^2}{1-\theta}\left\|\mathbf{X}_k - \mathbf{X}_{k-1}\right\|_F^2 + (1-\theta)^2\sigma^2 \\
&= \theta \left\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right\|_F^2 + \frac{L^2\eta^2}{1-\theta}\left\|\mathbf{P}_{k-1}\frac{\mathbf{P}_{k-1}^\top \mathbf{M}_{k-1}\mathbf{Q}_{k-1}}{\mathbf{D}_{k-1}}\mathbf{Q}_{k-1}^\top\right\|_F^2 + (1-\theta)^2\sigma^2 \\
&\overset{a}{=} \theta \left\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right\|_F^2 + \frac{L^2\eta^2}{1-\theta}\left\|\frac{\mathbf{P}_{k-1}^\top \mathbf{M}_{k-1}\mathbf{Q}_{k-1}}{\mathbf{D}_{k-1}}\right\|_F^2 + (1-\theta)^2\sigma^2 \\
&\overset{b}{\leq} \theta \left\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\right\|_F^2 + \frac{L^2\eta^2}{\delta(1-\theta)}\left\|\frac{\mathbf{P}_{k-1}^\top \mathbf{M}_{k-1}\mathbf{Q}_{k-1}}{\mathbf{D}_{k-1}^{\frac{1}{2}}}\right\|_F^2 + (1-\theta)^2\sigma^2,
\end{aligned}
$$

where we use the fact that $\|\cdot\|_F$ is a unitarily invariant norm such that $\|\mathbf{PAQ}^\top\|_F = \|\mathbf{A}\|_F$ for orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$ in $\overset{a}{=}$, the notation that $\frac{\mathbf{X}}{\mathbf{Y}}$ and $\mathbf{X}^{\frac{1}{2}}$ are the elementwise division and square-root and condition $\mathbf{D}_{k,i,j} \geq \delta$ in $\overset{b}{\leq}$. Specially, (11) holds for $\theta = 0$ trivially. ■

B PROOFS FOR SOAP

We now give the proof of Lemma 3. Instead of directly following the proof in (Li et al., 2025a), which makes use of an $\mathcal{F}_{k-1}$-measurable upper bound on $\widetilde{\mathbf{V}}_k$, we separate the main recursion term from the stochastic error term. This demonstrates how the error term affects the main term in a more straightforward way, and also provides an explicit recursion for both terms.

Proof 5 (Proof of Lemma 3) *Consider $\mathbf{B}_k$. Denote*

$$
\begin{aligned}
\phi_{k,i} &= \max_{1\leq j\leq n}\left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j}, \\
\varphi_{k,i} &= \max_{1\leq j\leq n}\left|\mathbf{P}_k^\top \left(\mathbf{G}_k - \nabla f\left(\mathbf{X}_k\right)\right)\mathbf{Q}_k\right|_{i,j}.
\end{aligned}
$$

In order to upper bound $\mathbf{V}_{k,i,j}$, define

$$
\begin{aligned}
\bar{\alpha}_{k,i} &= \beta\bar{\alpha}_{k-1,i} + 2(1-\beta)\left(\phi_{k,i}^2 + \varphi_{k,i}^2\right), \\
b_{k,i} &= \sqrt{\bar{\alpha}_{k,i} + \varepsilon},
\end{aligned}
$$

with $\bar{\alpha}_{0,i} = 0$. Therefore,

$$
\begin{aligned}
\left(\mathbf{P}_k^\top \mathbf{G}_k \mathbf{Q}_k\right)_{i,j}^2 &= \left(\mathbf{P}_k^\top \nabla f\left(\mathbf{X}_k\right)\mathbf{Q}_k + \mathbf{P}_k^\top\left(\mathbf{G}_k - \nabla f\left(\mathbf{X}_k\right)\right)\mathbf{Q}_k\right)_{i,j}^2 \\
&\leq 2\left(\left(\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right)_{i,j}^2 + \left(\mathbf{P}_k^\top\left(\mathbf{G}_k - \nabla f\left(\mathbf{X}_k\right)\right)\mathbf{Q}_k\right)_{i,j}^2\right) \\
&\leq 2\left(\phi_{k,i}^2 + \varphi_{k,i}^2\right).
\end{aligned}
$$

Compared to the definition of $\mathbf{V}_{k,i,j}$, by induction and $\mathbf{V}_{0,i,j} = \bar{\alpha}_{0,i} = 0$, we obtain,

$$
\bar{\alpha}_{k,i} \geq \mathbf{V}_{k,i,j} \quad \textit{and} \quad b_{k,i} \geq \sqrt{\mathbf{V}_{k,i,j} + \varepsilon} \quad \textit{for all } j = 1, 2, \cdots, n, \tag{12}
$$

Thus, the coefficients $b_{k,i}$ satisfy (4) since $\mathbf{D}_k = \sqrt{\mathbf{V}_k + \varepsilon}$ for SOAP. Thus, $b_{k,i}$ is a valid choice for the trace control process in the general framework. Then we separate the stochastic error contribution by recording the average error term:

$$
\begin{aligned}
\bar{\varphi}_{k,i} &= \beta\bar{\varphi}_{k-1,i} + 2(1-\beta)\varphi_{k,i}^2, \\
n_{k,i} &= \sqrt{\bar{\varphi}_{k,i} + \varepsilon},
\end{aligned}
$$

with $\bar{\varphi}_{0,i} = 0$. Therefore,

$$\begin{aligned} b_{k,i} - n_{k,i} &= \frac{\bar{\alpha}_{k,i} - \bar{\varphi}_{k,i}}{b_{k,i} + n_{k,i}} \\ &= \beta \frac{\bar{\alpha}_{k-1,i} - \bar{\varphi}_{k-1,i}}{b_{k,i} + n_{k,i}} + 2(1-\beta)\frac{\phi_{k,i}^2}{b_{k,i} + n_{k,i}} \\ &\overset{a}{\leq} \sqrt{\beta} \frac{\bar{\alpha}_{k-1,i} - \bar{\varphi}_{k-1,i}}{b_{k-1,i} + n_{k-1,i}} + 2(1-\beta)\frac{\phi_{k,i}^2}{b_{k,i}} \\ &= \sqrt{\beta}(b_{k-1,i} - n_{k-1,i}) + 2(1-\beta)\frac{\phi_{k,i}^2}{b_{k,i}}, \end{aligned}$$

where we use $b_{k,i}^2 = \bar{\alpha}_{k,i} + \varepsilon \geq \beta \bar{\alpha}_{k-1,i} + \beta\varepsilon = \beta b_{k-1,i}^2$ and a similar inequality for $n_{k,i}$ in $\overset{a}{\leq}$. Using this recursion and $b_{0,i} = n_{0,i} = \sqrt{\varepsilon}$, we have

$$b_{k,i} \leq n_{k,i} + 2(1-\beta)\sum_{t=1}^{k} \beta^{\frac{k-t}{2}} \frac{\phi_{t,i}^2}{b_{t,i}},$$

and in conclusion,

$$\begin{aligned} \sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr} \mathbf{B}_k\right] &= \sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[b_{k,i}\right] \\ &\leq \sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[n_{k,i}\right] + 2(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k} \beta^{\frac{k-t}{2}} \mathbb{E}\left[\frac{\phi_{t,i}^2}{b_{t,i}}\right] \\ &= \sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[\sqrt{\bar{\varphi}_{k,i} + \varepsilon}\right] + 2(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k} \beta^{\frac{k-t}{2}} \mathbb{E}\left[\frac{\phi_{t,i}^2}{b_{t,i}}\right] \\ &\overset{b}{\leq} \sqrt{Km}\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[\bar{\varphi}_{k,i}\right] + Km\varepsilon} + 2(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k} \beta^{\frac{k-t}{2}} \mathbb{E}\left[\frac{\phi_{t,i}^2}{b_{t,i}}\right] \\ &= \sqrt{Km}\sqrt{2(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k} \beta^{k-t}\mathbb{E}\left[\varphi_{t,i}^2\right] + Km\varepsilon} + 2(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k} \beta^{\frac{k-t}{2}} \mathbb{E}\left[\frac{\phi_{t,i}^2}{b_{t,i}}\right] \\ &\overset{c}{\leq} \sqrt{Km}\sqrt{2\sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[\varphi_{k,i}^2\right] + Km\varepsilon} + 4\sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[\frac{\phi_{k,i}^2}{b_{k,i}}\right] \\ &\overset{d}{\leq} \sqrt{2Km}\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[\varphi_{k,i}^2\right]} + Km\sqrt{\varepsilon} + 4\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\sqrt[4]{\mathbf{V}_k + \varepsilon}}\right)_{i,j}^2\right] \\ &\overset{e}{\leq} K\sqrt{2m}\sigma + Km\sqrt{\varepsilon} + 4\sum_{k=1}^{K} \mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\sqrt[4]{\mathbf{V}_k + \varepsilon}}\right\|_F^2\right], \end{aligned}$$

where $\overset{b}{\leq}$ uses the concavity of $\sqrt{x}$, $\overset{c}{\leq}$ uses the fact that

$$(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k} \beta^{k-t}\mathbb{E}\left[\varphi_{t,i}^2\right] = (1-\beta)\sum_{t=1}^{K}\sum_{i=1}^{m}\left(\mathbb{E}\left[\varphi_{t,i}^2\right]\sum_{k=t}^{K}\beta^{k-t}\right) \leq \sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\varphi_{k,i}^2\right],$$

as well as,

$$(1-\beta)\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{t=1}^{k} \beta^{\frac{k-t}{2}}\mathbb{E}\left[\frac{\phi_{t,i}^2}{b_{t,i}}\right] = (1-\beta)\sum_{t=1}^{K}\sum_{i=1}^{m}\left(\mathbb{E}\left[\frac{\phi_{t,i}^2}{b_{t,i}}\right]\sum_{k=t}^{K}\beta^{\frac{k-t}{2}}\right) \leq 2\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\frac{\phi_{k,i}^2}{b_{k,i}}\right].$$

In $\overset{d}{\leq}$, the first term uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and the second term follows

$$\frac{\phi_{k,i}^2}{b_{k,i}} \leq \sum_{j=1}^{n} \frac{\left(\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right)_{i,j}^2}{b_{k,i}} \leq \sum_{j=1}^{n} \left(\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\sqrt[4]{\mathbf{V}_k + \varepsilon}}\right)_{i,j}^2,$$

where we use (12). In $\overset{e}{\leq}$, we use

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{m} \mathbb{E}\left[\varphi_{k,i}^2\right] &\leq \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{E}\left[\left(\mathbf{P}_k^\top \left(\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right)_{i,j}^2\right] \\ &= \sum_{k=1}^{K} \mathbb{E}\left[\left\|\mathbf{P}_k^\top \left(\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right\|_F^2\right] = \sum_{k=1}^{K} \mathbb{E}\left[\left\|\mathbf{G}_k - \nabla f(\mathbf{X}_k)\right\|_F^2\right] \leq K\sigma^2,\end{aligned}$$

where we use the fact that $\|\cdot\|_F$ is a unitarily invariant norm such that $\|\mathbf{P}^\top\mathbf{A}\mathbf{Q}\|_F = \|\mathbf{A}\|_F$ for orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$. Similarly, we obtain

$$\sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right] \leq K\left(\sqrt{2n}\sigma + n\sqrt{\varepsilon}\right) + 4\sum_{k=1}^{K} \mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\sqrt[4]{\mathbf{V}_k + \varepsilon}}\right\|_F^2\right],$$

and the proof is completed by using $\sigma \leq \hat{\sigma}$. ■

The following gives the proof of Theorem 1.

Proof 6 (Proof of Theorem 1) *Plugging (6) in* (5)*, we have*

$$\begin{aligned}&\left(\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2 \\ \leq&\frac{1}{2}\left(\sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right] + \sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]\right)\mathcal{S}_K \\ \leq&\left(\frac{1}{2}K\left(\sqrt{2}\left(\sqrt{m}+\sqrt{n}\right)\hat{\sigma} + (m+n)\sqrt{\varepsilon}\right) + 4\mathcal{S}_K\right)\cdot\mathcal{S}_K \\ \leq&\left(K\sqrt{m+n}\left(\hat{\sigma} + \frac{1}{2}\sqrt{(m+n)\,\varepsilon}\right) + 4\mathcal{S}_K\right)\cdot\mathcal{S}_K \\ \overset{a}{\leq}&\left(2K\sqrt{m+n}\hat{\sigma} + 4\mathcal{S}_K\right)\cdot\mathcal{S}_K \\ \overset{b}{\leq}&\left(2K\sqrt{m+n}\hat{\sigma} + 28\frac{\sqrt{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}}{\delta}\right)\cdot 7\frac{\sqrt{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}}{\delta} \\ =&14K\sqrt{m+n}\hat{\sigma}\cdot\frac{\sqrt{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}}{\delta} + 196\frac{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}{\delta^2},\end{aligned}$$

where we let $\varepsilon = \frac{\tau\hat{\sigma}^2}{m+n}$ with $\tau \leq 1$ in $\overset{a}{\leq}$, while using the upper bound for $\mathcal{S}_K$ in Lemma 1 at $\overset{b}{\leq}$. Therefore,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq&\sqrt{14}\sqrt[4]{\frac{(m+n)\hat{\sigma}^2}{\delta^2}}\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)}{K}} + 14\sqrt{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)}{K\delta^2}} \\ \overset{c}{\leq}&\left(4\left(\sqrt{m+n}\frac{\hat{\sigma}}{\delta}\right)^{\frac{1}{2}} + 14\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}{K}} \\ \leq&\left(2\sqrt{m+n} + 16\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}{K}},\end{aligned}$$

where we use $\hat{\sigma}^4 \geq \frac{\hat{\sigma}^2 L(f(\mathbf{X}_1) - f^*)}{K}$ from the definition of $\hat{\sigma}$ in $\overset{c}{\leq}$. ■

C PROOFS FOR SPLUS

The SPlus algorithm proposed in (Frans et al., 2025) is shown in Algorithm 3. We extract the core idea of sign-type normalization, replace the sign-type update by a truncated modification, absorb the factor $\frac{2}{m+n}$ into η, and omit the iterate averaging, which leads to a suitable generalized Algorithm 6. We take a slightly different analytical approach for SPlus: since the matrix $\mathbf{\Phi}_k$ in this algorithm is computed directly from the current rotated momentum (unlike SOAP, which includes a historical averaging term $\mathbf{V}_k$), the entire analysis is simplified. We can deduce the following lemma from Lemma 7. It is crucial for algorithms with instantaneous normalizers, such as SPlus.

Lemma 8 *Suppose that Assumptions 1-3 and condition $\mathbf{D}_{k,i,j} \geq \delta$ hold. Let $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ with any $\gamma \in (0,1]$, $1-\theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$, and $\eta = \sqrt{\frac{\delta^2(f(\mathbf{X}_1)-f^*)}{4LK\hat{\sigma}^2}}$. Then for the general Algorithm 1, we have*

$$\sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\|_F\right] \leq \sqrt{13}K\hat{\sigma}.$$

Proof 7 *Taking expectation on both sides of (11) with respect to ζ_k conditioned on $\mathcal{F}_{k-1}$, summing over $k = 2, 3, \cdots, K$, and adding (10), we have*

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\|_F^2\right] \\
\leq&\theta \sum_{k=2}^{K} \mathbb{E}_{\mathcal{F}_{k-1}}\left[\|\mathbf{M}_{k-1} - \nabla f(\mathbf{X}_{k-1})\|_F^2\right] + \frac{L^2\eta^2}{\delta(1-\theta)} \sum_{k=2}^{K} \mathbb{E}_{\mathcal{F}_{k-1}}\left[\left\|\frac{\mathbf{P}_{k-1}^\top \mathbf{M}_{k-1}\mathbf{Q}_{k-1}}{\mathbf{D}_{k-1}^{\frac{1}{2}}}\right\|_F^2\right] \\
&+ (K-1)(1-\theta)^2\sigma^2 + 2L\left(f(\mathbf{X}_1) - f^*\right) + (1-\theta)^2\sigma^2 \\
\leq&\theta \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\|_F^2\right] + \frac{L^2\eta^2}{\delta(1-\theta)} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \mathbf{M}_k\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \\
&+ K(1-\theta)^2\sigma^2 + 2L\left(f(\mathbf{X}_1) - f^*\right) \\
\leq&\theta \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\|_F^2\right] + \frac{2L^2\eta^2}{\delta(1-\theta)} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \\
&+ \frac{2L^2\eta^2}{\delta(1-\theta)} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top (\mathbf{M}_k - \nabla f(\mathbf{X}_k))\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + K(1-\theta)^2\sigma^2 + 2L\left(f(\mathbf{X}_1) - f^*\right) \\
\overset{a}{\leq}&\theta \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\|_F^2\right] + \frac{2L^2\eta^2}{\delta(1-\theta)} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \\
&+ \frac{2L^2\eta^2}{\delta^2(1-\theta)} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\mathbf{P}_k^\top (\mathbf{M}_k - \nabla f(\mathbf{X}_k))\mathbf{Q}_k\right\|_F^2\right] + K(1-\theta)^2\sigma^2 + 2L\left(f(\mathbf{X}_1) - f^*\right) \\
\overset{b}{\leq}&\theta \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\|_F^2\right] + \frac{\delta(1-\theta)}{2} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \\
&+ \frac{1-\theta}{2} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\|_F^2\right] + K(1-\theta)^2\sigma^2 + 2L\left(f(\mathbf{X}_1) - f^*\right),
\end{aligned}$$

where we use $\mathbf{D}_{k,i,j} \geq \delta$ *in* $\overset{a}{\leq}$, $\frac{L^2\eta^2}{\delta^2(1-\theta)^2} \leq \frac{1}{4}$ *and the fact that* $\|\cdot\|_F$ *is a unitarily invariant norm in* $\overset{b}{\leq}$. *So we have*

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\|_F^2\right] \\
\leq &\frac{4L\left(f(\mathbf{X}_1) - f^*\right)}{1-\theta} + 2K(1-\theta)\sigma^2 + \delta \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] \\
\overset{c}{\leq} &\frac{4L\left(f(\mathbf{X}_1) - f^*\right)}{1-\theta} + 2K(1-\theta)\sigma^2 + 7\sqrt{K\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)} \\
\overset{d}{\leq} &13\sqrt{K\hat{\sigma}^2 L\left(f(\mathbf{X}_1) - f^*\right)} \overset{e}{\leq} 13K\hat{\sigma}^2,
\end{aligned}$$

where we use Lemma 1 in $\overset{c}{\leq}$, $1-\theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$ *and* $\sigma^2 \leq \hat{\sigma}^2$ *in* $\overset{d}{\leq}$, *and* $L\left(f(\mathbf{X}_1) - f^*\right) \leq K\hat{\sigma}^2$ *from the definition of* $\hat{\sigma}$ *in* $\overset{e}{\leq}$. *Using Cauchy-Schwarz inequality, we have*

$$\sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\|_F\right] \leq \sqrt{K\sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\|\nabla f(\mathbf{X}_k) - \mathbf{M}_k\|_F^2\right]} \leq \sqrt{13}K\hat{\sigma}.$$

■

The following lemma is an analogue of Lemma 3, but with a simpler proof.

Lemma 9 *Suppose that Assumptions 1-3 and condition* $\mathbf{D}_{k,i,j} \geq \delta$ *hold. Let* $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ *with any* $\gamma \in (0,1]$, $1-\theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$, *and* $\eta = \sqrt{\frac{\delta^2(f(\mathbf{X}_1)-f^*)}{4LK\hat{\sigma}^2}}$. *Then for Algorithm 6, we have*

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right] &\leq 2\sqrt{13m}K\hat{\sigma} + 2Km\delta + \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right], \\
\sum_{k=1}^{K} \mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right] &\leq 2\sqrt{13n}K\hat{\sigma} + 2Kn\delta + \sum_{k=1}^{K} \mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right].
\end{aligned} \tag{13}$$

Proof 8 *Different from the definition for SOAP, we instead define*

$$b_{k,i} = \max_{1\leq j\leq n} \left|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k\right|_{i,j} + \delta \geq \max_{1\leq j\leq n} \mathbf{D}_{k,i,j}$$

for (3). Denote

$$j_i^* = \operatorname*{argmax}_{1\leq j\leq n} \left|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k\right|_{i,j}.$$

Therefore,

$$b_{k,i} = \left|\mathbf{P}_k^\top \mathbf{M}_k \mathbf{Q}_k\right|_{i,j_i^*} + \delta \leq \left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j_i^*} + \left|\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right|_{i,j_i^*} + \delta.$$

Since

$$\left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j_i^*} \leq \frac{\left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j_i^*}^2}{2b_{k,i}} + \frac{b_{k,i}}{2},$$

plugging in gives

$$
\begin{aligned}
b_{k,i} &\leq \frac{\left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j_i^*}^2}{b_{k,i}} + 2\left|\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right|_{i,j_i^*} + 2\delta \\
&\leq \sum_{j=1}^{n} \frac{\left|\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k\right|_{i,j}^2}{b_{k,i}} + 2\left\|\left(\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right)_{i,:}\right\|_2 + 2\delta \\
&\leq \sum_{j=1}^{n} \left(\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right)_{i,j}^2 + 2\left\|\left(\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right)_{i,:}\right\|_2 + 2\delta.
\end{aligned}
$$

Summing over $i = 1, \ldots, m$ and $k = 1, \ldots, K$ gives

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right] = \sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[b_{k,i}\right] \\
&\leq \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{\mathcal{F}_k}\left[\left(\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right)_{i,j}^2\right] + 2\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\left(\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right)_{i,:}\right\|_2\right] + 2Km\delta \\
&\leq \sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + 2\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\sqrt{m\sum_{i=1}^{m}\left\|\left(\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right)_{i,:}\right\|_2^2}\right] + 2Km\delta \\
&= \sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + 2\sqrt{m}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\mathbf{P}_k^\top\left(\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right)\mathbf{Q}_k\right\|_F\right] + 2Km\delta \\
&= \sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + 2\sqrt{m}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\mathbf{M}_k - \nabla f(\mathbf{X}_k)\right\|_F\right] + 2Km\delta \\
&\leq \sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + 2\sqrt{13m}K\hat{\sigma} + 2Km\delta,
\end{aligned}
$$

where we use Lemma 8 in the last inequality. Similarly,

$$
\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right] \leq \sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_k}\left[\left\|\frac{\mathbf{P}_k^\top \nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right] + 2\sqrt{13n}K\hat{\sigma} + 2Kn\delta.
$$

■

Similar to the proof of Theorem 1, the following gives the proof of Theorem 2.

Proof 9 (Proof of Theorem 2) *Substituting (13) into* (5) *and using the bound of $\mathcal{S}_K$ in Lemma 1, we have*

$$
\begin{aligned}
&\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla f(\mathbf{X}_k)\right\|_*\right]\right)^2 \\
&\leq \left(\sqrt{13}\left(\sqrt{m}+\sqrt{n}\right)K\hat{\sigma} + K(m+n)\delta + \mathcal{S}_K\right)\cdot\mathcal{S}_K \\
&\leq \left(\sqrt{38(m+n)}K\hat{\sigma} + 7\frac{\sqrt{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}}{\delta}\right)\cdot 7\frac{\sqrt{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}}{\delta} \\
&= 7\sqrt{38(m+n)}K\hat{\sigma}\cdot\frac{\sqrt{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}}{\delta} + 49\frac{K\hat{\sigma}^2 L\left(f\left(\mathbf{X}_1\right) - f^*\right)}{\delta^2},
\end{aligned}
$$

and

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq & 7\sqrt[4]{\frac{(m+n)\hat{\sigma}^2}{\delta^2}}\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1)-f^*\right)}{K}}+7\sqrt{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1)-f^*\right)}{K\delta^2}} \\ & \leq \left(7\left(\sqrt{m+n}\frac{\hat{\sigma}}{\delta}\right)^{\frac{1}{2}}+7\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1)-f^*\right)}{K}} \\ & \leq \left(4\sqrt{m+n}+11\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2 L\left(f(\mathbf{X}_1)-f^*\right)}{K}},\end{aligned}$$

where we use $\delta=\frac{\tau\hat{\sigma}}{\sqrt{m+n}}$*. Substituting this setting yields the stated bound.* ■

D On One-Sided Matrix Control

We can use one-sided matrix-control to reduce the dimension dependence from $\sqrt{m+n}$ to $\sqrt{\min\{m,n\}}$. Namely, we consider $\mathbf{B}_k$ and $\mathbf{C}_k$ separately. Note that simply setting $\mathbf{C}_k=\mathbf{I}$ or $\mathbf{B}_k=\mathbf{I}$ in Lemma 2 does not suffice to prove the following Lemma.

Lemma 10 *For the general Algorithm 1 and* $\mathbf{B}_k$ *and* $\mathbf{C}_k$ *defined in (4), we have*

$$\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2\leq\min\left\{\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right],\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]\right\}\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right].$$

Proof 10 *Recall that we define*

$$\begin{aligned}b_{k,i}&\geq\max_{1\leq j\leq n}\mathbf{D}_{k,i,j},\quad \mathbf{B}_k=\operatorname{diag}\left(b_{k,1},b_{k,2},\ldots b_{k,m}\right),\\ c_{k,j}&\geq\max_{1\leq i\leq m}\mathbf{D}_{k,i,j},\quad \mathbf{C}_k=\operatorname{diag}\left(c_{k,1},c_{k,2}\cdots c_{k,n}\right).\end{aligned}$$

So we have

$$\begin{aligned}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2 &\overset{a}{=}\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right\|_*\right]\right)^2\\ &\overset{b}{\leq}\left(\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{\frac{1}{2}}\right\|_{S_2}\left\|\mathbf{B}_k^{-\frac{1}{2}}\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right\|_{S_2}\right]\right)^2\\ &\overset{c}{\leq}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{\frac{1}{2}}\right\|_{S_2}^2\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{-\frac{1}{2}}\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right\|_{S_2}^2\right]\\ &\overset{d}{=}\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{B}_k^{-\frac{1}{2}}\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right\|_F^2\right]\\ &=\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\left(\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k\right)_{i,j}}{\sqrt{b_{k,i}}}\right)^2\right]\\ &\overset{e}{\leq}\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right)_{i,j}^2\right]\\ &=\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right],\end{aligned}$$

where we use the fact that $\|\cdot\|_*$ *is a unitarily invariant norm such that* $\|\mathbf{P}^\top\mathbf{A}\mathbf{Q}\|_*=\|\mathbf{A}\|_*$ *for orthogonal matrices* $\mathbf{P}$ *and* $\mathbf{Q}$ *in* $\overset{a}{=}$*, the Schatten-Hölder inequality and* $\|\mathbf{X}\|_*=\|\mathbf{X}\|_{S_1}$ *in* $\overset{b}{\leq}$*,*

Hölder's inequality in $\overset{c}{\leq}$, Lemma 6 in $\overset{d}{=}$, and $b_{k,i} \geq \mathbf{D}_{k,i,j}$ in $\overset{e}{\leq}$. Same goes for $\mathbf{C}_k$, as we have

$$\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2 \leq \sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]\cdot\sum_{k=1}^{K}\mathbb{E}\left[\left\|\frac{\mathbf{P}_k^\top\nabla f(\mathbf{X}_k)\mathbf{Q}_k}{\mathbf{D}_k^{\frac{1}{2}}}\right\|_F^2\right]$$

Combining both inequalities, the proof is completed. ■

Theorem 3 *Suppose that Assumptions 1-3 and condition $\sqrt{\mathbf{V}_{k,i,j}+\varepsilon} \geq \delta$ hold. Let $\hat{\sigma}^2 = \max\left\{\sigma^2, \frac{L(f(\mathbf{X}_1)-f^*)}{K\gamma^2}\right\}$ with any $\gamma \in (0,1]$, $1-\theta = \sqrt{\frac{L(f(\mathbf{X}_1)-f^*)}{K\hat{\sigma}^2}}$, $0<\beta<1$, $\eta = \sqrt{\frac{\delta^2(f(\mathbf{X}_1)-f^*)}{4LK\hat{\sigma}^2}}$, and $\varepsilon = \frac{\tau\hat{\sigma}^2}{\min\{m,n\}}$, where τ is a tuning factor with $\tau \in (0,1]$. Then for generalized SOAP (Algorithm 4), we have*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq \left(3\sqrt{\min\{m,n\}}+17\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}{K}}.$$

Proof 11 *Combining Lemma 10 and Lemma 3, we have*

$$\begin{aligned}
&\left(\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right]\right)^2\\
\leq&\min\left\{\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{B}_k\right],\sum_{k=1}^{K}\mathbb{E}\left[\operatorname{tr}\mathbf{C}_k\right]\right\}\mathcal{S}_K\\
\leq&\left(K\left(\sqrt{2\min\{m,n\}}\hat{\sigma}+\min\{m,n\}\sqrt{\varepsilon}\right)+4\mathcal{S}_K\right)\cdot\mathcal{S}_K\\
\overset{a}{\leq}&\left((1+\sqrt{2})K\sqrt{\min\{m,n\}}\hat{\sigma}+4\mathcal{S}_K\right)\cdot\mathcal{S}_K\\
\overset{b}{\leq}&\left((1+\sqrt{2})K\sqrt{\min\{m,n\}}\hat{\sigma}+28\frac{\sqrt{K\hat{\sigma}^2L\left(f\left(\mathbf{X}_1\right)-f^*\right)}}{\delta}\right)\cdot7\frac{\sqrt{K\hat{\sigma}^2L\left(f\left(\mathbf{X}_1\right)-f^*\right)}}{\delta}\\
=&7(1+\sqrt{2})K\sqrt{\min\{m,n\}}\hat{\sigma}\cdot\frac{\sqrt{K\hat{\sigma}^2L\left(f\left(\mathbf{X}_1\right)-f^*\right)}}{\delta}+196\frac{K\hat{\sigma}^2L\left(f\left(\mathbf{X}_1\right)-f^*\right)}{\delta^2},
\end{aligned}$$

where we let $\varepsilon = \frac{\tau\hat{\sigma}^2}{\min\{m,n\}}$ with $\tau \leq 1$ in $\overset{a}{\leq}$, while using the upper bound for $\mathcal{S}_K$ in Lemma 1 at $\overset{b}{\leq}$. Therefore,

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{X}_k)\|_*\right] \leq&\sqrt{21}\sqrt[4]{\frac{\min\{m,n\}\hat{\sigma}^2}{\delta^2}}\sqrt[4]{\frac{\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}{K}}+14\sqrt{\frac{\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}{K\delta^2}}\\
\overset{c}{\leq}&\left(5\left(\sqrt{\min\{m,n\}}\frac{\hat{\sigma}}{\delta}\right)^{\frac{1}{2}}+14\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}{K}}\\
\leq&\left(3\sqrt{\min\{m,n\}}+17\frac{\hat{\sigma}}{\delta}\right)\cdot\sqrt[4]{\frac{\hat{\sigma}^2L\left(f(\mathbf{X}_1)-f^*\right)}{K}},
\end{aligned}$$

where we use $\hat{\sigma}^4 \geq \frac{\hat{\sigma}^2L(f(\mathbf{X}_1)-f^*)}{K}$ from the definition of $\hat{\sigma}$ in $\overset{c}{\leq}$. ■