Sharper Regret Bounds for Shampoo

Dahngoon Kim
Seoul National University
dahngoon.k@snu.ac.kr

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

SHAMPOO is a Kronecker-structured second-order optimizer widely used for large-scale deep learning, yet existing regret analyses exhibit an unfavorable dependence on the gradient rank, leading to guarantees that are uniformly weaker than those of optimizers such as ADAGRAD and ONE-SIDED SHAMPOO. We provide a substantially sharper regret bound for the full two-sided SHAMPOO update, yielding bounds that can compete with full-matrix ADAGRAD and more recent structured variants such as ONE-SIDED SHAMPOO. We also establish the first sublinear regret guarantee for the commonly implemented variant SHAMPOO^2 by introducing a principled scaling factor. Our analysis holds for a continuum of exponent choices $p \in (0, \frac{1}{2}]$, also providing the first sublinear guarantees for SHAMPOO^{4p} with exponents other than $p = 1/4$.

1 Introduction

As the size and depth of modern large language models continue to increase, the choice of optimization algorithm has become a central factor in training efficiency and stability. Curvature-aware (second-order) methods have therefore received renewed attention, with competitive results reported both in industrial-scale training [Anil et al., 2022, GeminiTeam, 2024] and in academic benchmarks [Dahl et al., 2025, Kasimbeg et al., 2025]. A key challenge in this line of work is to leverage curvature information while controlling the memory and computational overhead of full-matrix preconditioning methods such as full-matrix ADAGRAD [Duchi et al., 2011].

A particularly promising approach is SHAMPOO [Gupta et al., 2018], which exploits the native matrix (or tensor) structure of neural network parameters to construct Kronecker-factored preconditioners, rather than vectorizing parameters and operating in the ambient dimension. This structure is especially well matched to architectures dominated by matrix-shaped weights (e.g., Transformers). Notably, a distributed `PyTorch` implementation of SHAMPOO achieved the top result in the recent ALGOPERF: TRAINING ALGORITHMS competition, reducing wall-clock time by 28% [Kasimbeg et al., 2025]. SHAMPOO has also been used in industry-scale LLM training; for example, Gemini 1.5 Flash was trained with SHAMPOO as the optimizer of choice [GeminiTeam, 2024].

Despite its practical success, several theoretical aspects of SHAMPOO remain unclear. Most notably, the existing regret bound for SHAMPOO does not align with its empirical performance. Although SHAMPOO is a second-order method, the known regret analysis by Gupta et al. [2018] does not effectively exploit its second-order structure and yields guarantees that are uniformly weaker than those of many popular optimizers, including first-order methods [Xie et al., 2025].

There has also been longstanding ambiguity regarding the choice of exponent in the SHAMPOO preconditioner [Morwani et al., 2025, Eschenhagen et al., 2025]. In practice, the preconditioner exponent is treated as a tunable hyperparameter, as suggested by Anil et al. [2021]. Shi et al. [2023] reports that SHAMPOO^2, which uses doubled exponents, often outperforms SHAMPOO. The recently popular structured optimizer SOAP [Vyas et al., 2025] also uses the SHAMPOO^2 exponents.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Regret upper bounds for optimizers. Here $G_t \in \mathbb{R}^{m\times n}, d = mn$. We define $g_t = \overline{\mathrm{vec}}(G_t) \in \mathbb{R}^d$, $L_T = \epsilon I_m + \sum_{t=1}^T G_t G_t^\top$, $R_T = \epsilon I_n + \sum_{t=1}^T G_t^\top G_t$, $S_T = \epsilon I_d + \sum_{t=1}^T g_t g_t^\top$. Let $D_2 = \max_{t\in[T]} \|x_t - x^*\|_2$, $D_\infty = \max_{t\in[T]} \|x_t - x^*\|_\infty$, and $D_{\mathrm{op}} = \max_{t\in[T]} \|X_t - X^*\|_{\mathrm{op}}$, where $x_t = \overline{\mathrm{vec}}(X_t)$. Also $r_G = \max_{t\in[T]} \mathrm{rank}(G_t)$.

Algorithm	**Regret Bound**
ADAGRAD [Duchi et al., 2011]	$\sqrt{2}\, D_2\, \mathrm{tr}(S_T^{1/2})$
ADAGRAD (DIAG) [Streeter and McMahan, 2010]	$\sqrt{2}\, D_\infty \sum_{i=1}^d \Big(\sum_{t=1}^T g_{t,i}^2\Big)^{1/2}$
ADAGRAD-NORM [Streeter and McMahan, 2010]	$\sqrt{2}\, D_2\, \mathrm{tr}(S_T)^{1/2}$
ONE-SIDED SHAMPOO [Xie et al., 2025]	$\sqrt{2}\, D_{\mathrm{op}}\big(\mathrm{tr}(L_T^{1/2}) + m\sqrt{n\epsilon}\big)$
SHAMPOO [Gupta et al., 2018]	$\sqrt{2}\, D_2\, \sqrt{r_G}\, \mathrm{tr}(L_T^{1/4})\, \mathrm{tr}(R_T^{1/4})$
SHAMPOO [This paper]	$D_{\mathrm{op}}\big(\lambda_{\max}(L_T^{1/4})\, \mathrm{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4})\, \mathrm{tr}(L_T^{1/4})\big)$
SHAMPOO^2 [This paper]	$\sqrt{6}\, D_{\mathrm{op}}\big(\lambda_{\max}(L_T^{1/2})\, \mathrm{tr}(R_T^{1/2}) + \lambda_{\max}(R_T^{1/2})\, \mathrm{tr}(L_T^{1/2})\big)^{1/2} \times \big[\mathrm{tr}(\log(\epsilon^{-1} L_T))\big]^{1/4} \big[\mathrm{tr}(\log(\epsilon^{-1} R_T))\big]^{1/4}$

Another issue is the scaling of the preconditioners. A heuristic widely used in practice is *learning rate grafting* [Agarwal et al., 2020], which keeps only the SHAMPOO^2 update direction while adopting the step size of a first-order optimizer. For instance, the winning submission of the ALGOPERF competition uses SHAMPOO^2 with learning rate grafting [Shi et al., 2023].

Since Gupta et al. [2018] introduced the algorithm and provided the only known regret guarantee, a number of works have attempted to clarify these issues [Morwani et al., 2025, Eschenhagen et al., 2025, Xie et al., 2025]. However, these subsequent analyses do not provide explicit regret guarantees for the two-sided Kronecker-structured algorithm commonly used in practice, and there is no convergence guarantee for SHAMPOO^2.

In this paper, we address these ambiguities by deriving a sharper regret bound for SHAMPOO that enables comparison with other optimizers, and by providing the first regret bound for SHAMPOO with alternative exponent choices, including the practically preferred SHAMPOO^2. We also identify an appropriate scaling factor for SHAMPOO and SHAMPOO^2, which resolves the scaling issue.

1.1 Main Contributions

- **Sharper regret bound for SHAMPOO.** We provide a substantially sharper regret guarantee for the full two-sided SHAMPOO update. As seen in Table 1, our bound removes the explicit dependence on the gradient rank that appears in the original analysis and improves the Frobenius norm dependence D_2 to the sharper operator norm D_{op} by exploiting the matrix structure. The resulting guarantee is sharper by at least $\frac{D_2}{D_{\mathrm{op}}}\sqrt{r_G/2}$ and, unlike the bound of Gupta et al. [2018], competes with full-matrix ADAGRAD and structured variants such as ONE-SIDED SHAMPOO.
- **First regret guarantee for SHAMPOO^2 through principled preconditioner scaling.** We establish the first regret bound for SHAMPOO^2, the doubled-exponent variant that is often preferred in practice (Theorem 11). The central obstacle is that a naive application of standard adaptive gradient analyses yields an $\mathcal{O}(T)$ term, making the guarantee vacuous.

  We resolve this with a scaling of the preconditioner and a symmetric decomposition of the regret, giving an $\widetilde{\mathcal{O}}(\sqrt{T})$ guarantee and the first theoretical basis for the widespread use of SHAMPOO^2.
- **First regret guarantees for SHAMPOO^{4p} with exponents other than $p = 1/4$.** Our analysis holds for a continuum of exponent choices $p \in (0, \frac{1}{2}]$, which also yields the first sublinear regret guarantees for SHAMPOO^{4p} with exponents other than $p = 1/4$ (Theorem 12).

- **A general regret lemma for Kronecker-structured preconditioners.** One key ingredient of our improved bound is Lemma 10, which sharpens the standard adaptive mirror descent regret decomposition when the preconditioner has Kronecker product structure. This lemma isolates where existing bounds lose sharpness by treating structured preconditioners as unstructured vectors, and can be reused beyond SHAMPOO whenever the Kronecker structure of the preconditioner is maintained.
- **Comparison to prior bounds on adaptive methods.** We compare our bounds to classical adaptive methods and prior analyses, showing that the previously reported uniform inferiority of SHAMPOO to ADAGRAD and ADAGRAD-NORM is due to the proof technique rather than an inherent limitation of the algorithm. We provide examples demonstrating that the new SHAMPOO bound can be smaller or larger than the bounds for ADAGRAD and ADAGRAD-NORM. It remains larger than the bound for ONE-SIDED SHAMPOO, but by a factor that depends only on the gradient spectra rather than by at least $\frac{D_2}{D_{\text{op}}}\sqrt{r_G}$.

2 Related work

2.1 Shampoo

Gupta et al. [2018] originally proposed the SHAMPOO optimizer as a computationally efficient surrogate for full-matrix ADAGRAD. Following this work, several authors have noted that the regret bound in Gupta et al. [2018] is loose [An et al., 2025, Xie et al., 2025] and have sought sharper guarantees. However, these efforts did not provide an explicit regret bound for the two-sided SHAMPOO algorithm commonly used in practice.

Gupta et al. [2017] unifies a range of adaptive online and stochastic optimization algorithms through potential functions, but such a potential function is not available for SHAMPOO. Building on Gupta et al. [2017], Xie et al. [2025] develops a unified analysis via *well-conditioned* preconditioners and proposes ONE-SIDED SHAMPOO, reporting a regret bound that is substantially smaller than the bound for SHAMPOO stated in Gupta et al. [2018]. In this paper, we show that this improvement is largely attributable to the looseness of the original SHAMPOO bound. Moreover, the common two-sided SHAMPOO does not fall within the well-conditioned preconditioner framework of Xie et al. [2025], limiting the applicability of their analysis. Independently, An et al. [2025] proposes the same ONE-SIDED SHAMPOO algorithm under the name ASGO, with similar analysis and justification.

Bernstein and Newhouse [2024] offers an alternative perspective, showing that SHAMPOO without accumulation is equivalent to steepest descent under the spectral norm. This viewpoint motivated the development of the optimizer MUON, but does not yield an explicit guarantee for SHAMPOO due to the fact that the Kronecker PSD cone $\mathcal{K} = \{L \otimes R : L \in \mathbb{R}_+^{m\times m}, R \in \mathbb{R}_+^{n\times n}\}$ is not convex.

2.2 Shampoo^2

Morwani et al. [2025] suggests interpreting SHAMPOO^2 as an optimal Kronecker product approximation to the full ADAGRAD preconditioner. They show that a single step of power iteration for computing the optimal Kronecker approximation yields exactly the SHAMPOO^2 preconditioner. Their numerical results indicate that this approximation can be sharper than that of SHAMPOO, motivating the use of SHAMPOO^2. However, they do not provide a corresponding regret or convergence guarantee.

Yong et al. [2023] generalizes the analysis of Duchi et al. [2011] to algorithms that minimize the dual norm under cone constraints, similar in spirit to Xie et al. [2025]. They propose an algorithm ADABK, which is essentially SHAMPOO^2 in a stochastic deep-learning optimization setting, but do not obtain an explicit regret bound. One obstacle is that SHAMPOO^2 is only an approximate solution to their surrogate minimization problem, making their proof framework inapplicable.

2.3 Other adaptive gradient methods

Duchi et al. [2011] introduced ADAGRAD and the broader concept of adaptive optimization methods. They proposed both the full-matrix variant ADAGRAD and the diagonal variant ADAGRAD (DIAG). The diagonal approach subsequently inspired many efficient methods, most notably ADAM [Kingma and Ba, 2015] and ADAMW [Loshchilov and Hutter, 2019]. Streeter and McMahan [2010] studied

Algorithm 1 Adaptive Gradient Method

Require: $\eta > 0$
  Initialize $x_1 \in \mathbb{R}^d$.
  for $t = 1, 2, \ldots, T$ **do**
    Receive convex loss function $f_t : \mathbb{R}^d \to \mathbb{R}$.
    Compute gradient $g_t = \nabla f_t(x_t)$.
    Update preconditioner $H_t \succ 0$.
    Update:
      $x_{t+1} \leftarrow x_t - \eta\, H_t^{-1} g_t$
  end for

similar adaptive schemes concurrently and also proposed ADAGRAD-NORM, which only retains adaptive stepsizes while discarding second-order directional information.

3 Preliminaries

3.1 Online convex optimization

In an online convex optimization (OCO) problem, the learner predicts a vector in a convex domain $\mathcal{X} \subseteq \mathbb{R}^d$ for T rounds. After predicting $x_t \in \mathcal{X}$ on round t, a convex loss function $f_t : \mathcal{X} \to \mathbb{R}$ is revealed to the learner, potentially adversarially based on the learner's past predictions. The learner then incurs the loss $f_t(x_t)$ and receives its gradient $\nabla f_t(x_t)$ as feedback.

The goal of the learner is to achieve low regret, defined as the cumulative loss relative to the best fixed decision in $\mathcal{X}$. Mathematically, the learner aims to minimize

$$\mathcal{R}_T \;:=\; \sum_{t=1}^{T} f_t(x_t) \;-\; \min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x)\,.$$

Notation. Unless otherwise specified, we will work with the set $\mathcal{X} = \mathbb{R}^d$. We assume that x_t has a matrix structure, as is usually the case in neural networks. The gradient feedback at round t is denoted by $g_t = \nabla f_t(x_t)$. The parameter matrix $X_t \in \mathbb{R}^{m \times n}$ satisfies $x_t = \overline{\mathrm{vec}}(X_t)$, where $\overline{\mathrm{vec}}(A) = \mathrm{vec}(A^\top)$ denotes the row-wise vectorization of A. Similarly, the gradient matrix $G_t \in \mathbb{R}^{m \times n}$ satisfies $g_t = \overline{\mathrm{vec}}(G_t)$.

3.2 Adaptive gradient methods

In the online convex optimization setting, online mirror descent (OMD) is a well-studied family of algorithms. Let $\psi : \mathcal{X} \to \mathbb{R}$ be a regularizer (also known as the mirror potential) that is 1-strongly convex with respect to a norm $\|\cdot\|$ on $\mathbb{R}^d$. Let B_ψ be the Bregman divergence, defined by $B_\psi(x, y) \;=\; \psi(x) - \psi(y) - \langle \nabla\psi(y),\, x - y\rangle$.

Then the mirror descent update is given by [Orabona, 2025]

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x\rangle + B_\psi(x, x_t) \right\}.$$

Duchi et al. [2011] originally proposed adaptive gradient methods as adaptive variants of online mirror descent and primal-dual methods. Instead of a fixed regularizer ψ, they allow ψ_t to vary over time.

Let $\{H_t\}$ be a sequence of symmetric positive definite matrices. We choose the regularizer $\psi_t(x) := \frac{1}{2}\|x\|_{H_t}^2$, which is 1-strongly convex with respect to the Mahalanobis norm $\|\cdot\|_{H_t}$. The adaptive mirror descent update then becomes

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \eta \langle g_t, x\rangle + \frac{1}{2}\|x - x_t\|_{H_t}^2 \right\}. \tag{1}$$

For the unconstrained problem $\mathcal{X} = \mathbb{R}^d$, the solution of (1) is the preconditioned gradient update $x_{t+1} = x_t - \eta H_t^{-1} g_t$. The resulting Algorithm 1 admits the regret bound of Lemma 1, whose proof can be found in, e.g., Duchi et al. [2011, Proposition 3].

Algorithm 2 SHAMPOO

Require: $\eta > 0, \epsilon > 0$
  Initialize $X_1 = 0_{m\times n}$; $L_0 = \epsilon I_m$; $R_0 = \epsilon I_n$
  for $t = 1, \dots, T$ **do**
    Receive convex loss function $f_t : \mathbb{R}^{m\times n} \to \mathbb{R}$.
    Compute gradient $G_t = \nabla f_t(X_t)$.
    Update preconditioners:
      $L_t \leftarrow L_{t-1} + G_t G_t^\top,\ R_t \leftarrow R_{t-1} + G_t^\top G_t$
    Update:
      $X_{t+1} \leftarrow X_t - \eta L_t^{-1/4} G_t R_t^{-1/4}$
  end for

Lemma 1. *For any sequence of matrices $H_1, \dots, H_T \succ 0$, the regret of Algorithm 1 satisfies*

$$\mathcal{R}_T \le \frac{1}{2\eta}\sum_{t=2}^{T}\left(\|x_t - x^*\|_{H_t}^2 - \|x_t - x^*\|_{H_{t-1}}^2\right) + \frac{1}{2\eta}\|x_1 - x^*\|_{H_1}^2 + \frac{\eta}{2}\sum_{t=1}^{T}(\|g_t\|_{H_t}^*)^2\,.$$

When H_t are increasing in Löwner order, the following specialized bound is commonly employed. Notably, Gupta et al. [2018, equation 4] uses this inequality to prove their regret bound for SHAMPOO.

Lemma 2. *For any sequence of matrices $H_T \succeq \cdots \succeq H_1 \succ 0$, the regret of Algorithm 1 satisfies*

$$\mathcal{R}_T \le \frac{D_2^2}{2\eta}\ \mathrm{tr}(H_T)\ +\ \frac{\eta}{2}\sum_{t=1}^{T}(\|g_t\|_{H_t}^*)^2$$

where $D_2 = \max_{t\in[T]}\|x_t - x^*\|_2$.

3.3 SHAMPOO

SHAMPOO [Gupta et al., 2018] is a second-order optimization method that leverages curvature information via Kronecker-structured preconditioning. At each iteration, SHAMPOO maintains two positive semidefinite preconditioner matrices, $L_t \in \mathbb{R}^{m\times m}$ and $R_t \in \mathbb{R}^{n\times n}$, that capture left- and right-sided second-moment statistics of the gradient matrix $G_t \in \mathbb{R}^{m\times n}$ (Algorithm 2).

The SHAMPOO update given in Algorithm 2 can be rewritten as

$$x_{t+1} = x_t - \eta\,(L_t^{1/4} \otimes R_t^{1/4})^{-1} g_t$$

which is an adaptive gradient method with the choice $H_t = L_t^{1/4} \otimes R_t^{1/4}$. Since H_t is increasing in Löwner order, naively applying Lemma 2 gives

$$\mathcal{R}_T \le \frac{D_2^2}{2\eta}\,\mathrm{tr}(L_T^{1/4})\,\mathrm{tr}(R_T^{1/4}) + \frac{\eta}{2}\sum_{t=1}^{T}(\|g_t\|_{L_t^{1/4}\otimes R_t^{1/4}}^*)^2$$

Gupta et al. [2018] bounds the second term by $\eta\, r_G\,\mathrm{tr}(L_T^{1/4})\,\mathrm{tr}(R_T^{1/4})$ and proves the following data-dependent regret bound for SHAMPOO.

Theorem 3 (Gupta et al., 2018)**.** *Assume that the gradients $G_1, \dots, G_T$ are matrices of rank at most r_G. Choosing $\eta = D_2/\sqrt{2r_G}$, the regret of Algorithm 2 compared to any $x^* \in \mathbb{R}^{mn}$ is bounded as*

$$\mathcal{R}_T \le \sqrt{2r_G}\, D_2\,\mathrm{tr}\Big(L_T^{1/4}\Big)\,\mathrm{tr}\Big(R_T^{1/4}\Big)$$

where $D_2 = \max_{t\in[T]}\|x_t - x^*\|_2$.

3.3.1 Looseness of SHAMPOO's regret bound

However, the bound given in Theorem 3 is significantly looser than those available for other adaptive gradient methods. It contains an undesirable $\sqrt{r_G}$ factor and depends on the Frobenius norm of the parameter matrix, as $D_2 = \max_{t\in[T]}\|x_t - x^*\|_2 = \max_{t\in[T]}\|X_t - X^*\|_{\mathrm{F}}$.

Proposition 4 shows that the regret bound for SHAMPOO in Gupta et al. [2018] is uniformly worse than that of ADAGRAD, and uniformly worse than that of ADAGRAD-NORM by at least a factor of $\sqrt{r_G}$. Proofs for the propositions are provided in Appendix I.

Proposition 4. *Let $\mathcal{B}_{\text{SHAMPOO}}^{(\text{Gupta et al.})}$, $\mathcal{B}_{\text{ADAGRAD}}$, and $\mathcal{B}_{\text{ADAGRAD-NORM}}$ be the regret bounds in Table 1. Then*

$$\mathcal{B}_{\text{SHAMPOO}}^{(\text{Gupta et al.})} \geq \mathcal{B}_{\text{ADAGRAD}} \text{ and } \mathcal{B}_{\text{SHAMPOO}}^{(\text{Gupta et al.})} \geq \sqrt{r_G}\, \mathcal{B}_{\text{ADAGRAD-NORM}}\,.$$

Proposition 5 shows that the existing regret bound for SHAMPOO is consistently worse than that of ONE-SIDED SHAMPOO by at least a factor of $\frac{D_2}{D_{\text{op}}}\sqrt{r_G}$, and has been used to support the argument for using ONE-SIDED SHAMPOO in Xie et al. [2025].

Proposition 5 (Xie et al., 2025)**.** *Let $\mathcal{B}_{\text{SHAMPOO}}^{(\text{Gupta et al.})}$ and $\mathcal{B}_{\text{ONE-SIDED SHAMPOO}}$ be the regret bounds stated in Table 1. Assuming $\epsilon \approx 0$,*

$$\mathcal{B}_{\text{SHAMPOO}}^{(\text{Gupta et al.})} \geq \frac{D_2}{D_{\text{op}}}\sqrt{r_G}\, \mathcal{B}_{\text{ONE-SIDED SHAMPOO}}.$$

Note that $D_{\text{op}} \leq D_2$ since $\|\cdot\|_{\text{op}} \leq \|\cdot\|_F$. Since SHAMPOO was designed to leverage the matrix structure of the parameter matrix, it is natural to expect a dependence on D_{op} rather than D_2, as for ONE-SIDED SHAMPOO. This uniform weakness arises because Gupta et al. [2018] does not exploit the Kronecker product structure when controlling the norm and dual-norm terms. As we will see, the bounds for both terms can be significantly improved.

4 Sharper regret bound for SHAMPOO

We provide a regret bound for SHAMPOO that is substantially sharper than Theorem 3. The proofs of Theorem 6 and Corollary 7 are deferred to Appendix E.

Theorem 6 (Regret of SHAMPOO)**.** *Suppose we run* SHAMPOO *(Algorithm 2) with $\eta > 0$ and $\epsilon > 0$. Then its regret compared to any $X^* \in \mathbb{R}^{m\times n}$ is bounded as follows.*

$$\mathcal{R}_T \leq \frac{D_{\text{op}}^2}{2\eta}\left(\lambda_{\max}(L_T^{1/4})\,\text{tr}(R_T^{1/4}) + \text{tr}(L_T^{1/4})\lambda_{\max}(R_T^{1/4})\right) + \eta\,\text{tr}(L_T^{1/2})^{1/2}\text{tr}(R_T^{1/2})^{1/2}$$

where $D_{\text{op}} = \max_{t\in[T]}\|X_t - X^*\|_{\text{op}}$.

Corollary 7. *Running with $\eta = D_{\text{op}}$, the regret of* SHAMPOO *(Algorithm 2) is bounded as follows,*

$$\mathcal{R}_T \leq D_{\text{op}}\left(\lambda_{\max}(L_T^{1/4})\,\text{tr}(R_T^{1/4}) + \text{tr}(L_T^{1/4})\lambda_{\max}(R_T^{1/4})\right)$$

The bound extends directly to blockwise SHAMPOO, the form used in practice for transformers, by applying it to each block and summing (Appendix F).

Proposition 8 shows that the new regret bound in Corollary 7 is sharper than Theorem 3, the bound of Gupta et al. [2018], by at least a factor of $\frac{D_2}{D_{\text{op}}}\sqrt{r_G/2}$, which is at least one whenever $r_G \geq 2$.

Proposition 8. *Let $\mathcal{B}_{\text{SHAMPOO}}^{(\text{Gupta et al.})}$ and $\mathcal{B}_{\text{SHAMPOO}}^{(\text{Ours})}$ be the bounds in Theorem 3 and Corollary 7. Then*

$$\mathcal{B}_{\text{SHAMPOO}}^{(\text{Gupta et al.})} \geq \frac{D_2}{D_{\text{op}}}\sqrt{\frac{r_G}{2}}\, \mathcal{B}_{\text{SHAMPOO}}^{(\text{Ours})}$$

Concretely, by dividing the previous SHAMPOO bound [Gupta et al., 2018] by our new bound (Corollary 7), we obtain $\sqrt{r_G}\frac{D_2}{D_{\text{op}}}\frac{\sqrt{2}}{r_{\text{eff}}(L_T^{1/4})^{-1} + r_{\text{eff}}(R_T^{1/4})^{-1}}$. Here $r_{\text{eff}}(A) := \frac{\text{tr}(A)}{\|A\|_{\text{op}}} \geq 1$ is the *effective rank* of A [Varma et al., 2019, Koltchinskii et al., 2020].

This improvement ratio is therefore larger when at least one of the following holds: (i) the maximum gradient rank r_G is large; (ii) the effective ranks of $L_T^{1/4}$ and $R_T^{1/4}$ are not close to one; (iii) the iterate geometry is not concentrated in a single eigendirection.

Section 6 shows that all three conditions hold along real LLM training trajectories, as verified by the per-factor measurements in Table 2, and that the improvement ratio is 3–4 orders of magnitude.

Our sharper regret bound in Corollary 7 does not suffer from the relative suboptimality stated in Proposition 4. Depending on the gradient and iterate sequences, our bound can be smaller or larger than those of ADAGRAD and ADAGRAD-NORM. Our bound is still larger than that of ONE-SIDED SHAMPOO, but by a factor that depends only on the spectra of L_T and R_T, rather than by a factor of at least $\frac{D_2}{D_{\text{op}}}\sqrt{r_G}$ as in Proposition 5.

Algorithm 3 SHAMPOO^{4p} with scaling

Require: $p \in (0, \frac{1}{2}]$, $\eta > 0$, $\epsilon > 0$
  Initialize $X_1 = 0_{m\times n}$; $L_0 = \epsilon I_m$; $R_0 = \epsilon I_n$
  for $t = 1, \dots, T$ **do**
    Receive convex loss function $f_t : \mathbb{R}^{m\times n} \to \mathbb{R}$.
    Compute gradient $G_t = \nabla f_t(X_t)$.
    Update preconditioners:
      $L_t \leftarrow L_{t-1} + G_t G_t^\top,\ R_t \leftarrow R_{t-1} + G_t^\top G_t$
    Compute scaling factor α_t:
    $$\alpha_t = \begin{cases} \frac{\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}{(\lambda_{\max}(L_t^p)\mathrm{tr}(R_t^p)+\lambda_{\max}(R_t^p)\mathrm{tr}(L_t^p))^{1/2}} & \text{if } p \in (0, \frac{1}{2}) \\ \frac{\mathrm{tr}(\log(\epsilon^{-1}L_t))^{1/4}\mathrm{tr}(\log(\epsilon^{-1}R_t))^{1/4}}{(\lambda_{\max}(L_t^p)\mathrm{tr}(R_t^p)+\lambda_{\max}(R_t^p)\mathrm{tr}(L_t^p))^{1/2}} & \text{if } p = \frac{1}{2} \end{cases}$$
    Update:
      $X_{t+1} \leftarrow X_t - \eta\, \alpha_t^{-1}\, L_t^{-p} G_t R_t^{-p}$
  end for

Proposition 9. *(i) There exist sequences of gradients $\{G_t\}_{t=1}^T$ and iterates $\{X_t\}_{t=1}^T$ for which*

$$\mathcal{B}^{(\textit{Ours})}_{\text{SHAMPOO}} < \mathcal{B}_{\text{ADAGRAD}} \text{ and } \mathcal{B}^{(\textit{Ours})}_{\text{SHAMPOO}} < \mathcal{B}_{\text{ADAGRAD-NORM}}\,.$$

(ii) Assume $\epsilon \approx 0$. Then $\mathcal{B}^{(\textit{Ours})}_{\text{SHAMPOO}} \geq \frac{2^{3/2}}{3^{3/4}}\,\mathcal{B}_{\text{ONE-SIDED SHAMPOO}} \approx 1.24\,\mathcal{B}_{\text{ONE-SIDED SHAMPOO}}$ for all sequences of gradients and iterates. If, moreover, L_T and R_T have flat spectra with ranks r_L and r_R, that is, the nonzero eigenvalues of each matrix are all equal, then with $\gamma := r_R/r_L$,

$$\frac{\mathcal{B}^{(\textit{Ours})}_{\text{SHAMPOO}}}{\mathcal{B}_{\text{ONE-SIDED SHAMPOO}}} = \frac{\gamma^{3/4} + \gamma^{-1/4}}{\sqrt{2}}\,,$$

which equals $\sqrt{2}$ at $\gamma = 1$ and attains the lower bound at $\gamma = 1/3$.

Hence, for flat spectra with $r_L = r_R$, the two-sided bound is exactly $\sqrt{2}$ times the one-sided bound, instead of being worse by at least $\frac{D_2}{D_{\text{op}}}\sqrt{r_G}$ as the prior analysis suggested.

One key technical ingredient behind our improved regret bound is that, when the Löwner-increasing preconditioner admits a Kronecker product structure, adaptive gradient methods in fact enjoy a sharper bound than Lemma 2. We state this general result separately as Lemma 10.

Lemma 10 (Regret of structured AGM)**.** *For any sequence of matrices $A_T \succeq \cdots \succeq A_1 \succ 0$, $B_T \succeq \cdots \succeq B_1 \succ 0$, the regret of Algorithm 1 with $H_t = A_t \otimes B_t$ satisfies*

$$\mathcal{R}_T \leq \frac{D_{\text{op}}^2}{2\eta}\,\left(\lambda_{\max}(A_T)\,\mathrm{tr}(B_T) + \mathrm{tr}(A_T)\lambda_{\max}(B_T)\right) + \frac{\eta}{2}\sum_{t=1}^T (\|g_t\|_{H_t}^*)^2$$

where $D_{\text{op}} = \max_{t\in[T]} \|X_t - X^*\|_{\text{op}}$.

5 Regret bound for SHAMPOO^2

The main obstacle in obtaining a regret guarantee for SHAMPOO^2 is that using the preconditioner $H_t = L_t^{1/2} \otimes R_t^{1/2}$ causes the first term in Lemmas 2 and 10 to scale as $\mathcal{O}(T)$. The dual-norm term, in contrast, grows only as $\mathcal{O}(\log T)$, so the two terms decouple in scale and a fixed step size cannot recover sublinear regret without prior knowledge of T. Without it, the resulting regret bound is $\mathcal{O}(T)$, which is suboptimal against the standard bounds for adaptive gradient methods and the lower bound $\Omega(\sqrt{T})$ [Zinkevich, 2003].

Moreover, introducing a scaling factor to remedy this issue is not straightforward, since a decreasing scaling factor violates the assumption of Lemmas 2 and 10 that the preconditioners be increasing in Löwner order. For instance, this issue is overlooked in Yong et al. [2023] when invoking Lemma 2. Learning rate grafting can be viewed as such a heuristic scaling [Agarwal et al., 2020, Eschenhagen et al., 2025], but it comes with no convergence guarantee.

We provide the first regret bound for SHAMPOO^2 in Theorem 11. We resolve the scaling issue by choosing an appropriate scaling factor guided by Lemma 10, and by refining the analysis of both the norm and dual-norm terms. The full proofs are provided in Appendices G and H.

Theorem 11 (Regret of SHAMPOO^2)**.** *Suppose we run Algorithm 3 with $p = \frac{1}{2}$, $\eta = \sqrt{\frac{3}{2}} D_{\mathrm{op}}$, $\epsilon > 0$. Then its regret compared to any $X^* \in \mathbb{R}^{m \times n}$ is bounded as follows,*

$$\mathcal{R}_T \leq \sqrt{6} D_{\mathrm{op}} \left(\lambda_{\max}(L_T^{1/2}) \mathrm{tr}(R_T^{1/2}) + \lambda_{\max}(R_T^{1/2}) \mathrm{tr}(L_T^{1/2}) \right)^{1/2} \cdot \mathrm{tr}(\log(\epsilon^{-1} L_T))^{1/4} \mathrm{tr}(\log(\epsilon^{-1} R_T))^{1/4}$$

where $D_{\mathrm{op}} = \max_{t \in [T]} \|X_t - X^*\|_{\mathrm{op}}$.

More generally, our analysis applies to any exponent $p \in (0, \frac{1}{2})$, yielding the following theorem. This is also the first regret guarantee for exponents $p \in (0, \frac{1}{2})$ other than $p = \frac{1}{4}$.

Theorem 12 (Regret of SHAMPOO^{4p})**.** *Suppose we run Algorithm 3 with $p \in (0, \frac{1}{2})$, $\eta = \sqrt{\frac{3(1-2p)}{2}} D_{\mathrm{op}}$, $\epsilon > 0$. Then its regret compared to any $X^* \in \mathbb{R}^{m \times n}$ is bounded as*

$$\mathcal{R}_T \leq \sqrt{\frac{6}{1-2p}} D_{\mathrm{op}} \left(\lambda_{\max}(L_T^p) \mathrm{tr}(R_T^p) + \lambda_{\max}(R_T^p) \mathrm{tr}(L_T^p) \right)^{1/2} \mathrm{tr}(L_T^{1-2p})^{1/4} \mathrm{tr}(R_T^{1-2p})^{1/4}$$

where $D_{\mathrm{op}} = \max_{t \in [T]} \|X_t - X^*\|_{\mathrm{op}}$.

The choice of α_t is not ad hoc, it is reverse-engineered from the regret decomposition. In the proof, our symmetric decomposition of the norm-dependent regret into $\frac{\|L_t^p\|_{\mathrm{op}} \|R_t^p\|_* + \|R_t^p\|_{\mathrm{op}} \|L_t^p\|_*}{2}$ and the choice of α_t make the regret analysis go through for the non-monotonic Kronecker-structured preconditioners while preserving sublinear growth.

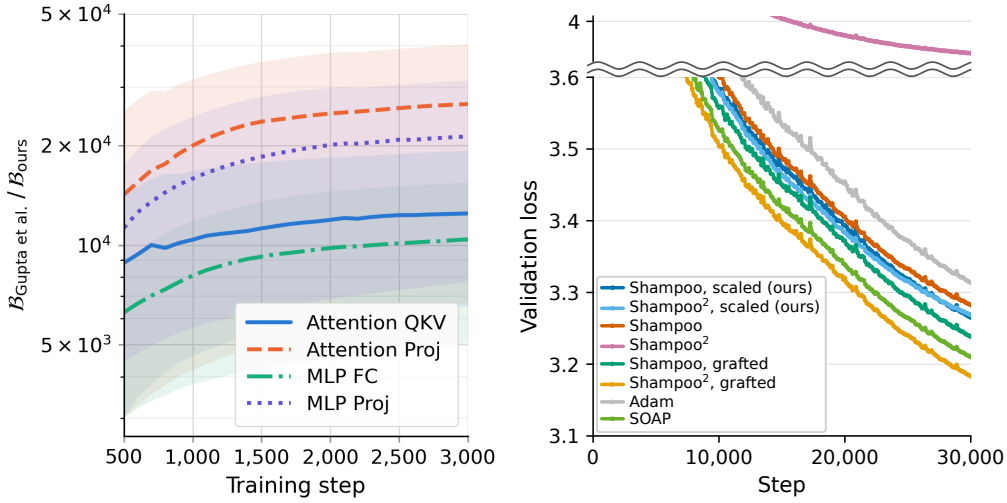
Finally, our regret bound for exponents $p \in (0, \frac{1}{2})$ (Theorem 12) scales as $\mathcal{O}(\sqrt{T})$, whereas the bound for SHAMPOO^2 (Theorem 11) scales as $\widetilde{\mathcal{O}}(\sqrt{T})$, or $\mathcal{O}(\sqrt{T \log T})$. For $p = \frac{1}{4}$, Theorem 12 also gives a bound for SHAMPOO with the scaling factor α_t, the scaled variant used in Section 6, and this bound is at most $\sqrt{6}$ times the bound of Corollary 7 for unscaled SHAMPOO (Corollary 18 and Proposition 19 in Appendix J).

Tightness. The $\mathcal{O}(\sqrt{T})$ rate is unimprovable in the worst case by the minimax lower bound of Abernethy et al. [2008], and the inequalities in our proofs are tight up to constant factors when the increments $G_t G_t^\top$ are small relative to the accumulated preconditioners and $X^* - X_t$ has a flat spectrum aligned with the preconditioner eigenbases. Whether the data-dependent spectral terms are optimal is open, as it would require a lower bound tailored to Kronecker-structured preconditioners, which to our knowledge does not exist.

Scope of the guarantees. Theorems 11 and 12 hold for Algorithm 3, in which the update is scaled by α_t. They do not cover implementations that use learning rate grafting, for which we are not aware of any regret bound. Computing α_t adds no matrix computation, since it only needs traces and largest eigenvalues that are byproducts of the eigendecompositions already performed (Appendix B.3).

6 Numerical experiments

While our contributions are primarily theoretical, we conduct preliminary experiments on language modeling. We train NanoGPT ($\sim$124M parameters) on the FineWeb-10B dataset for 32,000 steps, comparing SHAMPOO ($p = \frac{1}{4}$) and SHAMPOO^2 ($p = \frac{1}{2}$) under three scaling strategies (unscaled, ADAM-grafted, and Algorithm 3) against ADAM and SOAP [Vyas et al., 2025] baselines, and quantify the sharpness improvement of our bound along a SHAMPOO training trajectory. See Appendix A for more details on the experimental setup. Experiments on GPT-2 Medium ($\sim$345M parameters) are reported in Appendix B.



(a) Bound improvement ratio

(b) NanoGPT validation loss

Figure 1: (a) Ratio of the prior bound [Gupta et al., 2018] to ours along the SHAMPOO training trajectory, per 768×768 preconditioning block in NanoGPT (median over the blocks of each matrix type; shaded range over the blocks). Our bound is consistently 3–4 orders of magnitude smaller, and the gap widens as training progresses. (b) Validation loss on FineWeb-10B for NanoGPT trained for 32,000 steps with SHAMPOO ($p = \frac{1}{4}$) and SHAMPOO^2 ($p = \frac{1}{2}$) under three scaling strategies: unscaled, grafted (ADAM), and our principled scaling (Algorithm 3), together with ADAM and SOAP baselines. Note that the y-axis is broken above 3.6; unscaled SHAMPOO^2 appears in the strip at the top. Both scaled variants outperform their unscaled counterparts and ADAM, and trail the grafted variants and SOAP by at most 0.09.

Table 2: Bound improvement ratio $\mathcal{B}_{\text{Gupta et al.}}/\mathcal{B}_{\text{ours}}$ and its three factors after 3,000 steps of training on NanoGPT, reported as ranges over the 768×768 blocks of each matrix type. The effective-rank factor is $\sqrt{2}/\left(r_{\text{eff}}(L_T^{1/4})^{-1} + r_{\text{eff}}(R_T^{1/4})^{-1}\right)$.

Matrix type	r_G	D_2/D_{op}	Effective-rank factor	$\mathcal{B}_{\text{Gupta et al.}}/\mathcal{B}_{\text{ours}}$
Attention QKV	745–768	2.9–5.3	78–172	6.6×10^3–1.9×10^4
Attention Proj	759–768	2.7–8.6	89–187	6.7×10^3–4.0×10^4
MLP FC	647–765	2.2–3.4	85–178	5.1×10^3–1.5×10^4
MLP Proj	619–748	3.3–7.5	79–166	7.7×10^3–3.1×10^4

Bound comparison along training trajectories. We evaluate the prior regret bound [Gupta et al., 2018] and our improved bound of Corollary 7 along the (unscaled) SHAMPOO training trajectory. Figure 1(a) shows the ratio of the two bounds for every preconditioning block, grouped by matrix type, while training NanoGPT. After 3,000 steps of training, our bound is 5,100–40,000$\times$ smaller across all blocks. Table 2 decomposes this ratio into the three factors identified in Section 4, each reported as a range over the blocks of a matrix type, and shows that all three conditions hold.

The gradients are at or near full rank ($r_G \geq 619$ of 768), so $\sqrt{r_G}$ alone contributes a factor of at least 25. The iterates are spread across eigendirections (D_2/D_{op} between 2.2 and 8.6). The effective ranks of $L_T^{1/4}$ and $R_T^{1/4}$ are far from one, so the effective-rank factor contributes a further 78–187$\times$ improvement. Removing the r_G dependence is therefore one source of the improvement, but not the only one. The ratio grows along training (Appendix B.2), and the picture is the same at a larger scale; on GPT-2 Medium, after 2,000 steps of training, our bound is 3,100–51,000$\times$ smaller (Table 3).

Practical performance of the scaled variants. For SHAMPOO ($p = \frac{1}{4}$), Figure 1(b) shows that our scaled variant improves on the unscaled baseline and on ADAM, and trails the grafted variant by

Table 3: Bound improvement ratio $\mathcal{B}_{\text{Gupta et al.}}/\mathcal{B}_{\text{ours}}$ and its three factors after 2,000 steps of training on GPT-2 Medium, reported as ranges over the 1024×1024 blocks of each matrix type (full rank is 1024). The effective-rank factor is defined as in Table 2.

Matrix type	r_G	D_2/D_{op}	Effective-rank factor	$\mathcal{B}_{\text{Gupta et al.}}/\mathcal{B}_{\text{ours}}$
Attention QKV	703–1018	2.2–4.4	69–180	4.7×10^3–2.2×10^4
Attention Proj	988–1019	3.6–8.4	90–214	1.5×10^4–5.1×10^4
MLP FC	418–1005	1.7–4.9	62–139	3.1×10^3–2.1×10^4
MLP Proj	963–1006	1.5–5.9	81–132	4.7×10^3–2.5×10^4

at most 0.05. For SHAMPOO^2, the analysis predicts that the unscaled method should fail, since its norm term grows as $\mathcal{O}(T)$ while its dual-norm term grows as $\mathcal{O}(\log T)$, and that α_t should repair it. Figure 1(b) confirms this: unscaled SHAMPOO^2, even after tuning, plateaus at a validation loss of 3.731, whereas the scaled variant reaches 3.257, outperforming ADAM (3.300) and trailing the grafted variant (3.168) and SOAP (3.196) by at most 0.09. The remaining gap to grafting is expected, since our scaling minimizes the worst-case regret bound rather than being tuned for performance. GPT-2 Medium shows the same pattern (Appendix B).

7 Conclusion

Summary of contributions. Our first contribution is a substantially sharper regret bound for the standard (two-sided) SHAMPOO update, which removes the unfavorable dependence on the gradient rank present in prior analyses and improves the iterate-dependent factor D_2 to the matrix-form quantity D_{op}. This yields a bound that enables direct comparison with classical adaptive methods such as ADAGRAD and with structured variants such as ONE-SIDED SHAMPOO.

Our second contribution is the first regret guarantee for SHAMPOO^2. The main technical challenge is that naively using $H_t = L_t^{1/2} \otimes R_t^{1/2}$ leads to an $\mathcal{O}(T)$ term in the standard OMD-style bounds. We address this by introducing a scaling α_t and employing a symmetric decomposition of the regret, resulting in an $\widetilde{\mathcal{O}}(\sqrt{T})$ guarantee. Our analysis also yields $\mathcal{O}(\sqrt{T})$ regret bounds for any exponent $p \in (0, \frac{1}{2})$, the first sublinear guarantees for exponents other than $p = \frac{1}{4}$.

Future directions and limitations. Our sharpened regret bounds transfer directly to other settings through standard reductions. Online-to-batch conversion yields convergence rates for stochastic convex optimization, and online-to-nonconvex conversion [Cutkosky et al., 2023] yields stationarity guarantees for smooth nonconvex objectives. Several directions remain open. Including common deep-learning modifications such as momentum and weight decay in the analysis would bring the theory closer to the variants used in practice. Whether the same proof strategy gives sharper analyses for other Kronecker-structured optimizers is another open question. In particular, the preconditioner backbone of SOAP [Vyas et al., 2025] is exactly the SHAMPOO^2 structure $L_t^{1/2} \otimes R_t^{1/2}$, so Lemma 10 and our scaling technique provide a natural starting point for its analysis, the remaining step being the diagonal ADAM correction in the rotated basis. Moreover, whether it is possible to remove the logarithmic dependence in the SHAMPOO^2 bound, and whether unscaled SHAMPOO^2 admits any sublinear regret guarantee, remain open. Finally, works such as Yang et al. [2025], Anil et al. [2021] that base their regret analysis of SHAMPOO-based algorithms on Gupta et al. [2018] directly inherit sharper guarantees from our analysis.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant both funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023-00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008).

References

Jacob D. Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory (COLT 2008)*, pages 415–424. Omnipress, 2008. URL https://www.learningtheory.org/colt2008/papers/111-Abernethy.pdf.

Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling adaptive gradient methods from learning rates. *CoRR*, abs/2002.11803, 2020. URL https://arxiv.org/abs/2002.11803.

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=fru52tkjHf.

T. Ando. Concavity of certain maps on positive definite matrices and applications to hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(79)90179-4. URL https://www.sciencedirect.com/science/article/pii/0024379579901794.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning, 2021. URL https://arxiv.org/abs/2002.09018.

Rohan Anil, Sandra Gadanho, Da Huang, Nijith Jacob, Zhuoshu Li, Dong Lin, Todd Phillips, Cristina Pop, Kevin Regan, Gil I. Shamir, Rakesh Shivanna, and Qiqi Yan. On the factory floor: Ml engineering for industrial-scale ads recommendation models, 2022. URL https://arxiv.org/abs/2209.05310.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024. URL https://arxiv.org/abs/2409.20325.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6643–6670. PMLR, 2023. URL https://proceedings.mlr.press/v202/cutkosky23a.html.

George E. Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, Juhan Bae, Justin Gilmer, Abel L. Peirson, Bilal Khan, Rohan Anil, Mike Rabbat, Shankar Krishnan, Daniel Snider, Ehsan Amid, Kongtao Chen, Chris J. Maddison, Rakshith Vasudev, Michal Badura, Ankush Garg, and Peter Mattson. Benchmarking neural network training algorithms, 2025. URL https://arxiv.org/abs/2306.07179.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. URL https://api.semanticscholar.org/CorpusID:538820.

Runa Eschenhagen, Aaron Defazio, Tsung-Hsien Lee, Richard E. Turner, and Hao-Jun Michael Shi. Purifying shampoo: Investigating shampoo's heuristics by decomposing its preconditioner. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kePsKwxvaV.

Google GeminiTeam. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Vineet Gupta, Tomer Koren, and Yoram Singer. A unified approach to adaptive regularization in online and stochastic optimization, 2017. URL https://arxiv.org/abs/1706.06569.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024. URL https://github.com/KellerJordan/modded-nanogpt.

Priya Kasimbeg, Frank Schneider, Runa Eschenhagen, Juhan Bae, Chandramouli Shama Sastry, Mark Saroufim, BOYUAN FENG, Less Wright, Edward Z. Yang, Zachary Nado, Sourabh Medapati, Philipp Hennig, Michael Rabbat, and George E. Dahl. Accelerating neural network training: An analysis of the algoperf competition. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CtM5xjRSfm.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6980.

Vladimir Koltchinskii, Matthias Löffler, and Richard Nickl. Efficient estimation of linear functionals of principal components. *The Annals of Statistics*, 48(1):464 – 490, 2020. doi: 10.1214/19-AOS1816. URL https://doi.org/10.1214/19-AOS1816.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

K.T'. Löwner. Über monotone matrixfunktionen. *Mathematische Zeitschrift*, 38:177–216, 1934. URL http://eudml.org/doc/168495.

Depen Morwani, Itai Shapira, Nikhil Vyas, eran malach, Sham M. Kakade, and Lucas Janson. A new perspective on shampoo's preconditioner. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=c6zI3Cp8c6.

Francesco Orabona. A modern introduction to online learning, 2025. URL https://arxiv.org/abs/1912.13213.

Hao-Jun Michael Shi, Tsung-Hsien Lee, Shintaro Iwasaki, Jose Gallego-Posada, Zhijing Li, Kaushik Rangadurai, Dheevatsa Mudigere, and Michael Rabbat. A distributed data-parallel pytorch implementation of the distributed shampoo optimizer for training neural networks at-scale, 2023. URL https://arxiv.org/abs/2309.06497.

Matthew Streeter and H. Brendan McMahan. Less regret via online conditioning, 2010. URL https://arxiv.org/abs/1002.4862.

Paroma Varma, Frederic Sala, Ann He, Alexander Ratner, and Christopher Re. Learning dependency structures for weak supervision models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6418–6427. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/varma19a.html.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing shampoo using adam for language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IDxZhXrpNf.

Shuo Xie, Tianhao Wang, Sashank Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis, 2025. URL https://arxiv.org/abs/2503.10537.

Sifan Yang, Yuanyu Wan, Peijia Li, Yibo Wang, Xiao Zhang, Zhewei Wei, and Lijun Zhang. Dimension-free adaptive subgradient methods with frequent directions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=zpuKPuSYru.

Hongwei Yong, Ying Sun, and Lei Zhang. A general regret bound of preconditioned gradient method for dnn training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7866–7875, 2023. doi: 10.1109/CVPR52729.2023.00760.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 928–935. AAAI Press, 2003. ISBN 1577351894.

A Experimental details

Architecture and dataset. We train NanoGPT [Jordan et al., 2024] ($\sim$124M parameters, 12 layers, model dimension 768) on the FineWeb-10B dataset.

Hardware and compute budget. All NanoGPT runs use a single NVIDIA H100 GPU with FlexAttention. Each optimizer step processes a training sequence of $48 \times 1024 = 49{,}152$ tokens, and the validation loss is computed over 10,485,760 tokens. Each run uses 32,000 gradient steps, consuming approximately 1.57B training tokens.

Optimizer setup. Following Shi et al. [2023], we apply SHAMPOO only to hidden matrix parameters (attention QKV, projection and MLP weight matrices), and the embeddings, the language modeling head, and scalar parameters are updated with ADAM, as standard in the SHAMPOO literature for transformer training. For non-square weight matrices, we apply square-block decomposition, matching Shi et al. [2023].

We compare SHAMPOO ($p = \frac{1}{4}$) and SHAMPOO^2 ($p = \frac{1}{2}$), each under three scaling strategies: (i) unscaled, (ii) grafted, where the update direction is taken from the preconditioned update but the per-parameter step magnitude is matched to that of ADAM [Agarwal et al., 2020], and (iii) our proposed *principled scaling* from Algorithm 3 with the corresponding exponent. We also train ADAM and SOAP [Vyas et al., 2025] baselines under identical conditions. Preconditioners are updated every 100 steps, following Shi et al. [2023].

Hyperparameter tuning. We tune the hyperparameters of every method using Optuna, tuning peak learning rate, learning-rate cooldown fraction, second-moment decay β_2, and weight decay. For the learning-rate schedule we use the linear cooldown schedule of the source implementation [Jordan et al., 2024], with the cooldown fraction treated as a tuned hyperparameter.

Bound evaluation. For the bound-comparison experiment in Figure 1(a) and Appendix B.2, we run the tuned unscaled SHAMPOO in the setup above, except that preconditioning starts at step 500, and evaluate both bounds along 3,000 steps of training. At each gradient step, we record the gradient matrix G_t of every 768×768 preconditioning block of the hidden parameter matrices (Appendix F), its rank, and the accumulators $L_t = \epsilon I + \sum_{s \le t} G_s G_s^\top$ and $R_t = \epsilon I + \sum_{s \le t} G_s^\top G_s$. We then compute both the prior bound from Theorem 3 and our bound from Corollary 7 for each block. The diameter terms D_2 and D_{op} are computed empirically using the final iterate as X^*, that is, $D = \max_{s \le t} \|X_s - X^*\|$ over the iterates recorded every 50 steps. The bound evaluation on GPT-2 Medium (Appendix B.1) follows the same protocol with the tuned unscaled SHAMPOO of that experiment on a single H100 with 65,536 tokens per step and 1024×1024 blocks, with iterates recorded every 100 steps, and evaluates both bounds along 2,000 steps of training.

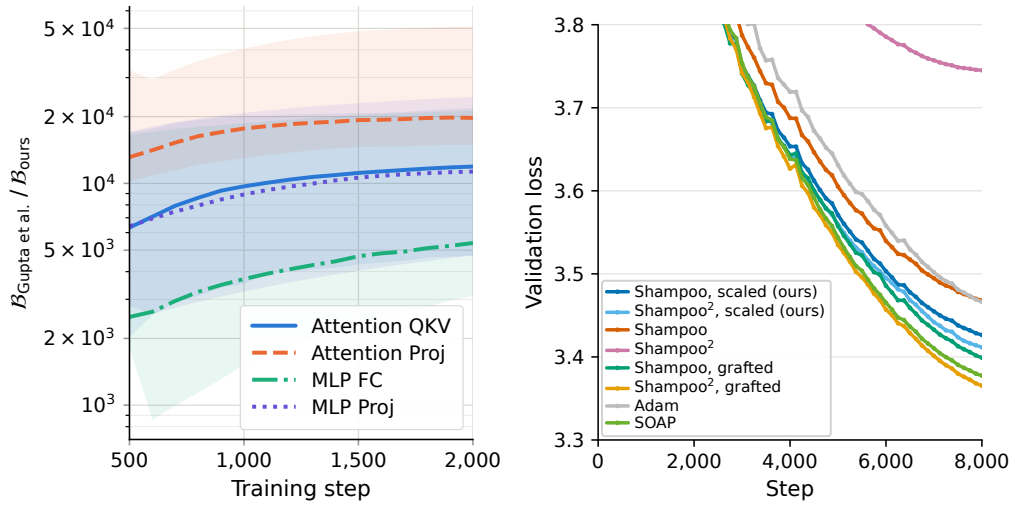
Implementation. Our NanoGPT training pipeline is mostly based on Jordan et al. [2024], with our SHAMPOO implementation following Shi et al. [2023] and SOAP following Vyas et al. [2025].

B Additional experiments

B.1 GPT-2 Medium

To verify that these findings hold at a larger scale, we train GPT-2 Medium ($\sim$345M parameters) on FineWeb-10B for 8,000 gradient steps and compare SHAMPOO ($p = \frac{1}{4}$) and SHAMPOO^2 ($p = \frac{1}{2}$) under the three scaling strategies against ADAM and SOAP, with all methods tuned via Optuna. Figure 2(b) reports the validation loss. The pattern is unchanged at the larger scale. Our scaled variants outperform their unscaled counterparts at every measured step. Unscaled SHAMPOO^2 trails every other method at every checkpoint and plateaus at 3.745, whereas unscaled SHAMPOO remains comparable to ADAM. The scaled variants stay within 0.05 validation loss of the grafted variants at every checkpoint while requiring no auxiliary optimizer state, and they outperform ADAM throughout. SOAP performs comparably to grafted SHAMPOO^2.

We also evaluate the two regret bounds along the GPT-2 Medium training trajectory, following the protocol of Appendix A. Figure 2(a) shows the ratio of the two bounds for every 1024×1024



(a) Bound improvement ratio

(b) GPT-2 Medium validation loss

Figure 2: (a) Ratio of the prior bound [Gupta et al., 2018] to ours along the SHAMPOO training trajectory of GPT-2 Medium, per 1024×1024 preconditioning block (median over the blocks of each matrix type; shaded range over the blocks). As for NanoGPT, our bound is 3–4 orders of magnitude smaller and the gap widens along training. (b) Validation loss for GPT-2 Medium trained for 8,000 steps with SHAMPOO ($p = \frac{1}{4}$) and SHAMPOO^2 ($p = \frac{1}{2}$) under the three scaling strategies of Appendix A, together with ADAM and SOAP. The scaled variants outperform their unscaled counterparts and ADAM, and stay close to the grafted variants and SOAP.

preconditioning block along 2,000 steps of training, and Table 3 in Section 6 reports the three factors of Section 4 and the resulting ratio after 2,000 steps of training, as Table 2 does for NanoGPT. The gradients of the attention projection and MLP projection blocks are at or near full rank ($r_G \geq 963$ of 1024), while those of the QKV and MLP FC blocks range from 418 to 1018; the iterates are spread across eigendirections (D_2/D_{op} between 1.5 and 8.4); and the effective-rank factor contributes 62–214×. Our bound improves upon the previous bound by 3–4 orders of magnitude (3,100–51,000×), as at the smaller scale, and both r_G and the improvement ratio grow along training.

B.2 Empirical verification of conditions (i)–(iii)

The improvement ratio discussed in Section 4 factorizes exactly as

$$\sqrt{r_G} \times \frac{D_2}{D_{\text{op}}} \times \frac{\sqrt{2}}{r_{\text{eff}}(L_T^{1/4})^{-1} + r_{\text{eff}}(R_T^{1/4})^{-1}},$$

and the three factors correspond to conditions (i), (iii), and (ii), respectively. We measure each factor per preconditioning block along the SHAMPOO trajectory of the bound-evaluation run of Appendix A (NanoGPT, ∼124M parameters, FineWeb-10B). Table 2 in Section 6 reports the three factors and the resulting ratio after 3,000 steps of training, and Table 3 reports the same quantities for GPT-2 Medium after 2,000 steps of training.

Condition (i): gradients are at or near full rank. Attention blocks reach $r_G \in [745, 768]$ and MLP blocks $r_G \in [619, 765]$, all within 20% of the full rank 768. This is structurally expected, since the gradient of a linear layer decomposes into one rank-one outer product per token and each optimizer step in our run aggregates 49,152 tokens.

Condition (ii): effective ranks are much larger than one. The effective ranks $r_{\text{eff}}(L_T^{1/4})$ and $r_{\text{eff}}(R_T^{1/4})$ range from 76 to 331. Although some raw accumulators L_T, R_T have ill-conditioned, heavy-tailed spectra with condition numbers up to 3×10^8, the fourth root compresses them to $\kappa(L_T^{1/4}) = \kappa(L_T)^{1/4} \leq 134$, which makes the spectrum of $L_T^{1/4}$ far flatter than that of L_T.

Condition (iii): iterate geometry is spread across eigendirections. The ratio D_2/D_{op} ranges from 2.2 to 8.6, out of a maximum of $\sqrt{768} \approx 27.7$, which shows that the iterates are not concentrated in a single eigendirection.

The regime strengthens along training. For the attention QKV blocks, between steps 1,000 and 3,000 the per-block improvement ratio grows from 4.6×10^3–1.7×10^4 to 6.6×10^3–1.9×10^4, driven by the growing r_G (719–767 to 745–768) and effective-rank factor (56–147 to 78–172). The gap thus widens along training, and the previous rank-dependent bound becomes increasingly unfavorable.

B.3 Computational overhead of the scaling factor

The scaling factor adds no matrix computation. The quantities used to compute α_t, namely $\lambda_{\max}(L_t^p)$, $\mathrm{tr}(L_t^p)$, $\mathrm{tr}(L_t^{1-2p})$ (or $\mathrm{tr}(\log(\epsilon^{-1}L_t))$ when $p = \frac{1}{2}$) and their R_t counterparts, are byproducts of the eigendecompositions that practical implementations such as that of Shi et al. [2023] already compute to form the inverse matrix roots. In distributed settings, the only added communication is a constant number of scalars per parameter block, which is negligible relative to the communication of the preconditioner factors themselves.

Empirically, we observe less than 1% wall-clock difference between our scaling and grafting at update frequency 10 in our preliminary runs on 8 H100 GPUs. SOAP replaces the eigendecomposition with a QR decomposition and power iteration for computational gain, and in our GPT-2 Medium experiments with update frequency 100, scaled SHAMPOO^2 was about 6% slower than SOAP in wall-clock time.

C Auxiliary lemmas

C.1 Mathematical preliminaries

Let $\otimes$ denote the Kronecker product, defined by

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix},$$

for $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$, so that $A \otimes B \in \mathbb{R}^{mp \times nq}$.

Following are some basic properties of Kronecker products. For the proof and further details, see, e.g., Horn and Johnson [1991].

Lemma 13. *Let $A \in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{n \times n}$. The following properties hold for Kronecker products.*

1. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$
2. $A \otimes B \succeq 0 \iff A \succeq 0 \wedge B \succeq 0$
3. $A \succeq C, B \succeq D \implies A \otimes B \succeq C \otimes D$
4. $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\,\mathrm{tr}(B)$

For the following Lemma 14, the proof of the first equality can be found in Gupta et al. [2018], and the second equality is a direct consequence of the first.

Lemma 14. *Let $G \in \mathbb{R}^{m \times n}$, $L \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{n \times n}$. Then one has*

$$(L \otimes R)\,\overline{\mathrm{vec}}(G) = \overline{\mathrm{vec}}(LGR^\top) \quad \textit{and} \quad \overline{\mathrm{vec}}(G)^\top (L \otimes R)\,\overline{\mathrm{vec}}(G) = \mathrm{tr}(G^\top LGR^\top).$$

The following Lemmas 15 and 16 are classic results from matrix analysis used in our analysis.

Lemma 15 (Löwner, 1934)**.** *The map $x \mapsto x^\alpha$ is operator monotone for $\alpha \in [0, 1]$. That is, if $0 \preceq X \preceq Y$ then $X^\alpha \preceq Y^\alpha$.*

Lemma 16 (Ando, 1979)**.** *The map $A \mapsto A^\alpha$ on $\mathbb{H}_+^n$ is operator concave for $\alpha \in (0, 1]$. That is, if $X, Y \succeq 0$ and $\theta \in [0, 1]$, then $(\theta X + (1-\theta)Y)^\alpha \succeq \theta X^\alpha + (1-\theta)Y^\alpha$.*

Lemma 17. *For $X \in \mathbb{R}^{m\times n}$ and positive semidefinite $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$,*

$$\operatorname{tr}(X^\top AXB) \;\le\; \|X\|_{\mathrm{op}}^2 \|A\|_{\mathrm{op}}\|B\|_* \quad \textit{and} \quad \operatorname{tr}(X^\top AXB) \;\le\; \|X\|_{\mathrm{op}}^2 \|B\|_{\mathrm{op}}\|A\|_*,$$

where $\|\cdot\|_*$ denotes the nuclear norm.

Proof. We prove the first inequality. By the duality of $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_*$,

$$\operatorname{tr}(X^\top AXB) = \langle X^\top AX,\ B\rangle_{\mathrm{F}} \le \|X^\top AX\|_{\mathrm{op}}\, \|B\|_*.$$

By submultiplicativity of the operator norm,

$$\|X^\top AX\|_{\mathrm{op}} \le \|X^\top\|_{\mathrm{op}}\, \|A\|_{\mathrm{op}}\, \|X\|_{\mathrm{op}} = \|X\|_{\mathrm{op}}^2\, \|A\|_{\mathrm{op}},$$

giving the first inequality. The second inequality follows similarly by writing

$$\operatorname{tr}(X^\top AXB) = \operatorname{tr}(AXBX^\top) = \langle A,\ XBX^\top\rangle_{\mathrm{F}} \le \|A\|_*\, \|XBX^\top\|_{\mathrm{op}},$$

and

$$\|XBX^\top\|_{\mathrm{op}} \le \|X\|_{\mathrm{op}}\, \|B\|_{\mathrm{op}}\, \|X^\top\|_{\mathrm{op}} = \|X\|_{\mathrm{op}}^2\, \|B\|_{\mathrm{op}}.$$

□

D Proof of lemmas

Proof of Lemma 2. Since $H_t - H_{t-1} \succeq 0$,

$$\sum_{t=2}^{T} \left(\|x_t - x^*\|_{H_t}^2 - \|x_t - x^*\|_{H_{t-1}}^2\right) = \sum_{t=2}^{T} (x_t - x^*)^\top (H_t - H_{t-1})(x_t - x^*)$$
$$\le \sum_{t=2}^{T} \|x_t - x^*\|_2^2\, \operatorname{tr}(H_t - H_{t-1}) \;\le D_2^2(\operatorname{tr}(H_T) - \operatorname{tr}(H_1)).$$

Combining with $\|x_1 - x^*\|_{H_1}^2 \le D_2^2 \operatorname{tr}(H_1)$ and Lemma 1 gives the desired result. □

Proof of Lemma 10. $H_T, \ldots, H_1 \succ 0$ by Lemma 13, thus we begin from Lemma 1.

$$\begin{aligned}
&\sum_{t=2}^{T} \left(\|x_t - x^*\|_{H_t}^2 - \|x_t - x^*\|_{H_{t-1}}^2\right) \\
&= \sum_{t=2}^{T} (x_t - x^*)^\top (A_t \otimes B_t - A_{t-1} \otimes B_{t-1})(x_t - x^*) \\
&= \sum_{t=2}^{T} (x_t - x^*)^\top \left((A_t - A_{t-1}) \otimes B_t + A_{t-1} \otimes (B_t - B_{t-1})\right)(x_t - x^*) \\
&\le \sum_{t=2}^{T} (x_t - x^*)^\top \left((A_t - A_{t-1}) \otimes B_T + A_T \otimes (B_t - B_{t-1})\right)(x_t - x^*) \\
&= \sum_{t=2}^{T} \operatorname{tr}\left((X_t - X^*)^\top (A_t - A_{t-1})(X_t - X^*)B_T\right) \\
&\qquad\qquad + \sum_{t=2}^{T} \operatorname{tr}\left((X_t - X^*)^\top A_T (X_t - X^*)(B_t - B_{t-1})\right) \\
&\le \lambda_{\max}(B_T) \sum_{t=2}^{T} \|X_t - X^*\|_{\mathrm{op}}^2 \operatorname{tr}(A_t - A_{t-1}) \\
&\qquad\qquad + \lambda_{\max}(A_T) \sum_{t=2}^{T} \|X_t - X^*\|_{\mathrm{op}}^2 \operatorname{tr}(B_t - B_{t-1}) \\
&\le D_{\mathrm{op}}^2 \lambda_{\max}(B_T)(\operatorname{tr}(A_T) - \operatorname{tr}(A_1)) + D_{\mathrm{op}}^2 \lambda_{\max}(A_T)(\operatorname{tr}(B_T) - \operatorname{tr}(B_1))
\end{aligned}$$

where the second to last inequality is due to Lemma 17. Moreover,

$$\begin{aligned}\|x_1 - x^*\|_{H_1}^2 = \operatorname{tr}\left((X_1 - X^*)^\top A_1 (X_1 - X^*) B_1\right) &\le D_{\mathrm{op}}^2 \lambda_{\max}(A_1)\operatorname{tr}(B_1) \\ &\le D_{\mathrm{op}}^2 \lambda_{\max}(A_T)\operatorname{tr}(B_1)\end{aligned}$$

Combining these two inequalities with Lemma 1 gives the desired result. □

E Proof for the regret bound of SHAMPOO

Proof of Theorem 6. Since $L_T^{1/4} \succeq \cdots \succeq L_1^{1/4} \succ 0$, $R_T^{1/4} \succeq \cdots \succeq R_1^{1/4} \succ 0$, we begin from the conclusion of Lemma 10. It only remains to bound the second term $\frac{\eta}{2}\sum_{t=1}^T (\|g_t\|^*_{L_t^{1/4}\otimes R_t^{1/4}})^2$. By Cauchy–Schwarz inequality for Frobenius inner products and for sequences,

$$\begin{aligned}\sum_{t=1}^T (\|g_t\|^*_{L_t^{1/4}\otimes R_t^{1/4}})^2 &= \sum_{t=1}^T \operatorname{tr}(G_t^\top L_t^{-1/4} G_t R_t^{-1/4}) \\ &\le \sum_{t=1}^T \|L_t^{-1/4} G_t\|_F \|G_t R_t^{-1/4}\|_F \\ &= \sum_{t=1}^T \operatorname{tr}(L_t^{-1/2} G_t G_t^\top)^{1/2} \operatorname{tr}(R_t^{-1/2} G_t^\top G_t)^{1/2} \\ &\le \left[\sum_{t=1}^T \operatorname{tr}(L_t^{-1/2} G_t G_t^\top)\right]^{1/2} \left[\sum_{t=1}^T \operatorname{tr}(R_t^{-1/2} G_t^\top G_t)\right]^{1/2}\end{aligned}$$

By Lemma 16 the map $\operatorname{tr}(A^{1/2})$ is concave on $\mathbb{H}_+^n$. Thus

$$\sum_{t=1}^T \operatorname{tr}(L_t^{-1/2} G_t G_t^\top) \le 2\sum_{t=1}^T \left(\operatorname{tr}(L_t^{1/2}) - \operatorname{tr}(L_{t-1}^{1/2})\right) \le 2\,\operatorname{tr}(L_T^{1/2})$$

and similarly for R_t, which leads to the inequality

$$\sum_{t=1}^T (\|g_t\|^*_{L_t^{1/4}\otimes R_t^{1/4}})^2 \le 2\operatorname{tr}(L_T^{1/2})^{1/2} \operatorname{tr}(R_T^{1/2})^{1/2}.$$

Combining this inequality with Lemma 10, we obtain the desired bound. □

Proof of Corollary 7. By AM-GM, we can bound the second term in Theorem 6 by

$$\begin{aligned}\operatorname{tr}(L_T^{1/2})^{1/2} \operatorname{tr}(R_T^{1/2})^{1/2} &\le \left(\lambda_{\max}(L_T^{1/4})\operatorname{tr}(L_T^{1/4}) \lambda_{\max}(R_T^{1/4})\operatorname{tr}(R_T^{1/4})\right)^{1/2} \\ &\le \frac{1}{2}\left(\lambda_{\max}(L_T^{1/4})\operatorname{tr}(R_T^{1/4}) + \operatorname{tr}(L_T^{1/4})\lambda_{\max}(R_T^{1/4})\right).\end{aligned}$$

Plugging in $\eta = D_{\mathrm{op}}$ in Theorem 6 gives the desired bound. □

F Extension to blockwise SHAMPOO

Blockwise SHAMPOO updates each block $X^{(b)}$ independently using only that block's gradient statistics, so each block is an independent instance of two-sided SHAMPOO on the linearized losses $\langle G_t^{(b)}, \cdot\rangle$. Since the linearized regret decomposes exactly across blocks, we can apply Theorem 6 to each block. With a shared step size η and $\bar{D} := \max_b D_{\mathrm{op}}^{(b)}$, the AM-GM step in the proof of Corollary 7 gives

$$\mathcal{R}_T \le \left(\frac{\bar{D}^2}{2\eta} + \frac{\eta}{2}\right) \sum_b \left(\lambda_{\max}\big((L_T^{(b)})^{1/4}\big)\operatorname{tr}\big((R_T^{(b)})^{1/4}\big) + \lambda_{\max}\big((R_T^{(b)})^{1/4}\big)\operatorname{tr}\big((L_T^{(b)})^{1/4}\big)\right),$$

where the superscript (b) denotes the quantities of block b. Choosing $\eta = \bar{D}$ yields

$$\mathcal{R}_T \le \bar{D} \sum_b \left(\lambda_{\max}\big((L_T^{(b)})^{1/4}\big)\operatorname{tr}\big((R_T^{(b)})^{1/4}\big) + \lambda_{\max}\big((R_T^{(b)})^{1/4}\big)\operatorname{tr}\big((L_T^{(b)})^{1/4}\big)\right).$$

G Proof for the regret bound of properly scaled SHAMPOO^2

Proof of Theorem 11. Throughout this proof we write $\bar{L}_t := \epsilon^{-1} L_t$ and $\bar{R}_t := \epsilon^{-1} R_t$. Since $L_t \succeq \epsilon I_m$ and $R_t \succeq \epsilon I_n$, both $\mathrm{tr}(\log \bar{L}_t)$ and $\mathrm{tr}(\log \bar{R}_t)$ are nonnegative.

By Lemma 1, we have

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|_{H_1}^2 + \frac{1}{2\eta}\sum_{t=2}^{T}\left(\|x_t - x^*\|_{H_t}^2 - \|x_t - x^*\|_{H_{t-1}}^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}(\|g_t\|_{H_t}^*)^2$$

We first aim to bound the second term.

$$\begin{aligned}
&\sum_{t=2}^{T}(\|x_t - x^*\|_{H_t}^2 - \|x_t - x^*\|_{H_{t-1}}^2) \\
&= \sum_{t=2}^{T}(x_t - x^*)^\top(\alpha_t\, L_t^{1/2} \otimes R_t^{1/2} - \alpha_{t-1}\, L_{t-1}^{1/2} \otimes R_{t-1}^{1/2})(x_t - x^*) \\
&= \sum_{t=2}^{T}(x_t - x^*)^\top\left(\alpha_t(L_t^{1/2} \otimes R_t^{1/2} - L_{t-1}^{1/2} \otimes R_{t-1}^{1/2}) + (\alpha_t - \alpha_{t-1})L_{t-1}^{1/2} \otimes R_{t-1}^{1/2}\right)(x_t - x^*) \\
&= \sum_{t=2}^{T}\alpha_t(x_t - x^*)^\top\left(L_t^{1/2} \otimes (R_t^{1/2} - R_{t-1}^{1/2}) + (L_t^{1/2} - L_{t-1}^{1/2}) \otimes R_{t-1}^{1/2}\right)(x_t - x^*) \\
&\quad + \sum_{t=2}^{T}(\alpha_t - \alpha_{t-1})(x_t - x^*)^\top\left(L_{t-1}^{1/2} \otimes R_{t-1}^{1/2}\right)(x_t - x^*) \\
&\leq \sum_{t=2}^{T}\alpha_t\|X_t - X^*\|_{\mathrm{op}}^2\|L_t^{1/2}\|_{\mathrm{op}}\|R_t^{1/2} - R_{t-1}^{1/2}\|_* + \sum_{t=2}^{T}\alpha_t\|X_t - X^*\|_{\mathrm{op}}^2\|R_{t-1}^{1/2}\|_{\mathrm{op}}\|L_t^{1/2} - L_{t-1}^{1/2}\|_* \\
&\quad + \sum_{t=2}^{T}|\alpha_t - \alpha_{t-1}|\,\|X_t - X^*\|_{\mathrm{op}}^2\left(\frac{\|L_{t-1}^{1/2}\|_{\mathrm{op}}\|R_{t-1}^{1/2}\|_* + \|R_{t-1}^{1/2}\|_{\mathrm{op}}\|L_{t-1}^{1/2}\|_*}{2}\right) \\
&\leq D_{\mathrm{op}}^2\sum_{t=2}^{T}\alpha_t\lambda_{\max}(L_t^{1/2})\,\mathrm{tr}(R_t^{1/2} - R_{t-1}^{1/2}) + D_{\mathrm{op}}^2\sum_{t=2}^{T}\alpha_t\lambda_{\max}(R_{t-1}^{1/2})\,\mathrm{tr}(L_t^{1/2} - L_{t-1}^{1/2}) \\
&\quad + D_{\mathrm{op}}^2\sum_{t=2}^{T}|\alpha_t - \alpha_{t-1}|\left(\frac{\lambda_{\max}(L_{t-1}^{1/2})\,\mathrm{tr}(R_{t-1}^{1/2}) + \lambda_{\max}(R_{t-1}^{1/2})\,\mathrm{tr}(L_{t-1}^{1/2})}{2}\right) \qquad (2)
\end{aligned}$$

where the second to last inequality is due to Lemma 17.

With $s_t := \lambda_{\max}(L_t^{1/2})\mathrm{tr}(R_t^{1/2}) + \lambda_{\max}(R_t^{1/2})\mathrm{tr}(L_t^{1/2})$, the first two terms in (2) satisfy

$$\begin{aligned}
&D_{\mathrm{op}}^2\sum_{t=2}^{T}\alpha_t\lambda_{\max}(L_t^{1/2})\,\mathrm{tr}(R_t^{1/2} - R_{t-1}^{1/2}) + D_{\mathrm{op}}^2\sum_{t=2}^{T}\alpha_t\lambda_{\max}(R_{t-1}^{1/2})\,\mathrm{tr}(L_t^{1/2} - L_{t-1}^{1/2}) \\
&\leq D_{\mathrm{op}}^2\sum_{t=2}^{T}\alpha_t\left(\lambda_{\max}(L_t^{1/2})\,\mathrm{tr}(R_t^{1/2}) - \lambda_{\max}(L_{t-1}^{1/2})\,\mathrm{tr}(R_{t-1}^{1/2}) + \lambda_{\max}(R_t^{1/2})\,\mathrm{tr}(L_t^{1/2}) - \lambda_{\max}(R_{t-1}^{1/2})\,\mathrm{tr}(L_{t-1}^{1/2})\right) \\
&= D_{\mathrm{op}}^2\sum_{t=2}^{T}\frac{\mathrm{tr}(\log\bar{L}_t)^{1/4}\mathrm{tr}(\log\bar{R}_t)^{1/4}}{s_t^{1/2}}(s_t - s_{t-1}) \\
&\leq D_{\mathrm{op}}^2\sum_{t=2}^{T}\mathrm{tr}(\log\bar{L}_t)^{1/4}\mathrm{tr}(\log\bar{R}_t)^{1/4}2(s_t^{1/2} - s_{t-1}^{1/2}) \\
&\leq D_{\mathrm{op}}^2\,\mathrm{tr}(\log\bar{L}_T)^{1/4}\mathrm{tr}(\log\bar{R}_T)^{1/4}\sum_{t=2}^{T}2(s_t^{1/2} - s_{t-1}^{1/2}) \\
&= 2D_{\mathrm{op}}^2\,\mathrm{tr}(\log\bar{L}_T)^{1/4}\mathrm{tr}(\log\bar{R}_T)^{1/4}(s_T^{1/2} - s_1^{1/2})\,.
\end{aligned}$$

The last term of equation (2) can be bounded as

$$
\begin{aligned}
&D_{\mathrm{op}}^2 \sum_{t=2}^{T} |\alpha_t - \alpha_{t-1}| \left(\frac{\lambda_{\max}(L_{t-1}^{1/2}) \operatorname{tr}(R_{t-1}^{1/2}) + \lambda_{\max}(R_{t-1}^{1/2}) \operatorname{tr}(L_{t-1}^{1/2})}{2} \right) \\
&= D_{\mathrm{op}}^2 \sum_{t=2}^{T} \left| \frac{\operatorname{tr}(\log \bar{L}_t)^{1/4} \operatorname{tr}(\log \bar{R}_t)^{1/4}}{s_t^{1/2}} - \frac{\operatorname{tr}(\log \bar{L}_{t-1})^{1/4} \operatorname{tr}(\log \bar{R}_{t-1})^{1/4}}{s_{t-1}^{1/2}} \right| \frac{s_{t-1}}{2} \\
&\le D_{\mathrm{op}}^2 \sum_{t=2}^{T} \left| \frac{\operatorname{tr}(\log \bar{L}_t)^{1/4} \operatorname{tr}(\log \bar{R}_t)^{1/4}}{s_t^{1/2}} - \frac{\operatorname{tr}(\log \bar{L}_t)^{1/4} \operatorname{tr}(\log \bar{R}_t)^{1/4}}{s_{t-1}^{1/2}} \right| \frac{s_{t-1}}{2} \\
&\quad + D_{\mathrm{op}}^2 \sum_{t=2}^{T} \left| \frac{\operatorname{tr}(\log \bar{L}_t)^{1/4} \operatorname{tr}(\log \bar{R}_t)^{1/4}}{s_{t-1}^{1/2}} - \frac{\operatorname{tr}(\log \bar{L}_{t-1})^{1/4} \operatorname{tr}(\log \bar{R}_{t-1})^{1/4}}{s_{t-1}^{1/2}} \right| \frac{s_{t-1}}{2} \\
&\le D_{\mathrm{op}}^2 \operatorname{tr}(\log \bar{L}_T)^{1/4} \operatorname{tr}(\log \bar{R}_T)^{1/4} \sum_{t=2}^{T} \left(\frac{1}{s_{t-1}^{1/2}} - \frac{1}{s_t^{1/2}} \right) \frac{s_{t-1}^{1/2} s_t^{1/2}}{2} \\
&\quad + \frac{1}{2} D_{\mathrm{op}}^2 s_T^{1/2} \sum_{t=2}^{T} \left(\operatorname{tr}(\log \bar{L}_t)^{1/4} \operatorname{tr}(\log \bar{R}_t)^{1/4} - \operatorname{tr}(\log \bar{L}_{t-1})^{1/4} \operatorname{tr}(\log \bar{R}_{t-1})^{1/4} \right) \\
&\le \frac{1}{2} D_{\mathrm{op}}^2 \operatorname{tr}(\log \bar{L}_T)^{1/4} \operatorname{tr}(\log \bar{R}_T)^{1/4} (s_T^{1/2} - s_1^{1/2}) \\
&\quad + \frac{1}{2} D_{\mathrm{op}}^2 s_T^{1/2} \left(\operatorname{tr}(\log \bar{L}_T)^{1/4} \operatorname{tr}(\log \bar{R}_T)^{1/4} - \operatorname{tr}(\log \bar{L}_1)^{1/4} \operatorname{tr}(\log \bar{R}_1)^{1/4} \right)
\end{aligned}
$$

By Lemma 17, the first term satisfies $\|x_1 - x^*\|_{H_1}^2 \le \frac{1}{2} D_{\mathrm{op}}^2 s_1^{1/2} \operatorname{tr}(\log \bar{L}_1)^{1/4} \operatorname{tr}(\log \bar{R}_1)^{1/4}$.

Combining, so far we have

$$
\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \le \frac{1}{2\eta} \cdot 3 D_{\mathrm{op}}^2 \operatorname{tr}(\log \bar{L}_T)^{1/4} \operatorname{tr}(\log \bar{R}_T)^{1/4} s_T^{1/2} + \frac{\eta}{2} \sum_{t=1}^{T} (\|g_t\|_{H_t}^*)^2,
$$

and it only remains to bound the last term. By Cauchy–Schwarz inequality for Frobenius inner products and for sequences,

$$
\begin{aligned}
\sum_{t=1}^{T} (\|g_t\|_{H_t}^*)^2 &= \sum_{t=1}^{T} \alpha_t^{-1} \operatorname{tr}(G_t^\top L_t^{-1/2} G_t R_t^{-1/2}) \\
&\le \sum_{t=1}^{T} \alpha_t^{-1} \|L_t^{-1/2} G_t\|_F \|G_t R_t^{-1/2}\|_F \\
&= \sum_{t=1}^{T} \frac{s_t^{1/2}}{\operatorname{tr}(\log \bar{L}_t)^{1/4} \operatorname{tr}(\log \bar{R}_t)^{1/4}} \operatorname{tr}(L_t^{-1} G_t G_t^\top)^{1/2} \operatorname{tr}(R_t^{-1} G_t^\top G_t)^{1/2} \\
&\le s_T^{1/2} \sum_{t=1}^{T} \left[\frac{\operatorname{tr}(L_t^{-1} G_t G_t^\top)}{\operatorname{tr}(\log \bar{L}_t)^{1/2}} \right]^{1/2} \left[\frac{\operatorname{tr}(R_t^{-1} G_t^\top G_t)}{\operatorname{tr}(\log \bar{R}_t)^{1/2}} \right]^{1/2} \\
&\le s_T^{1/2} \left[\sum_{t=1}^{T} \frac{\operatorname{tr}(L_t^{-1} G_t G_t^\top)}{\operatorname{tr}(\log \bar{L}_t)^{1/2}} \right]^{1/2} \left[\sum_{t=1}^{T} \frac{\operatorname{tr}(R_t^{-1} G_t^\top G_t)}{\operatorname{tr}(\log \bar{R}_t)^{1/2}} \right]^{1/2}
\end{aligned}
$$

The map $A \mapsto \operatorname{tr}(\log A) = \log\det A$ is concave on positive definite matrices. Thus

$$
\operatorname{tr}(L_t^{-1} G_t G_t^\top) \le \operatorname{tr}(\log \bar{L}_t) - \operatorname{tr}(\log \bar{L}_{t-1})
$$

and a telescoping argument yields

$$
\sum_{t=1}^{T} \frac{\operatorname{tr}(L_t^{-1} G_t G_t^\top)}{\operatorname{tr}(\log \bar{L}_t)^{1/2}} \le 2 \operatorname{tr}(\log \bar{L}_T)^{1/2}
$$

and similarly for R_t, which leads to the inequality

$$\sum_{t=1}^{T}(\|g_t\|_{H_t}^*)^2 \leq 2\,\mathrm{tr}(\log \bar{L}_T)^{1/4}\mathrm{tr}(\log \bar{R}_T)^{1/4}s_T^{1/2}$$

Combining this inequality with the previous inequalities, we obtain

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \leq \left(\frac{3}{2\eta}D_{\mathrm{op}}^2 + \eta\right)\,\mathrm{tr}(\log \bar{L}_T)^{1/4}\mathrm{tr}(\log \bar{R}_T)^{1/4}s_T^{1/2}\,.$$

Choosing $\eta = \sqrt{\frac{3}{2}}D_{\mathrm{op}}$ gives the bound

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \leq \sqrt{6}D_{\mathrm{op}}\left(\lambda_{\max}(L_T^{1/2})\mathrm{tr}(R_T^{1/2}) + \lambda_{\max}(R_T^{1/2})\mathrm{tr}(L_T^{1/2})\right)^{1/2} \cdot \mathrm{tr}(\log \bar{L}_T)^{1/4}\mathrm{tr}(\log \bar{R}_T)^{1/4}\,.$$

□

H Proof for the regret bound of SHAMPOO^{4p} with scaling

Proof of Theorem 12. By Lemma 1, we have

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \leq \frac{1}{2\eta}\|x_1 - x^*\|_{H_1}^2 + \frac{1}{2\eta}\sum_{t=2}^{T}\left(\|x_t - x^*\|_{H_t}^2 - \|x_t - x^*\|_{H_{t-1}}^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}(\|g_t\|_{H_t}^*)^2.$$

We first aim to bound the second term.

$$\begin{aligned}
&\sum_{t=2}^{T}\left(\|x_t - x^*\|_{H_t}^2 - \|x_t - x^*\|_{H_{t-1}}^2\right)\\
&= \sum_{t=2}^{T}(x_t - x^*)^\top(\alpha_t\, L_t^p \otimes R_t^p - \alpha_{t-1}\, L_{t-1}^p \otimes R_{t-1}^p)(x_t - x^*)\\
&= \sum_{t=2}^{T}(x_t - x^*)^\top\Big(\alpha_t(L_t^p \otimes R_t^p - L_{t-1}^p \otimes R_{t-1}^p) + (\alpha_t - \alpha_{t-1})L_{t-1}^p \otimes R_{t-1}^p\Big)(x_t - x^*)\\
&= \sum_{t=2}^{T}\alpha_t(x_t - x^*)^\top\Big(L_t^p \otimes (R_t^p - R_{t-1}^p) + (L_t^p - L_{t-1}^p) \otimes R_{t-1}^p\Big)(x_t - x^*)\\
&\qquad + \sum_{t=2}^{T}(\alpha_t - \alpha_{t-1})(x_t - x^*)^\top\Big(L_{t-1}^p \otimes R_{t-1}^p\Big)(x_t - x^*)\\
&\leq \sum_{t=2}^{T}\alpha_t\|X_t - X^*\|_{\mathrm{op}}^2\|L_t^p\|_{\mathrm{op}}\|R_t^p - R_{t-1}^p\|_* + \sum_{t=2}^{T}\alpha_t\|X_t - X^*\|_{\mathrm{op}}^2\|R_{t-1}^p\|_{\mathrm{op}}\|L_t^p - L_{t-1}^p\|_*\\
&\qquad + \sum_{t=2}^{T}|\alpha_t - \alpha_{t-1}|\,\|X_t - X^*\|_{\mathrm{op}}^2\left(\frac{\|L_{t-1}^p\|_{\mathrm{op}}\|R_{t-1}^p\|_* + \|R_{t-1}^p\|_{\mathrm{op}}\|L_{t-1}^p\|_*}{2}\right)\\
&\leq D_{\mathrm{op}}^2\sum_{t=2}^{T}\alpha_t\lambda_{\max}(L_t^p)\,\mathrm{tr}(R_t^p - R_{t-1}^p) + D_{\mathrm{op}}^2\sum_{t=2}^{T}\alpha_t\lambda_{\max}(R_{t-1}^p)\,\mathrm{tr}(L_t^p - L_{t-1}^p)\\
&\qquad + D_{\mathrm{op}}^2\sum_{t=2}^{T}|\alpha_t - \alpha_{t-1}|\left(\frac{\lambda_{\max}(L_{t-1}^p)\,\mathrm{tr}(R_{t-1}^p) + \lambda_{\max}(R_{t-1}^p)\,\mathrm{tr}(L_{t-1}^p)}{2}\right) \qquad (3)
\end{aligned}$$

where the second to last inequality is due to Lemma 17.

Defining

$$s_t := \lambda_{\max}(L_t^p)\mathrm{tr}(R_t^p) + \lambda_{\max}(R_t^p)\mathrm{tr}(L_t^p), \qquad \alpha_t = \frac{\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}{s_t^{1/2}},$$

for convenience, the first two terms in (3) satisfy

$$\begin{aligned}
&D_{\mathrm{op}}^2 \sum_{t=2}^{T} \alpha_t \lambda_{\max}(L_t^p)\,\mathrm{tr}(R_t^p - R_{t-1}^p) + D_{\mathrm{op}}^2 \sum_{t=2}^{T} \alpha_t \lambda_{\max}(R_{t-1}^p)\,\mathrm{tr}(L_t^p - L_{t-1}^p) \\
&\le D_{\mathrm{op}}^2 \sum_{t=2}^{T} \alpha_t \Big(\lambda_{\max}(L_t^p)\,\mathrm{tr}(R_t^p) - \lambda_{\max}(L_{t-1}^p)\,\mathrm{tr}(R_{t-1}^p) + \lambda_{\max}(R_t^p)\,\mathrm{tr}(L_t^p) - \lambda_{\max}(R_{t-1}^p)\,\mathrm{tr}(L_{t-1}^p) \Big) \\
&= D_{\mathrm{op}}^2 \sum_{t=2}^{T} \frac{\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}{s_t^{1/2}} (s_t - s_{t-1}) \\
&\le D_{\mathrm{op}}^2 \sum_{t=2}^{T} \mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}\, 2(s_t^{1/2} - s_{t-1}^{1/2}) \\
&\le D_{\mathrm{op}}^2\, \mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4} \sum_{t=2}^{T} 2(s_t^{1/2} - s_{t-1}^{1/2}) \\
&= 2D_{\mathrm{op}}^2\, \mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4}(s_T^{1/2} - s_1^{1/2}).
\end{aligned}$$

The last term of (3) can be bounded as

$$\begin{aligned}
&D_{\mathrm{op}}^2 \sum_{t=2}^{T} |\alpha_t - \alpha_{t-1}| \left(\frac{\lambda_{\max}(L_{t-1}^p)\,\mathrm{tr}(R_{t-1}^p) + \lambda_{\max}(R_{t-1}^p)\,\mathrm{tr}(L_{t-1}^p)}{2} \right) \\
&= D_{\mathrm{op}}^2 \sum_{t=2}^{T} \left| \frac{\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}{s_t^{1/2}} - \frac{\mathrm{tr}(L_{t-1}^{1-2p})^{1/4}\mathrm{tr}(R_{t-1}^{1-2p})^{1/4}}{s_{t-1}^{1/2}} \right| \frac{s_{t-1}}{2} \\
&\le D_{\mathrm{op}}^2 \sum_{t=2}^{T} \left| \frac{\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}{s_t^{1/2}} - \frac{\mathrm{tr}(L_{t-1}^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}{s_{t-1}^{1/2}} \right| \frac{s_{t-1}}{2} \\
&\quad + D_{\mathrm{op}}^2 \sum_{t=2}^{T} \left| \frac{\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}{s_{t-1}^{1/2}} - \frac{\mathrm{tr}(L_{t-1}^{1-2p})^{1/4}\mathrm{tr}(R_{t-1}^{1-2p})^{1/4}}{s_{t-1}^{1/2}} \right| \frac{s_{t-1}}{2} \\
&\le D_{\mathrm{op}}^2 \mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4} \sum_{t=2}^{T} \left(\frac{1}{s_{t-1}^{1/2}} - \frac{1}{s_t^{1/2}} \right) \frac{s_{t-1}^{1/2} s_t^{1/2}}{2} \\
&\quad + \frac{1}{2} D_{\mathrm{op}}^2 s_T^{1/2} \sum_{t=2}^{T} \Big(\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4} - \mathrm{tr}(L_{t-1}^{1-2p})^{1/4}\mathrm{tr}(R_{t-1}^{1-2p})^{1/4} \Big) \\
&\le \frac{1}{2} D_{\mathrm{op}}^2 \mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4}(s_T^{1/2} - s_1^{1/2}) \\
&\quad + \frac{1}{2} D_{\mathrm{op}}^2 s_T^{1/2} \Big(\mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4} - \mathrm{tr}(L_1^{1-2p})^{1/4}\mathrm{tr}(R_1^{1-2p})^{1/4} \Big)
\end{aligned}$$

By Lemma 17, the first term satisfies $\|x_1 - x^*\|_{H_1}^2 \le \frac{1}{2} D_{\mathrm{op}}^2 s_1^{1/2} \mathrm{tr}(L_1^{1-2p})^{1/4}\mathrm{tr}(R_1^{1-2p})^{1/4}$.

Combining, so far we have

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*) \le \frac{1}{2\eta} \cdot 3D_{\mathrm{op}}^2\, \mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4} s_T^{1/2} + \frac{\eta}{2} \sum_{t=1}^{T} (\|g_t\|_{H_t}^*)^2,$$

and it only remains to bound the last term.

By Cauchy–Schwarz inequality for Frobenius inner products and for sequences,

$$\begin{aligned}
\sum_{t=1}^{T}(\|g_t\|_{H_t}^*)^2 &= \sum_{t=1}^{T}\alpha_t^{-1}\,\mathrm{tr}(G_t^\top L_t^{-p}G_t R_t^{-p}) \\
&\le \sum_{t=1}^{T}\alpha_t^{-1}\|L_t^{-p}G_t\|_F\|G_t R_t^{-p}\|_F \\
&= \sum_{t=1}^{T}\frac{s_t^{1/2}}{\mathrm{tr}(L_t^{1-2p})^{1/4}\mathrm{tr}(R_t^{1-2p})^{1/4}}\mathrm{tr}(L_t^{-2p}G_tG_t^\top)^{1/2}\mathrm{tr}(R_t^{-2p}G_t^\top G_t)^{1/2} \\
&\le s_T^{1/2}\sum_{t=1}^{T}\left[\frac{\mathrm{tr}(L_t^{-2p}G_tG_t^\top)}{\mathrm{tr}(L_t^{1-2p})^{1/2}}\right]^{1/2}\left[\frac{\mathrm{tr}(R_t^{-2p}G_t^\top G_t)}{\mathrm{tr}(R_t^{1-2p})^{1/2}}\right]^{1/2} \\
&\le s_T^{1/2}\left[\sum_{t=1}^{T}\frac{\mathrm{tr}(L_t^{-2p}G_tG_t^\top)}{\mathrm{tr}(L_t^{1-2p})^{1/2}}\right]^{1/2}\left[\sum_{t=1}^{T}\frac{\mathrm{tr}(R_t^{-2p}G_t^\top G_t)}{\mathrm{tr}(R_t^{1-2p})^{1/2}}\right]^{1/2}.
\end{aligned}$$

By Lemma 16, the map $\mathrm{tr}(A^{1-2p})$ is concave on $\mathbb{H}_+^n$ for $p \in (0, \frac{1}{2})$. Thus

$$(1-2p)\ \mathrm{tr}(L_t^{-2p}G_tG_t^\top) \le \mathrm{tr}(L_t^{1-2p}) - \mathrm{tr}(L_{t-1}^{1-2p}),$$

and a telescoping argument yields

$$\sum_{t=1}^{T}\frac{\mathrm{tr}(L_t^{-2p}G_tG_t^\top)}{\mathrm{tr}(L_t^{1-2p})^{1/2}} \le \frac{2}{1-2p}\ \mathrm{tr}(L_T^{1-2p})^{1/2},$$

and similarly for R_t. Therefore,

$$\sum_{t=1}^{T}(\|g_t\|_{H_t}^*)^2 \le \frac{2}{1-2p}\ \mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4}s_T^{1/2}.$$

Combining this inequality with the previous inequalities, we obtain

$$\sum_{t=1}^{T}f_t(x_t) - \sum_{t=1}^{T}f_t(x^*) \le \left(\frac{3}{2\eta}D_{\mathrm{op}}^2 + \frac{\eta}{1-2p}\right)\ \mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4}s_T^{1/2}.$$

Choosing $\eta = \sqrt{\frac{3(1-2p)}{2}}D_{\mathrm{op}}$ gives

$$\sum_{t=1}^{T}f_t(x_t) - \sum_{t=1}^{T}f_t(x^*) \le \sqrt{\frac{6}{1-2p}}\,D_{\mathrm{op}}\ s_T^{1/2}\mathrm{tr}(L_T^{1-2p})^{1/4}\mathrm{tr}(R_T^{1-2p})^{1/4}.$$

Recalling the definition of s_T completes the proof. □

I Proof of propositions regarding bound comparisons

Proof of Proposition 4. Let $S_T = \epsilon I_d + \sum_{t=1}^{T} g_t g_t^\top$ and $H_T = L_T^{1/4} \otimes R_T^{1/4}$.

By Gupta et al. [2018, Lemma 8], we have $S_T^{1/2} \preceq \sqrt{r_G}\, H_T$ and

$$\mathrm{tr}(S_T^{1/2}) \le \sqrt{r_G}\ \mathrm{tr}(L_T^{1/4})\ \mathrm{tr}(R_T^{1/4}),$$

hence $\mathcal{B}_{\text{AdaGrad}} \le \mathcal{B}_{\text{Shampoo}}^{(\textit{Gupta et al.})}$.

By subadditivity of $x \mapsto x^{1/4}$ on $[0,\infty)$, $\mathrm{tr}(L_T)^{1/4} \le \mathrm{tr}(L_T^{1/4})$ and $\mathrm{tr}(R_T)^{1/4} \le \mathrm{tr}(R_T^{1/4})$. Since $\mathrm{tr}(L_T) = \mathrm{tr}(R_T) = \mathrm{tr}(S_T)$, we get

$$\mathrm{tr}(S_T)^{1/2} \le \mathrm{tr}(L_T^{1/4})\ \mathrm{tr}(R_T^{1/4}).$$

Multiplying by $\sqrt{2}D_2\sqrt{r_G}$ yields $\sqrt{r_G}\,\mathcal{B}_{\text{AdaGrad-Norm}} \le \mathcal{B}_{\text{Shampoo}}^{(\textit{Gupta et al.})}$. □

Proof of Proposition 8. Consider the ratio

$$\frac{\mathcal{B}^{(\textit{Gupta et al.})}_{\textsc{Shampoo}}}{\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}}} = \frac{\sqrt{2}\, D_2\, \sqrt{r_G}\ \operatorname{tr}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4})}{D_{\mathrm{op}}\left(\lambda_{\max}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4})\ \operatorname{tr}(L_T^{1/4})\right)}$$
$$= \frac{D_2}{D_{\mathrm{op}}}\sqrt{r_G}\cdot \frac{\sqrt{2}\ \operatorname{tr}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4})}{\lambda_{\max}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4})\ \operatorname{tr}(L_T^{1/4})}.$$

Since $L_T^{1/4} \succeq 0$ and $R_T^{1/4} \succeq 0$, we have $\lambda_{\max}(A) \le \operatorname{tr}(A)$ for any $A \succeq 0$, hence

$$\lambda_{\max}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4})\ \operatorname{tr}(L_T^{1/4}) \le 2\ \operatorname{tr}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4}).$$

Therefore the last fraction is at least $1/\sqrt{2}$, and

$$\frac{\mathcal{B}^{(\textit{Gupta et al.})}_{\textsc{Shampoo}}}{\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}}} \ge \frac{D_2}{D_{\mathrm{op}}}\sqrt{\frac{r_G}{2}}.$$

□

Proof of Proposition 9. *(i)* Take $T = 1$, $m = n \ge 3$, and $G_1 := e_1 e_1^\top \in \mathbb{R}^{m\times m}$, and let $\epsilon \to 0$. Then $L_1 = R_1 = e_1 e_1^\top$, so $\lambda_{\max}(L_1^{1/4}) = \operatorname{tr}(L_1^{1/4}) = \lambda_{\max}(R_1^{1/4}) = \operatorname{tr}(R_1^{1/4}) = 1$ and $\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}} = 2\, D_{\mathrm{op}}$. Similarly, $g_1 = \overline{\mathrm{vec}}(G_1) = e_1 \in \mathbb{R}^{m^2}$, so $S_1 = e_1 e_1^\top$ satisfies $\operatorname{tr}(S_1) = \operatorname{tr}(S_1^{1/2}) = 1$ and $\mathcal{B}_{\textsc{AdaGrad}} = \mathcal{B}_{\textsc{AdaGrad-Norm}} = \sqrt{2}\, D_2$. With the comparator $X^* = I_m$ and $X_1 = 0$, we get $D_{\mathrm{op}} = 1$ and $D_2 = \sqrt{m}$, hence

$$\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}} = 2 \ <\ \sqrt{2m} = \mathcal{B}_{\textsc{AdaGrad-Norm}} = \mathcal{B}_{\textsc{AdaGrad}} \qquad \text{since } m \ge 3.$$

(ii) Let $\epsilon \to 0$ and write $s := \operatorname{tr}(L_T) = \operatorname{tr}(R_T) = \sum_{t=1}^T \|G_t\|_{\mathrm{F}}^2$, so that $\mathcal{B}_{\textsc{One-sided Shampoo}} = \sqrt{2}\, D_{\mathrm{op}}\ \operatorname{tr}(L_T^{1/2})$. Let $\ell_1, \dots, \ell_m$ denote the eigenvalues of L_T, and write $a := \operatorname{tr}(L_T^{1/4}) = \sum_i \ell_i^{1/4}$ and $u := \lambda_{\max}(R_T^{1/4}) = \lambda_{\max}(R_T)^{1/4}$.

Lower bound. By Hölder's inequality with exponents $\frac{3}{2}$ and 3,

$$\operatorname{tr}(L_T^{1/2}) = \sum_i (\ell_i^{1/4})^{2/3}\, \ell_i^{1/3} \le \Big(\sum_i \ell_i^{1/4}\Big)^{2/3} \Big(\sum_i \ell_i\Big)^{1/3} = a^{2/3}\, s^{1/3}.$$

Since $\ell_i^{1/4} \ge \ell_i\, \lambda_{\max}(L_T)^{-3/4}$ for every i, summing over i gives $a \ge s\, \lambda_{\max}(L_T)^{-3/4}$, that is, $\lambda_{\max}(L_T^{1/4}) \ge (s/a)^{1/3}$. In the same way, $\operatorname{tr}(R_T^{1/4}) \ge s\, \lambda_{\max}(R_T)^{-3/4} = s\, u^{-3}$. Therefore

$$\lambda_{\max}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4})\ \operatorname{tr}(L_T^{1/4}) \ge \frac{s^{4/3}}{a^{1/3}u^3} + \frac{au}{3} + \frac{au}{3} + \frac{au}{3}$$
$$\ge 4\Big(\frac{s^{4/3}a^{8/3}}{27}\Big)^{1/4} = \frac{4}{3^{3/4}}\, a^{2/3} s^{1/3},$$

where the second inequality is the AM-GM inequality for four terms. Combining the two displays,

$$\frac{\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}}}{\mathcal{B}_{\textsc{One-sided Shampoo}}} = \frac{\lambda_{\max}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4})\ \operatorname{tr}(L_T^{1/4})}{\sqrt{2}\ \operatorname{tr}(L_T^{1/2})} \ge \frac{4}{\sqrt{2}\cdot 3^{3/4}} = \frac{2^{3/2}}{3^{3/4}}.$$

Flat spectra. Under the flat-spectrum assumption, the nonzero eigenvalues of L_T all equal s/r_L and those of R_T all equal s/r_R. Hence

$$\lambda_{\max}(L_T^{1/4})\ \operatorname{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4})\ \operatorname{tr}(L_T^{1/4}) = \Big(\frac{s}{r_L}\Big)^{1/4} r_R \Big(\frac{s}{r_R}\Big)^{1/4} + \Big(\frac{s}{r_R}\Big)^{1/4} r_L \Big(\frac{s}{r_L}\Big)^{1/4}$$
$$= \frac{s^{1/2}\,(r_L + r_R)}{(r_L r_R)^{1/4}},$$
$$\operatorname{tr}(L_T^{1/2}) = r_L \Big(\frac{s}{r_L}\Big)^{1/2} = s^{1/2}\, r_L^{1/2},$$

and therefore

$$\frac{\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}}}{\mathcal{B}_{\textsc{One-sided Shampoo}}} = \frac{r_L + r_R}{\sqrt{2}\, r_L^{3/4} r_R^{1/4}} = \frac{\gamma^{3/4} + \gamma^{-1/4}}{\sqrt{2}}, \qquad \gamma = \frac{r_R}{r_L}.$$

This equals $\sqrt{2}$ at $\gamma = 1$ and $\frac{3^{-3/4}+3^{1/4}}{\sqrt{2}} = \frac{2^{3/2}}{3^{3/4}}$ at $\gamma = 1/3$, so the lower bound is attained, for instance by $T = 3$ and $G_t = e_t e_1^\top$ with $m \geq 3$, for which $L_T = \sum_{t=1}^3 e_t e_t^\top$ and $R_T = 3\, e_1 e_1^\top$. $\square$

J Regret bound for scaled SHAMPOO

We also state Theorem 12 for $p = \frac{1}{4}$, that is, SHAMPOO with its scaling factor, as a corollary.

Corollary 18 (Regret of scaled SHAMPOO)**.** *Suppose we run Algorithm 3 with $p = \frac{1}{4}$, $\eta = \sqrt{\frac{3}{4}} D_{\mathrm{op}}$, $\epsilon > 0$. Then its regret compared to any $X^* \in \mathbb{R}^{m \times n}$ is bounded as follows,*

$$\mathcal{R}_T \leq\ 2\sqrt{3} D_{\mathrm{op}} \left(\lambda_{\max}(L_T^{1/4}) \mathrm{tr}(R_T^{1/4}) + \lambda_{\max}(R_T^{1/4}) \mathrm{tr}(L_T^{1/4}) \right)^{1/2} \mathrm{tr}(L_T^{1/2})^{1/4} \mathrm{tr}(R_T^{1/2})^{1/4}$$

where $D_{\mathrm{op}} = \max_{t \in [T]} \|X_t - X^*\|_{\mathrm{op}}$.

Note that by AM-GM and $\mathrm{tr}(A) \leq \lambda_{\max}(A^{1/2})\, \mathrm{tr}(A^{1/2})$, the second multiplicative factor in Corollary 18, $\sqrt{2}\, \mathrm{tr}(L_T^{1/2})^{1/4}\, \mathrm{tr}(R_T^{1/2})^{1/4}$, is at most the first factor. Hence the bound in Corollary 18 is at most $\sqrt{6}$ times the bound in Corollary 7, which we state as Proposition 19.

Proposition 19. *Let $\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}}$ and $\mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo-Scaled}}$ be the bounds of Corollaries 7 and 18. Then*

$$\frac{1}{\sqrt{6}} \mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo-Scaled}} \leq \mathcal{B}^{(\textit{Ours})}_{\textsc{Shampoo}}$$

Note that the algorithms are not the same due to the time-varying scaling factor α_t.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract is to provide and prove a sharper regret bound for the full two-sided SHAMPOO, provide the first regret guarantees for SHAMPOO^2, and more broadly, SHAMPOO^{4p}. We summarized the main contributions of this paper again in Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of this work in the **Conclusion**.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide the assumptions for our theoretical results in Section 3, Section 4 and Section 5. We provide the proofs for every main result in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide numerical experimental results in Section 6 and the detailed information of experiments in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not yet provide open access to the code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed explanations of the experimental setting in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our experimental result is from 80 random runs per method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the specific computer resources required to reproduce the experiments in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper adheres to the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no expected negative societal impacts of the work since this research focuses on the theoretical aspects.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The research conducted in this paper does not pose any such risks.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research did not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.