ADAPTIVE REGULARIZATION THROUGH COUPLED KRONECKER FACTORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive regularization based optimization algorithms have been widely used in deep learning. While diagonal preconditioners such as Adam and Adagrad are scalable for large numbers of parameters, methods utilizing Kronecker factoring have recently found success. Notably, Shampoo, CASPR, and SOAP share identical axes preconditioners which are developed independently to optimize regret. However, there remain theoretical gaps regarding how to develop these preconditioner factors jointly and in an online fashion to minimize regret in online convex optimization. To address this, we first build a *novel Loewner order constraint for approximate preconditioners*, when followed, leads to a regret upper bound in online convex optimization. We show a *correlation between LogDet matrix divergence of approximate preconditioner with full-matrix Adagrad and our regret upperbound*. We then utilize both the Loewner order constraint and LogDet divergence to develop a Kronecker-structured preconditioner with factors jointly optimized to approximate full-matrix Adagrad. We show that a fully coupled Kronecker product is equivalent to Shampoo-KL, an existing algorithm, and propose a novel partially coupled version which uses the same number of matmuls as Shampoo, unlike Shampoo-KL. The Loewner order constraint and further analysis of the coupling between the axes preconditioners sets us up to show that fully coupled and partially coupled optimizers are optimal in online convex optimization. Our method improves generalization by up to 8% and accelerates training, improving time-to-target by up to 20%.

1 INTRODUCTION

Adaptive regularization (Hazan, 2016) is a well-known optimization technique with practical significance in deep learning optimization. ADAGRAD (Duchi et al., 2011) is an algorithm originally developed for convex objectives, that preconditions/multiplies the gradient with a diagonal matrix – preconditioner, which changes at each iteration, thus scaling each coordinate with its own adaptive learning rate. Several variants such as ADAM (Kingma & Ba, 2014) and NADAM (Dozat, 2016) have been developed to use the ADAGRAD preconditioner to scale the momentum and Nesterov based momentum of gradients respectively for deep neural network training. FULL-MATRIX ADAGRAD, also proposed in (Duchi et al., 2011), is a matrix generalization of ADAGRAD, where the gradient is preconditioned with a full-matrix preconditioner. The diagonal preconditioner can only address ill-conditioning in coordinate axes/basis, while a full-matrix preconditioner has the potential to address ill-conditioning in any arbitrary basis. However, full-matrix preconditioners can be expensive to store and compute as the the number of optimization variables/parameters increase, which is a common trend in deep neural networks.

Existing methods leverage tensor/matrix structure of gradients in deep neural networks to precondition each axis separately, thus reducing the computation and memory. These methods use Kronecker based approximations to FULL-MATRIX ADAGRAD: SHAMPOO (Gupta et al., 2018; Anil et al., 2020) which uses Kronecker product, CASPR (Duvvuri et al., 2024) which uses Kronecker sum. CASPR derives axes preconditioners independently from the gradients by approximating FULL-MATRIX ADAGRAD, and then combine them via Kronecker sum. In this paper, we discuss a coupled Kronecker product structure, where the Kronecker factors, one for each axis, are jointly tuned to approach FULL-MATRIX ADAGRAD. Two key challenges remain: a) Identifying what matrix Bregman divergence (Dhillon & Tropp, 2007) with full-matrix Adagrad improves convergence, and b) establishing coupled Kronecker approximations as adaptive regularization algorithm with optimal regret in online convex

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

optimization framework. The former helps us devise new preconditioners by reducing the matrix Bregman divergence with the full-matrix Adagrad, while solving the latter ensures that the proposed algorithm has optimal asymptotic dependence $\mathcal{O}(\sqrt{T})$ on number of optimization steps T, making it scalable to large number of optimization steps.

To solve the aforementioned challenges, we develop a novel theoretical framework for a sequence of approximate positive definite preconditioners which increase in Loewner order with optimization step and serve as an upperbound of full-matrix Adagrad preconditioner at each step. We show that such preconditioners admit a regret bound dependent on approximation error with full-matrix Adagrad measured by LogDet matrix divergence. Then, we devise LogDet divergence (Dhillon & Tropp, 2007) subproblems which jointly optimize for the factors in the Kronecker product to approach the full-matrix Adagrad preconditioner resulting in the *fully-coupled Kronecker product preconditioner*. But, this algorithm is expensive as it uses two additional matrix multiplications compared to Shampoo. So, we introduce a *partially coupled Kronecker product approximation* where only one factor is optimized while the other uses a factor from the Shampoo algorithm, while keeping the total number of matmuls per optimization step the same as Shampoo. To show that these coupled Kronecker approximations follow the aforementioned Loewner order conditions and admit optimal regret bound, we develop a novel proof strategy which uses our theoretical framework and a careful singular value analysis. Our main contributions are:

- We introduce a **Loewner order constraint** $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ for approximate preconditioners, yielding regret bounds that depend on $\operatorname{tr}(\tilde{H}_T^{1/2})$ and the LogDet divergence with FULL-MATRIX ADAGRAD (Theorem 3.2).
- We prove a **gap theorem** (Theorem A.5): minimizing LogDet divergence between $\tilde{H}_t$ and $\tilde{H}_{t-1} + g_t g_t^\top$ at each step controls global approximation error, enabling a greedy approach to approximate FULL-MATRIX ADAGRAD.
- We derive **coupled Kronecker preconditioners** by solving the Loewner constraint and minimizing LogDet divergence for $\tilde{H}_t = L_t \otimes R_t$. The partially coupled variant matches SHAMPOO's computational cost.
- We establish the **first** $\mathcal{O}(\sqrt{T})$ **regret bounds** for coupled Kronecker factorizations in online convex optimization (Theorem A.15).
- On **AlgoPerf benchmarks**(Dahl et al., 2023), the partially coupled preconditioner achieves 0.7–4.65% better generalization and 2.6–20% faster time-to-target than SHAMPOO, which leads the AlgoPerf benchmark.

1.1 PRELIMINARIES

Positive definite matrices and Kronecker products. We use $A \otimes B$ to denote Kronecker product of matrices $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mm}B \end{pmatrix}$$

where a_{ij} denotes the elements of matrix A.

A symmetric matrix $A \in \mathbb{R}^{m\times m}$ is positive definite if $x^\top Ax > 0$ for any non-zero $x \in \mathbb{R}^m$ (positive semi-definite if $x^\top Ax \geq 0$). The spectral radius is $\rho(A) = \max_i |\lambda_i(A)|$. We use Loewner orderings $\succ, \succeq, \preceq, \prec$: $A_1 \succeq A_2$ iff $A_1 - A_2$ is positive semi-definite. We use row-major vectorization $g = \overline{\operatorname{vec}}(G)$ for $G \in \mathbb{R}^{m\times n}$.

Standard Kronecker product identities (Lemma A.1), the rank-r outer-product bound $\frac{1}{r} gg^\top \preceq I_m \otimes G^\top G$ (Lemma A.4, Gupta et al., 2018), notation for block-diagonal matrices, and the general Bregman divergence framework are summarized in Appendix A.1.

LogDet matrix divergence. For positive definite $A, B \succ 0$, the LogDet matrix divergence (Kulis et al., 2009; Dhillon & Tropp, 2007) is the Bregman divergence (Bregman, 1967; Dhillon & Tropp,

2007) induced by $\phi(X) = -\log\det(X)$:

$$\mathrm{D}_{\ell\mathrm{d}}(A, B) = \mathrm{tr}(AB^{-1}) - \log\det(AB^{-1}) - d, \tag{1}$$

where d is the dimension of the matrices. When $A \succeq B$, all eigenvalues of AB^{-1} are at least 1, and the LogDet divergence captures how much A exceeds B.

Jeffreys LogDet divergence. Since $\mathrm{D}_{\ell\mathrm{d}}$ is asymmetric, we will also use its symmetrization, known as the Jeffreys (LogDet) divergence (Cichocki et al., 2015):

$$\delta_{\mathrm{LD}}{}^2(A, B) \;\coloneqq\; \mathrm{D}_{\ell\mathrm{d}}(A, B) + \mathrm{D}_{\ell\mathrm{d}}(B, A) \;=\; \mathrm{tr}(AB^{-1}) + \mathrm{tr}(BA^{-1}) - 2d. \tag{2}$$

This quantity appears directly in the regret bound (Theorem 3.2) as $S_t = \delta_{\mathrm{LD}}{}^2(\tilde{H}_t, H_t)$, measuring the symmetric approximation error between the current preconditioner and full-matrix Adagrad.

S-divergence (symmetric Stein divergence). The S-divergence (Sra, 2016) is a metric on positive definite matrices defined as:

$$\delta_{\mathrm{S}}(A, B) = \sqrt{\log\det\left(\frac{A+B}{2}\right) - \frac{1}{2}\log\det(AB)} \tag{3}$$

$$= \sqrt{\mathrm{D}_{\ell\mathrm{d}}(A, (A+B)/2) + \mathrm{D}_{\ell\mathrm{d}}(B, (A+B)/2)} \tag{4}$$

Unlike the LogDet divergence, δ_{S} satisfies the triangle inequality, making it a true metric on the cone of positive definite matrices.

$\delta_{\mathrm{S}}{}^2(A, B)$ and the Jeffreys LogDet divergence $\delta_{\mathrm{LD}}{}^2(A, B)$ are distinct quantities (Sra, 2016); the S-divergence admits the variational form (Sra, 2011) $\delta_{\mathrm{S}}{}^2(A, B) = \min_{Z \succ 0}\{\mathrm{D}_{\ell\mathrm{d}}(A, Z) + \mathrm{D}_{\ell\mathrm{d}}(B, Z)\}$. Inspired by this observation, we use the one-sided LogDet divergence $\mathrm{D}_{\ell\mathrm{d}}$ as the objective for our preconditioner development in the following sections.

2 Related Work

Adaptive regularization methods multiply the gradient with a preconditioner constructed from gradient history. Diagonal-Adagrad (Duchi et al., 2011) maintains accumulated squared gradients for coordinate-wise learning rates. Full-matrix Adagrad (Duchi et al., 2011) generalizes to full-matrix preconditioners, addressing ill-conditioning in arbitrary bases rather than just coordinate axes.

Kronecker product approximations such as Shampoo (Gupta et al., 2018; Anil et al., 2020; Shi et al., 2023) and CASPR (Duvvuri et al., 2024) approximate Full-matrix Adagrad by constructing separate preconditioners for each axis. Shampoo demonstrates superior performance over diagonal methods like AdamW, Adagrad, and RMSprop (Hinton et al., 2012), with industry-scale adoption at Google (Anil et al., 2022). CASPR shows theoretical improvements over Shampoo by improving its regret bound and approximation error. Morwani et al. (2024) propose coupled Kronecker factoring via Frobenius norm minimization, showing Shampoo emerges from such coupling.

Shampoo-KL (Lin et al., 2025) reformulates second-moment estimation through KL divergence minimization between Gaussian distributions, arriving at coupled update rules equivalent to our fully-coupled preconditioner (Section 3.5). While Shampoo-KL focuses on empirical performance, our work provides the first regret analysis for such preconditioners via a Loewner order constraint.

AlgoPerf (Dahl et al., 2023) provides a methodology for rigorous optimizer comparison across eight benchmarks. The Distributed Shampoo submission (Shi et al., 2023) won the competition (MLCommons, 2024) with 28% improvement over NAdamW. We demonstrate further improvements through coupled Kronecker factoring.

3 Adaptive Regularization via Coupled Kronecker Products

3.1 Full-matrix adaptive regularization

The framework of online convex optimization (OCO) (Hazan, 2016) involves making predictions $w_t \in \mathcal{W} \subseteq \mathbb{R}^d$ at each round/iteration t, where $\mathcal{W}$ is a bounded convex set with diameter $D =$

$\sup_{w,w' \in \mathcal{W}} \|w - w'\|$ (Duchi et al., 2011). The objective functions $f_t : \mathcal{W} \to \mathbb{R}$, $t \in \{1, \ldots, T\}$ are to be minimized. Due to the online nature of the optimization, the following formulation of regret is used to assess the predictions w_t with respect to the optimum $w^* = \arg\min_{w \in \mathcal{W}} \sum_{t=1}^{T} f_t(w)$:

$$\mathcal{R}(T) = \sum_{t=1}^{T} f_t(w_t) - f_t(w^*). \tag{5}$$

Regret minimization has been used to develop convergence bounds for stochastic non-convex optimization in Cutkosky et al. (2023); Ahn & Cutkosky (2024). Additionally, there are online-to-convex reductions from Cesa-Bianchi et al. (2004), which show that regret upper bound - $\mathcal{R}(T)$ translates to a convergence upper bound of $\mathcal{R}(T)/T$ in stochastic convex optimization. Hence, the condition for convergence is $\mathcal{R}(T) = o(T)$.

Since the empirical workloads in Section 4 are non-convex, we note that an $\mathcal{O}(\sqrt{T})$ OCO regret bound converts to an $\mathcal{O}(1/\sqrt{T})$ stationary-point convergence rate via the reductions of Cutkosky et al. (2023); Ahn & Cutkosky (2024); see Section A.5 for details.

Adaptive regularization introduced in (Duchi et al., 2011) includes FULL-MATRIX ADAGRAD, which makes updates to the weights at each iteration t by accumulating outer products of the observed gradients $g_t = \nabla f_t(w_t)$ into a second-moment matrix $H_t \in \mathbb{R}^{d \times d}$ using the following update rule:

$$\begin{aligned} H_t &\coloneqq H_{t-1} + g_t g_t^\top, \\ w_{t+1} &\coloneqq w_t - \eta H_t^{-1/2} g_t. \end{aligned} \tag{6}$$

The diagonal version of the above preconditioner - $\mathrm{diag}(H_t)^{-1/2}$ - ADAGRAD is commonly used to train deep learning models. Some of the well known algorithms include ADAM (Kingma & Ba, 2014), RMSProp (Hinton et al., 2012), which use variants of ADAGRAD to precondition the gradient. Unlike diagonal preconditioners, full-matrix preconditioners have the potential to navigate anisotropic curvature of the objective function faster. The winning optimizer of the AlgoPerf competition (Dahl et al., 2023) - Distributed Shampoo (Shi et al., 2023; Gupta et al., 2018; Anil et al., 2020) uses an approximation of FULL-MATRIX ADAGRAD, and has demonstrated accelerated convergence compared to diagonal methods.

A standard regret decomposition (Lemma A.6, Appendix A.4) splits $\mathcal{R}(T)$ into a preconditioner-monotonicity term $\mathcal{R}_1(T) = \frac{1}{2\eta} \sum_t (w_t - w^*)^\top (H_t^{1/2} - H_{t-1}^{1/2})(w_t - w^*)$ and a gradient-norm term $\mathcal{R}_2(T) = \frac{\eta}{2} \sum_t g_t^\top H_t^{-1/2} g_t$. The first term encourages small changes between consecutive preconditioners H_t and H_{t-1}, in contrast, the second term encourages larger H_t. In the following lemma, FULL-MATRIX ADAGRAD is shown to balance and bound both terms with similar values.

Lemma 3.1 (FULL-MATRIX ADAGRAD regret bound (Hazan, 2016))**.** *FULL-MATRIX ADAGRAD predictions (6) over $\mathcal{W}$ with step size $\eta = D/\sqrt{2}$, where D is the diameter of $\mathcal{W}$, satisfies:*

$$\begin{aligned} \mathcal{R}_1(T), \mathcal{R}_2(T) &\leq \frac{D}{\sqrt{2}} \operatorname{tr}(H_T^{1/2}), \\ \mathcal{R}(T) &\leq \sqrt{2} D \operatorname{tr}(H_T^{1/2}) = \mathcal{O}(\sqrt{T}) \end{aligned}$$

The proof for this lemma can be found in (Hazan, 2016; Duchi et al., 2011). In the next section, we use Lemma A.6 and 3.1 to develop a regret upper bound for a general sequence of approximations of second-moment matrices.

3.2 APPROXIMATE FULL-MATRIX ADAPTIVE REGULARIZATION

At each iteration, FULL-MATRIX ADAGRAD (Duchi et al., 2011) has access to the current gradient g_t obtained from back-propagation and previous second-moment matrix H_{t-1} which can be infeasible to hold in memory for large dimension d, due to a memory complexity of $\mathcal{O}(d^2)$. Furthermore, finding the preconditioner $H_t^{-1/2}$ requires inverse square root which admits a $\mathcal{O}(d^3)$ complexity. This behooves us to develop practical approximations $\tilde{H}_t$ which a) admit a regret upper bound b) which are feasible to store in memory, c) which have fast inverse square root - $\tilde{H}_t^{-1/2}$, d) fast to

precondition the gradient via the transformation $g_t \to \tilde{H}_t^{-1/2} g_t$. We first address a) by developing the following theorem:

Theorem 3.2 (Regret bound of general approximations)**.** *For sequence of approximations $\tilde{H}_t$ which satisfy the Loewner order conditions: (i) $\tilde{H}_t \succeq \tilde{H}_{t-1}$ (monotonically non-decreasing), and (ii) $\tilde{H}_t \succeq H_t$ (dominating the true second-moment matrix), with $\operatorname{tr}(\tilde{H}_T^{1/2}) = o(T)$, the updates $w_{t+1} := w_t - \eta \tilde{H}_t^{-1/2} g_t$ with $\eta = \frac{D}{\sqrt{2}}$ yield the regret bound:*

$$\tilde{\mathcal{R}}(T) \leq \frac{D}{\sqrt{2}} \left(\operatorname{tr}(\tilde{H}_T^{1/2}) + \sqrt{\tilde{\kappa}} \operatorname{tr}(H_T^{1/2}) \right),$$

where D is the diameter of the constraint set $\mathcal{W}$, $\tilde{\kappa} = \max_t \rho(\tilde{H}_t^{-1} H_t) \leq 1$ is the inverse inflation factor*, and $\rho(\cdot)$ denotes spectral radius. By Lemma A.8:*

$$\tilde{\kappa} \leq \frac{(\bar{S}_T + 2) + \sqrt{\bar{S}_T(\bar{S}_T + 4)}}{2}$$

where $\bar{S}_T = \max_{t \leq T} S_t$ and $S_t = \mathrm{D}_{\ell\mathrm{d}}(\tilde{H}_t, H_t) + \mathrm{D}_{\ell\mathrm{d}}(H_t, \tilde{H}_t) = {\delta_{\mathrm{LD}}}^2(\tilde{H}_t, H_t)$ is the Jeffreys LogDet divergence (Cichocki et al., 2015) at step t.

Similar to Lemma A.6, the regret $\tilde{\mathcal{R}}(T)$ decomposes into two terms $\tilde{\mathcal{R}}_1(T)$ and $\tilde{\mathcal{R}}_2(T)$ with $\tilde{H}_t$ replacing H_t. The first term $\tilde{\mathcal{R}}_1(T)$ is bounded using monotonicity of $\tilde{H}_t$, while $\tilde{\mathcal{R}}_2(T)$ is bounded via a Hölder-type argument using the inverse inflation factor $\tilde{\kappa}$.

Proof sketch for $\tilde{\mathcal{R}}_1(T)$: Using monotonicity $\tilde{H}_t \succeq \tilde{H}_{t-1}$ and properties of matrix square roots: $\sum_t \operatorname{tr}(\tilde{H}_t^{1/2} - \tilde{H}_{t-1}^{1/2}) = \operatorname{tr}(\tilde{H}_T^{1/2})$, giving $\tilde{\mathcal{R}}_1(T) \leq \frac{D}{\sqrt{2}} \operatorname{tr}(\tilde{H}_T^{1/2})$.

Proof sketch for $\tilde{\mathcal{R}}_2(T)$: For each t, define $\tilde{\kappa}_t = \rho(\tilde{H}_t^{-1} H_t) \leq 1$. By a Hölder-type argument, $g_t^\top \tilde{H}_t^{-1/2} g_t \leq \sqrt{\tilde{\kappa}_t} \cdot g_t^\top H_t^{-1/2} g_t$. Taking $\tilde{\kappa} = \max_t \tilde{\kappa}_t$ and applying Lemma 3.1: $\sum_t g_t^\top \tilde{H}_t^{-1/2} g_t \leq \sqrt{\tilde{\kappa}} \sum_t g_t^\top H_t^{-1/2} g_t \leq 2\sqrt{\tilde{\kappa}} \operatorname{tr}(H_T^{1/2})$.

Connection to LogDet divergence: By Lemma A.8, for each t, $\tilde{\kappa}_t \leq \frac{(S_t+2)+\sqrt{S_t(S_t+4)}}{2}$. Taking the maximum over t gives the bound with $\bar{S}_T = \max_t S_t$. The full proof is in Section A.4.

A sufficient condition is $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ for all t, which implies both Loewner conditions of Theorem 3.2 (Lemma A.10, Appendix A.4). In Section 3.3, we use Lemma A.10 to verify that our coupled Kronecker product preconditioners satisfy the Loewner order conditions, thereby inheriting the regret bounds of Theorem 3.2.

The bound in Theorem 3.2 reveals that closeness between $\tilde{H}_t$ and H_t reduces the regret bound through the inverse inflation factor $\tilde{\kappa} = \max_t \rho(\tilde{H}_t^{-1} H_t) \leq 1$, which measures how close the approximation is to the true second-moment matrix (eigenvalue ratio $H_t / \tilde{H}_t$). When $\tilde{H}_t = H_t$ for all t (exact FULL-MATRIX ADAGRAD), we have $\tilde{\kappa} = 1$ and the bound recovers $\sqrt{2} D \operatorname{tr}(H_T^{1/2})$, matching Lemma 3.1. When $\bar{S}_T = \max_t S_t$ is small (good approximation uniformly over time), $\tilde{\kappa}$ is close to 1.

We state Theorem 3.2 to highlight *why* minimizing the LogDet divergence with FULL-MATRIX ADAGRAD is the right objective for approximate preconditioners — the inflation factor $\tilde{\kappa}$ links approximation quality to regret. To prove optimality of our partially coupled and fully coupled algorithms, however, we do not invoke the inflation-factor refinement: the trace term $\operatorname{tr}(\tilde{H}_T^{1/2})$ together with the Loewner constraint of Lemma A.10 (verified for our updates in Section 3.3) is sufficient to obtain the optimal $\mathcal{O}(\sqrt{T})$ rate, as carried out in Theorem A.15 and Theorem 3.4.

3.3 COUPLED KRONECKER PRODUCT

In deep neural networks, matrix shaped gradients $G_t \in \mathbb{R}^{m \times n} = \nabla_{W_t} f_t(W_t)$ occur frequently. An emerging preconditioning approach in Gupta et al. (2018) and Duvvuri et al. (2024) is to precondition the gradient on the left $\hat{L}_t G_t$ and right $G_t \hat{R}_t$ via matrix multiplication with the *axes preconditioners* $\hat{L}_t \in \mathbb{R}^{m \times m}$, $\hat{R}_t \in \mathbb{R}^{n \times n}$, as a result, $\hat{L}_t$ and $\hat{R}_t$ precondition the columns and

rows of G_t respectively. Such preconditioning is motivated by the fast matrix multiplication kernels in accelerators, lower memory $\mathcal{O}(m^2 + n^2)$ and computation $\mathcal{O}(m^3 + n^3)$ than FULL-MATRIX ADAGRAD eq. (6). Here we develop more accurate axes preconditioners which are closer to FULL-MATRIX ADAGRAD by introducing *coupled Kronecker factoring*.

3.3.1 LOGDET DIVERGENCE MINIMIZATION.

To construct coupled Kronecker approximations $\tilde{H}_t = L_t \otimes R_t$ that satisfy the sufficient condition in Lemma A.10, we seek to minimize the gap between $\tilde{H}_t$ and $\tilde{H}_{t-1} + g_t g_t^\top$. Intuitively, keeping $\tilde{H}_t$ close to $\tilde{H}_{t-1} + g_t g_t^\top$ at each step prevents the approximation error from accumulating, thereby ensuring that $\tilde{H}_T$ remains close to $H_T = \sum_{t=1}^T g_t g_t^\top$. The S-divergence δ_{S} satisfies a contraction property under additive shifts (Sra, 2016), which provides theoretical justification for this approach. This motivates the following subproblem at each iteration:

$$\min_{L,R} \mathrm{D}_{\ell\mathrm{d}}(\tilde{H}_{t-1} + g_t g_t^\top, L \otimes R). \tag{7}$$

While the symmetrized divergence $\mathrm{D}_{\ell\mathrm{d}}(A, B) + \mathrm{D}_{\ell\mathrm{d}}(B, A)$ captures total discrepancy, we optimize the one-directional LogDet divergence $\mathrm{D}_{\ell\mathrm{d}}(\tilde{H}_{t-1} + g_t g_t^\top, L \otimes R)$ because this problem is convex in L^{-1} and R^{-1} individually (Dhillon & Tropp, 2007), and admits explicit closed-form solutions via alternating minimization.

Via the metric properties of δ_{S} (triangle inequality and additive contraction), minimizing the local gap $\delta_{\mathrm{S}}(\tilde{H}_t, \tilde{H}_{t-1} + g_t g_t^\top)$ at each step controls the global S-divergence between $\tilde{H}_T$ and H_T (Theorem A.5, Appendix A.3).

Theorem A.5 motivates the greedy LogDet objective (7) via the metric properties of δ_{S} (triangle inequality and additive contraction), which the Jeffreys LogDet divergence does not satisfy. The formal regret bound (Theorem A.15) is established directly by the Loewner constraint of Lemma A.10. The variational form of δ_{S} furthermore implies the upper bound ${\delta_{\mathrm{S}}}^2(A, B) \leq \mathrm{D}_{\ell\mathrm{d}}(A, B) + \mathrm{D}_{\ell\mathrm{d}}(B, A) = {\delta_{\mathrm{LD}}}^2(A, B)$, so reducing the one-sided $\mathrm{D}_{\ell\mathrm{d}}(\tilde{H}_{t-1} + g_t g_t^\top, L \otimes R)$ also reduces an upper bound on ${\delta_{\mathrm{S}}}^2$.

Theorem A.5 reduces the problem of approximating H_T to many smaller greedy subproblems—minimizing the gap between $\tilde{H}_t$ and $\tilde{H}_{t-1} + g_t g_t^\top$ at each step. This decomposition is crucial: rather than solving a global optimization problem over all T steps, we solve T local subproblems. Each greedy subproblem admits a tractable solution when $\tilde{H}_t$ has Kronecker structure, enabling the derivation of optimized coupled Kronecker factors.

Lemma 3.3 (Alternating minimization for coupled Kronecker factors)**.** *Let $\tilde{H}_{t-1} = L_{t-1} \otimes R_{t-1}$. Solving (7) via alternating minimization yields:*

$$\begin{aligned} L_t &= \arg\min_L \mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top, L \otimes R_{t-1}) \\ &= L_{t-1} + G_t R_{t-1}^{-1} G_t^\top / n, \\ R_t &= \arg\min_R \mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top, L_t \otimes R) \\ &= \frac{\mathrm{tr}(L_{t-1} L_t^{-1})}{m} R_{t-1} + \frac{1}{m} G_t^\top L_t^{-1} G_t, \end{aligned}$$

where $g_t = \overline{\mathrm{vec}}(G_t)$.

The proof uses transformation invariance of LogDet divergence and is given in Section A.6. When $R_{t-1} = I_n$, the update simplifies to $L_t = L_{t-1} + G_t G_t^\top / n$, recovering the Shampoo statistic.

Based on Lemma 3.3, we propose two variants of LogDet-based coupled Kronecker preconditioners, detailed below.

LogDet-based partially-coupled Kronecker product. In this variant, we fix $R_{t-1} = I_n$ in the L-subproblem while allowing R_t to evolve according to the second update rule:

$$\begin{aligned} L_t &\coloneqq L_{t-1} + G_t G_t^\top / n, \\ R_t &\coloneqq \frac{\mathrm{tr}(L_{t-1} L_t^{-1})}{m} R_{t-1} + \frac{1}{m} G_t^\top L_t^{-1} G_t. \end{aligned} \tag{8}$$

This uses the standard Shampoo update for L_t while introducing coupling through R_t.

LogDet-based fully-coupled Kronecker product. In this variant, both factors evolve according to the alternating minimization updates:

$$L_t \coloneqq L_{t-1} + G_t R_{t-1}^{-1} G_t^\top / n,$$
$$R_t \coloneqq \frac{\operatorname{tr}(L_{t-1} L_t^{-1})}{m} R_{t-1} + \frac{1}{m} G_t^\top L_t^{-1} G_t. \tag{9}$$

Here the L_t update incorporates information from R_{t-1}, creating bidirectional coupling between the two factors. This yields a tighter approximation to the LogDet divergence minimization subproblem (7).

3.3.2 Coupled Kronecker Product from Loewner Order Constraint

In the previous section, both fully-coupled and partially-coupled variants produced are from LogDet divergence minimization. However, these updates do not directly satisfy the Loewner constraint $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ required by Lemma A.10. To obtain regret bounds, we construct sequences L_t, R_t that satisfy the Loewner constraint.

We now recall Theorem 3.2, where the sufficient condition in Lemma A.10 requires $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ to obtain regret bounds. Substituting $\tilde{H}_t = L_t \otimes R_t$, we seek sequences of $L_t, R_t \succeq 0$ satisfying:

$$L_t \otimes R_t \succeq L_{t-1} \otimes R_{t-1} + g_t g_t^\top \tag{10}$$

Loewner order for partially coupled preconditioner. For the partially coupled case (8), we establish (10) by transforming via $(L_t^{1/2} \otimes I_n)$:

$$L_{t-1} \otimes R_{t-1} + g_t g_t^\top = (L_t^{1/2} \otimes I_n)\big(L_t^{-1/2} L_{t-1} L_t^{-1/2} \otimes R_{t-1} + (L_t^{-1/2} \otimes I_n) g_t g_t^\top (L_t^{-1/2} \otimes I_n)\big)(L_t^{1/2} \otimes I_n)$$

Setting $R_t \coloneqq \|L_t^{-1/2} L_{t-1} L_t^{-1/2}\| \, R_{t-1} + \tilde{r} \, G_t^\top L_t^{-1} G_t$ (with $\tilde{r} = \max_t \operatorname{rank}(G_t)$) satisfies the Loewner constraint (10) for $\tilde{H}_t = L_t \otimes R_t$ (Lemma A.12, Appendix A.6).

Loewner order for fully coupled preconditioner. For the fully coupled case (9), we split $g_t g_t^\top$ into two halves and handle each with a different transformation. This technique accounts for the bidirectional coupling. For the fully coupled updates $L_t \coloneqq L_{t-1} + \tilde{r} \, G_t R_{t-1}^{-1} G_t^\top$ and $R_t \coloneqq R_{t-1} + \tilde{r} \, G_t^\top L_{t-1}^{-1} G_t$, the Loewner constraint (10) holds (Lemma A.13, Appendix A.6).

Given that the update rules satisfy (10), Theorem 3.2 can be applied. Both updates yield regret $\sum_t f_t(W_t) - f(W^*) \leq \sqrt{2} D \, \operatorname{tr}(L_T^{1/2}) \operatorname{tr}(R_T^{1/2})$ (Theorems A.14, A.15, Appendix A.7).

3.4 Scaling analysis and optimal $\mathcal{O}(\sqrt{T})$ regret

The regret bound in Theorem A.15 depends on $\operatorname{tr}(L_T^{1/2}) \operatorname{tr}(R_T^{1/2})$. We analyze the scaling of these quantities under the bounded-gradient assumption $G_t G_t^\top \preceq g^2 I$.

(Scalar intuition for the recurrences is given in Appendix A.8.)

Distributed Shampoo (Anil et al., 2021; Shi et al., 2023) adds a small diagonal δI to each factor before inversion for limited-precision (fp32) numerical stability. We analyze this same regularization with a step-dependent schedule $\delta_t I$, which controls the condition number and yields the optimal $\mathcal{O}(\sqrt{T})$ regret rate.

Theorem 3.4 (Regularized scaling and regret of fully and partially coupled preconditioners)**.** *Under the bounded-gradient assumption $G_t G_t^\top \preceq g^2 I_m$, both regularized variants below achieve*

$$\operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) = \mathcal{O}\big(mn \, g \sqrt{\tilde{r} T}\big) \quad \textit{and hence} \quad \mathcal{R}(T) = \mathcal{O}\big(D \cdot mn \cdot g \sqrt{\tilde{r} T}\big),$$

matching the optimal $\mathcal{O}(\sqrt{T})$ rate.

(a) Fully coupled regularized. *For $c > 0$ with the updates*

$$\bar{L}_t \coloneqq \bar{L}_{t-1} + \tilde{r}\, G_t \bar{R}_{t-1}^{-1} G_t^\top + \tfrac{c}{\sqrt{t}}\, I_m, \qquad \bar{R}_t \coloneqq \bar{R}_{t-1} + \tilde{r}\, G_t^\top \bar{L}_{t-1}^{-1} G_t + \tfrac{c}{\sqrt{t}}\, I_n,$$

the choice $c = g\sqrt{\tilde{r}}$ achieves the bound; both factors satisfy $\lambda_{\max}(\bar{L}_T), \lambda_{\max}(\bar{R}_T) = \mathcal{O}(\sqrt{T})$.

(b) Partially coupled regularized. *For $\rho > 0$ with the updates*

$$\bar{L}_t \coloneqq \bar{L}_{t-1} + G_t G_t^\top + \rho\, I_m, \quad \bar{R}_t \coloneqq \bar{c}_t \bar{R}_{t-1} + \tilde{r}\, G_t^\top \bar{L}_t^{-1} G_t, \quad \bar{c}_t = \|\bar{L}_t^{-1/2} \bar{L}_{t-1} \bar{L}_t^{-1/2}\|,$$

the choice $\rho = g^2$ achieves the bound; here $\lambda_{\max}(\bar{L}_T) = \mathcal{O}(T)$ but $\lambda_{\max}(\bar{R}_T) = \mathcal{O}(1)$ via shrinkage induced by the ρI regularization.

Proof sketch (full proofs in Section A.8). For both variants we run a 3-step argument. *Step 1 (eigenvalue floor):* the regularization contributes a PSD increment to $\bar{L}_t$ at each step, giving $\lambda_{\min}(\bar{L}_t) \geq c\sqrt{t}$ in (a) and $\geq \rho t$ in (b) without any gradient-side assumption. *Step 2 (max eigenvalue):* combining Step 1 with $G_t G_t^\top \preceq g^2 I$ yields $\bar{\Lambda}_T = \mathcal{O}(\sqrt{T})$ in (a) and $\bar{\Lambda}_T = \mathcal{O}(T)$ in (b); for $\bar{R}_T$, the symmetric argument in (a) gives $\mathcal{O}(\sqrt{T})$, while in (b) the regularization-induced shrinkage $\bar{c}_t \leq 1 - \mathcal{O}(1/t)$ keeps $\lambda_{\max}(\bar{R}_T) = \mathcal{O}(1)$. *Step 3 (trace product):* substituting the eigenvalue bounds yields the trace product, and applying Theorem A.15 gives the regret bound. □

3.5 Equivalence with Shampoo-KL

Concurrent work by Lin et al. (2025) derives coupled updates via KL divergence minimization between Gaussian distributions, yielding $S_a \leftarrow (1-\beta_2)S_a + \frac{\beta_2}{n}\mathbb{E}[G S_b^{-1} G^\top]$ and $S_b \leftarrow (1-\beta_2)S_b + \frac{\beta_2}{m}\mathbb{E}[G^\top S_a^{-1} G]$. These are equivalent to our fully-coupled updates (Lemma A.13), differing only in accumulation scheme (EMA vs. direct summation). Our work provides the first $\mathcal{O}(\sqrt{T})$ regret bounds for such preconditioners.

3.6 Computational cost

A natural concern is whether coupled updates introduce additional overhead. We show that with $c_t = 1$, the partially coupled preconditioner requires the *same 4 matrix multiplications per iteration* as Shampoo: $G_t G_t^\top$ for updating L_t (1 matmul), $L_t^{-1/2} G_t$ for left preconditioning (1 matmul), $(L_t^{-1/2} G_t)^\top (L_t^{-1/2} G_t)$ for updating R_t (1 matmul) and $(L_t^{-1/2} G_t) R_t^{-1/2}$ for right preconditioning (1 matmul). The key observation is that $G_t^\top L_t^{-1} G_t = (L_t^{-1/2} G_t)^\top (L_t^{-1/2} G_t)$ reuses the left-preconditioned gradient already computed for the output $L_t^{-1/2} G_t R_t^{-1/2}$. For large parameter matrices, blocking into smaller sub-matrices with per-block Kronecker factors (Anil et al., 2021; Shi et al., 2023) keeps factor sizes bounded regardless of model scale.

4 Experiments

We evaluate the partially coupled preconditioner on 7 AlgoPerf benchmarks (Dahl et al., 2023): ImageNet ViT, ImageNet ResNet-50, OGBG GNN, WMT Transformer, Criteo 1TB DLRMSmall, LibriSpeech Conformer, and LibriSpeech DeepSpeech. We compare against NAdamW, Shampoo, and CASPR using grafting (Agarwal et al., 2020) for fair comparison. Implementation details, including practical update rules with exponential moving averages and hyperparameter search spaces, are in Section A.10. We add δI to L_t, R_t before each inverse so that $\kappa(L_t), \kappa(R_t) \leq 10^7$, the maximum tolerated by fp32 inverse operations (Anil et al., 2021).

4.1 Generalization performance

Table 1 shows validation and test metrics after training for a fixed number of steps with 50 random hyperparameter configurations per optimizer. The partially coupled preconditioner outperforms baselines on 5 of 7 benchmarks, with notable improvements on LibriSpeech Conformer (val 3.15%, test 8.08%), LibriSpeech DeepSpeech (val 4.65%, test 2.44%), OGBG GNN (val 2.16%, test 2.76%), and ImageNet ViT (val 0.7%, test 1.8%).

Table 1: Performance comparison of NADAMW, SHAMPOO, CASPR, and the partially coupled preconditioner (Ours) on various benchmarks. The left table shows the results for validation metrics, while the right table presents test metrics. The best results in each benchmark are highlighted in bold. For average precision and BLEU score, higher is better, for other metrics lower is better.

Benchmark	NADAMW	SHAMPOO	CASPR	Ours	Validation Metric
IMAGENET VIT	0.25048	0.23016	0.23282	**0.22844**	Error Rate
OGBG GNN	0.29296	0.29655	0.29544	**0.30297**	Average Precision
WMT TRANSFORMER	31.133	**31.319**	31.297	31.208	BLEU Score
CRITEO 1TB DLRMSMALL	0.12382	0.12376	**0.12362**	**0.12361**	Loss
LIBRISPEECH CONFORMER	0.08294	0.08102	0.08210	**0.07846**	Word Error Rate
LIBRISPEECH DEEPSPEECH	0.11629	0.11201	0.11271	**0.10680**	Word Error Rate
IMAGENET RESNET-50	0.25048	0.22000	**0.21900**	0.22196	Error Rate

Benchmark	NADAMW	SHAMPOO	CASPR	Ours	Test Metric
IMAGENET VIT	0.37120	0.35080	0.35200	**0.34430**	Error Rate
OGBG GNN	0.27716	0.28144	0.27899	**0.28943**	Average Precision
WMT TRANSFORMER	**31.3934**	31.3348	31.0362	31.1958	BLEU Score
CRITEO 1TB DLRMSMALL	0.12621	0.12618	**0.12600**	**0.12601**	Loss
LIBRISPEECH CONFORMER	0.05017	0.04948	0.05019	**0.04548**	Word Error Rate
LIBRISPEECH DEEPSPEECH	0.06946	0.06812	0.06637	**0.06475**	Word Error Rate
IMAGENET RESNET-50	0.37120	**0.33290**	0.33840	0.34420	Error Rate

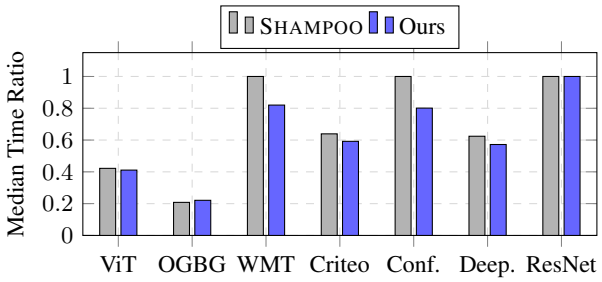

Figure 1: Median time ratio (time to reach target / time budget) for SHAMPOO and the partially coupled preconditioner across AlgoPerf benchmarks. Lower is better; ratio of 1.0 indicates failure to reach the target within the time budget. The partially coupled preconditioner outperforms SHAMPOO on 5 of 7 benchmarks, with notable improvements on WMT (18%), LibriConf (20%), and LibriDeep (8%).

4.2 TIME TO TARGET

We compare against the AlgoPerf competition winner—Distributed Shampoo (MLCommons, 2024; Shi et al., 2023)—measuring median time ratio (time to target / time budget) across 3 seeds. Figure 1 shows the partially coupled preconditioner outperforms SHAMPOO on 5 of 7 benchmarks: WMT TRANSFORMER (18% faster), LIBRISPEECH CONFORMER (20%), and LIBRISPEECH DEEPSPEECH (8%).

5 CONCLUSION

We introduce a Loewner order framework for approximate preconditioners that admits regret bounds dependent on LogDet divergence with FULL-MATRIX ADAGRAD. This motivates coupled Kronecker approximations: the fully coupled preconditioner (equivalent to Shampoo-KL) and the partially coupled preconditioner, which maintains the same computational cost as Shampoo. We establish the first optimal $\mathcal{O}(\sqrt{T})$ regret bounds for both, and demonstrate improved generalization (0.7–8%) and time-to-target (2.6–20%) on AlgoPerf benchmarks.

ETHICS STATEMENT

This paper presents foundational work in optimization whose goal is to advance the field of machine learning. Our experiments use only publicly available AlgoPerf benchmark datasets and models (Dahl et al., 2023); no human subjects, private data, or personally identifiable information are involved. We do not foresee societal consequences of this work that must be specifically highlighted beyond those common to general-purpose improvements in deep learning optimization.

REPRODUCIBILITY STATEMENT

All theoretical results are stated with their assumptions in the main text, and complete proofs are provided in the appendix: Kronecker-product preliminaries and supporting lemmas (Appendices A.1 and A.2), the gap theorem (Appendix A.3), the general regret framework (Appendix A.4), the Loewner

order proofs and regret bounds for the coupled preconditioners (Appendices A.6 and A.7), and the scaling analysis (Appendix A.8). For the experiments, Appendix A.10 lists the benchmarks, the practical update rules (including exponential moving averages, grafting, and blocking), the hardware used, the full hyperparameter search space (Table 2), the AlgoPerf targets and budgets, and the fixed optimizer settings (Table 4). All benchmarks are from the public AlgoPerf suite (Dahl et al., 2023).

AI USE STATEMENT

In this work, we used generative AI tools to reformat the manuscript to the ICLR template and to edit the paper for readability, and to format references. We have not used generative AI tools to generate synthetic datasets, to develop the theoretical framework, to formulate mathematical claims, to provide critical ingredients for or to assist in the writing of proofs, to propose or refine hypotheses, to design or provide feedback on the research methodology or experiments, to implement the methods, or to interpret results; assistance with translation, dataset cleaning or reformatting, and qualitative or thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work: all AI-assisted edits were checked by the authors against the original manuscript, and no technical content, claims, or results were altered. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling adaptive gradient methods from learning rates. *arXiv preprint arXiv:2002.11803*, 2020.

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. *arXiv preprint arXiv:2405.18199*, 2024.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Towards practical second-order optimization for deep learning. 2021.

Rohan Anil, Sandra Gadanho, Da Huang, Nijith Jacob, Zhuoshu Li, Dong Lin, Todd Phillips, Cristina Pop, Kevin Regan, Gil I Shamir, et al. On the factory floor: Ml engineering for industrial-scale ads recommendation models. *arXiv preprint arXiv:2209.05310*, 2022.

Rajendra Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. Findings of the 2017 conference on machine translation (wmt 2017). Association for Computational Linguistics, 2017.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences. *Entropy*, 17(5):2988–3034, 2015.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *arXiv preprint arXiv:2302.03775*, 2023.

George E Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, et al. Benchmarking neural network training algorithms. *arXiv preprint arXiv:2306.07179*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Inderjit S Dhillon and Joel A Tropp. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(61):2121–2159, 2011.

Sai Surya Duvvuri, Fnu Devvrit, Rohan Anil, Cho-Jui Hsieh, and Inderjit S Dhillon. Combining axes preconditioners through kronecker approximation for deep learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

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

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

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

Brian Kulis, Mátyás A Sustik, and Inderjit S Dhillon. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research*, 10(13):341–376, 2009.

Criteo AI Lab. Criteo 1tb click logs dataset. `https://ailab.criteo.com/download-criteo-1tb-click-logs-dataset/`, 2014.

Wu Lin, Scott C. Lowe, Felix Dangel, Runa Eschenhagen, Zikun Xu, and Roger B. Grosse. Understanding and improving shampoo and soap via Kullback-Leibler minimization. *arXiv preprint arXiv:2509.03378*, 2025.

MLCommons. Announcing the results of the inaugural algoperf: Training algorithms benchmark competition. https://mlcommons.org/2024/08/mlc-algoperf-benchmark-competition/, 2024. Accessed: 2024-09-10.

Depen Morwani, Itai Shapira, Nikhil Vyas, Eran Malach, Sham Kakade, and Lucas Janson. A new perspective on shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*, 2024.

Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*, 2019.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

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. *arXiv preprint arXiv:2309.06497*, 2023.

Suvrit Sra. Positive definite matrices and the symmetric stein divergence. *arXiv preprint arXiv:1110.1773*, 2011.

Suvrit Sra. Positive definite matrices and the S-divergence. *Proceedings of the American Mathematical Society*, 144(7):2787–2797, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A Appendix

Table of Contents

A.1 Preliminaries and Kronecker Properties p.12
A.2 Supporting Lemmas p.14
A.3 Gap Theorem Proof p.15
A.4 General Regret Framework p.17
A.5 From OCO Regret to Non-Convex Stationarity p.20
A.6 Loewner Order Proofs p.20
A.7 Regret Bounds for Coupled Preconditioners p.22
A.8 Scaling Analysis and Optimal Regret Proofs p.24
A.9 Frobenius Norm Approximation p.27
A.10 Experimental Details and Hyperparameters p.27

A.1 Preliminaries and Kronecker Properties

Notation. Let $\mathrm{vec}(G)$ denote column-major vectorization of matrix G (stacking columns) and $\overline{\mathrm{vec}}(G)$ denote row-major vectorization (stacking rows). Throughout this paper we use the row-major form: $g = \overline{\mathrm{vec}}(G)$ for gradient matrices $G \in \mathbb{R}^{m \times n}$, which satisfies $(A \otimes B)\,\overline{\mathrm{vec}}(G) = \overline{\mathrm{vec}}(AGB)$ for symmetric A, B.

Block diagonal matrices. $\text{block-diag}(A_1, \ldots, A_n)$ denotes a block-diagonal matrix with blocks $\{A_1, \ldots, A_n\}$. When all blocks are identical, $\text{block-diag}(A, \ldots, A) = I_n \otimes A$ where $I_n \in \mathbb{R}^{n \times n}$ is the identity. Note that $A \otimes I_n$ is also block-diagonal up to a symmetric permutation: by Lemma A.1(v), $P^\top (A \otimes I_n) P = I_n \otimes A$ for the appropriate permutation P.

Bregman matrix divergence. The Bregman matrix divergence (Bregman, 1967; Dhillon & Tropp, 2007) between $A, B \succ 0$ induced by a strictly convex function ϕ is $\mathrm{D}_\phi(A, B) = \phi(A) - \phi(B) - \operatorname{tr}(\nabla\phi(B)^\top (A - B))$. For strictly convex ϕ, $\mathrm{D}_\phi(A, B) \geq 0$ with equality iff $A = B$. For example, with $\phi(X) = \|X\|_F^2$ the divergence is the squared Frobenius distance $\|A - B\|_F^2$. The LogDet matrix divergence used in this paper (Eq. (1)) corresponds to $\phi(X) = -\log\det(X)$.

Lemma A.1. *Let $A,\ B,\ C,\ D,\ G,\ A_1,\ A_2,\ B_1,\ B_2$ be matrices with appropriate dimensions then*

(v) If P is a permutation matrix such that $P \operatorname{vec}(G^\top) = \operatorname{vec}(G)$, then $(A \otimes B) = P(B^\top \otimes A)P^\top$.

(vii) For a positive definite matrix A, $x^\top A x \leq \|x\|_2^2 \operatorname{tr}(A)$.

Proof. Proof of (v). We can expand the $(A \otimes B) \operatorname{vec}(G)$ for $G \in \mathbb{R}^{m \times n}$ as follows:

$$\begin{aligned}
(A \otimes B) \operatorname{vec}(G) &= \operatorname{vec}(BGA^\top) (\text{Using Lemma A.1(ii)}) \\
&= \operatorname{vec}((AG^\top B)^\top) \\
&= P \operatorname{vec}(AG^\top B) \\
&= P(B^\top \otimes A) \operatorname{vec}(G^\top) \\
&= P(B^\top \otimes A) P^\top \operatorname{vec}(G)
\end{aligned}$$

Thus $A \otimes B = P(B^\top \otimes A)P^\top$

Proof of (vii). Let $A = V \Sigma V^\top$ be the eigenvalue decomposition with eigenpairs (σ_i, v_i), then

$$\begin{aligned}
x^\top A x &= x^\top V \Sigma V^\top x \\
&= \sum_{i=1}^n \sigma_i \langle v_i, x \rangle^2 \\
&\leq (\sum_{i=1}^n \sigma_i)(\max_i (\langle v_i, x \rangle^2)) (\text{Since } \sigma_i \text{ are positive}) \\
&= (\sum_{i=1}^n \sigma_i) \|V^\top x\|_\infty^2 \\
&\leq \operatorname{tr}(A) \|V^\top x\|_2^2 \\
&= \operatorname{tr}(A) \|x\|_2^2
\end{aligned}$$

□

Lemma A.2 (Concavity of matrix square root)**.** *For $A \succ 0$ and $B \succeq 0$:*

$$\operatorname{tr}((A + B)^{1/2}) \leq \operatorname{tr}(A^{1/2}) + \frac{1}{2} \operatorname{tr}(A^{-1/2} B).$$

Proof. The function $f(X) = X^{1/2}$ is operator concave on positive definite matrices (Bhatia, 2009). By operator concavity, for $A \succ 0$ and $B \succeq 0$:

$$(A + B)^{1/2} \preceq A^{1/2} + D_A f(B),$$

where $D_A f(B)$ is the Fréchet derivative of f at A in direction B. The Fréchet derivative of the matrix square root satisfies the Sylvester equation $A^{1/2} D_A f(B) + D_A f(B) A^{1/2} = B$, and has the integral representation:

$$D_A f(B) = \int_0^\infty (sI + A^{1/2})^{-1} B (sI + A^{1/2})^{-1} ds.$$

Taking traces and using the cyclic property:

$$\begin{aligned}\operatorname{tr}(D_A f(B)) &= \int_0^\infty \operatorname{tr}((sI + A^{1/2})^{-2} B)\, ds \\ &= \operatorname{tr}\left(B \int_0^\infty (sI + A^{1/2})^{-2} ds\right) \\ &= \tfrac{1}{2} \operatorname{tr}(A^{-1/2} B).\end{aligned}$$

The integral evaluates to $\frac{1}{2} A^{-1/2}$ since for a positive definite matrix M, $\int_0^\infty (sI + M)^{-2} ds = M^{-1}$. Applying the trace to both sides of the operator concavity inequality yields the result. □

A.2 Supporting Lemmas

Lemma A.3 (Update Rule for R_t)**.** *Let $R_t \in \mathbb{R}^{n \times n}$ be defined as follows*

$$R_t = \arg\min_R \mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top, L_t \otimes R),$$

minimizing the LogDet divergence between the greedy approximation $H_t' = (L_{t-1} \otimes R_{t-1} + g_t g_t^\top)$ and $L_t \otimes R$ (using row-major vectorization $g_t = \overline{\mathrm{vec}}(G_t)$), then optimal R^* takes the following closed form:

$$R^* = (\operatorname{tr}(L_t^{-1} L_{t-1}) R_{t-1} + G_t^\top L_t^{-1} G_t)/n$$

Proof. The optimization problem can be simplified as follows. Let $A = L_{t-1} \otimes R_{t-1} + g_t g_t^\top$ and $B = L_t \otimes R$ (using row-major vectorization $g_t = \overline{\mathrm{vec}}(G_t)$):

$$\begin{aligned}&\min_R \mathrm{D}_{\ell\mathrm{d}}(A, B) \\ &= \min_R -\log\det((L_t^{-1} \otimes R^{-1})A) \\ &\quad + \operatorname{tr}((L_t^{-1} \otimes R^{-1})A) - mn \\ &= \min_R -\log\det(L_t^{-1} \otimes R^{-1}) \\ &\quad + \operatorname{tr}(L_t^{-1} L_{t-1} \otimes R^{-1} R_{t-1}) + \operatorname{tr}((L_t^{-1} \otimes R^{-1}) g_t g_t^\top) \\ &= \min_R -n \log\det(R^{-1}) + \operatorname{tr}(L_t^{-1} L_{t-1}) \operatorname{tr}(R^{-1} R_{t-1}) \\ &\quad + g_t^\top (L_t^{-1} \otimes R^{-1}) g_t \\ &= \min_R -n \log\det(R^{-1}) + \operatorname{tr}(L_t^{-1} L_{t-1}) \operatorname{tr}(R^{-1} R_{t-1}) \\ &\quad + \operatorname{tr}(R^{-1} G_t^\top L_t^{-1} G_t) \\ &= \min_R -n \log\det(R^{-1}) \\ &\quad + \operatorname{tr}(R^{-1}(\operatorname{tr}(L_t^{-1} L_{t-1}) R_{t-1} + G_t^\top L_t^{-1} G_t)) \\ &= \min_R -n \cdot \mathrm{D}_{\ell\mathrm{d}}((\operatorname{tr}(L_t^{-1} L_{t-1}) R_{t-1} \\ &\quad + G_t^\top L_t^{-1} G_t)/n, R)\end{aligned}$$

The second equality uses $(A \otimes B)(C \otimes D) = (AC \otimes BD)$. The third equality uses $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \operatorname{tr}(B)$. The fourth equality uses row-major identity $g_t^\top (L^{-1} \otimes R^{-1}) g_t = \operatorname{tr}(R^{-1} G_t^\top L^{-1} G_t)$. The optimization problem after last equality is solved by

$$R^* = (\operatorname{tr}(L_t^{-1} L_{t-1}) R_{t-1} + G_t^\top L_t^{-1} G_t)/n$$

□

Lemma A.4 (Gupta et al. (2018))**.** *Let G be an $m \times n$ matrix of rank at most r and denote $g = \overline{\mathrm{vec}}(G)$ (row-major). Then,*

$$\frac{1}{r} g g^\top \preceq I_m \otimes G^\top G \quad \textit{and} \quad \frac{1}{r} g g^\top \preceq G G^\top \otimes I_n.$$

Proof. Write the singular value decomposition $G = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$, where $\sigma_i \geq 0$ for all i, and $u_1, \ldots, u_r \in \mathbb{R}^m$ and $v_1, \ldots, v_r \in \mathbb{R}^n$ are orthonormal sets of vectors. For row-major vectorization, $g = \overline{\text{vec}}(G) = \sum_{i=1}^{r} \sigma_i (u_i \otimes v_i)$ and hence,

$$gg^\top = \left(\sum_{i=1}^{r} \sigma_i (u_i \otimes v_i)\right)\left(\sum_{i=1}^{r} \sigma_i (u_i \otimes v_i)\right)^\top.$$

Next, we use the fact that for any set of vectors $w_1, \ldots, w_r$,

$$\left(\sum_{i=1}^{r} w_i\right)\left(\sum_{i=1}^{r} w_i\right)^\top \preceq r \sum_{i=1}^{r} w_i w_i^\top,$$

which holds since given a vector x we can write $\alpha_i = x^\top w_i$, and use the convexity of $\alpha \mapsto \alpha^2$ to obtain

$$\begin{aligned} x^\top \Big(\sum_{i=1}^{r} w_i\Big)\Big(\sum_{i=1}^{r} w_i\Big)^\top x &= \Big(\sum_{i=1}^{r} \alpha_i\Big)^2 \\ &\preceq r \sum_{i=1}^{r} \alpha_i^2 = r x^\top \Big(\sum_{i=1}^{r} w_i w_i^\top\Big) x. \end{aligned}$$

Using this fact and $(A \otimes B)(C \otimes D) = (AC \otimes BD)$, we can rewrite,

$$\begin{aligned} gg^\top &= \left(\sum_{i=1}^{r} \sigma_i (u_i \otimes v_i)\right)\left(\sum_{i=1}^{r} \sigma_i (u_i \otimes v_i)\right)^\top \\ &\preceq r \sum_{i=1}^{r} \sigma_i^2 (u_i \otimes v_i)(u_i \otimes v_i)^\top \\ &= r \sum_{i=1}^{r} \sigma_i^2 (u_i u_i^\top) \otimes (v_i v_i^\top). \end{aligned}$$

Now, since $GG^\top = \sum_{i=1}^{r} \sigma_i^2 u_i u_i^\top$ and $v_i v_i^\top \preceq I_n$ for all i, we have

$$\frac{1}{r} gg^\top \preceq \sum_{i=1}^{r} \sigma_i^2 (u_i u_i^\top) \otimes I_n = (GG^\top) \otimes I_n.$$

Similarly, using $G^\top G = \sum_{i=1}^{r} \sigma_i^2 v_i v_i^\top$ and $u_i u_i^\top \preceq I_m$ for all i, we obtain the second matrix inequality. □

A.3 Gap Theorem Proof

Theorem A.5 (Gap theorem for approximate preconditioners)**.** *Let $\{H_t\}_{t=0}^{T}$ and $\{\tilde{H}_t\}_{t=0}^{T}$ be sequences of positive definite matrices satisfying:*

- ***Initial condition:*** *$H_0 = \tilde{H}_0 = \epsilon I$ for some $\epsilon > 0$*
- ***Exact update:*** *$H_t = H_{t-1} + g_t g_t^\top$*
- ***Approximate update:*** *$\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ (Loewner order condition)*

Let $B_t = \tilde{H}_{t-1} + g_t g_t^\top$ be the intermediate target at step t. Then:

$$\delta_{\mathrm{S}}(\tilde{H}_T, H_T) \leq \sum_{t=1}^{T} \delta_{\mathrm{S}}(\tilde{H}_t, B_t),$$

$$\delta_{\mathrm{S}}{}^2(\tilde{H}_T, H_T) \leq T \sum_{t=1}^{T} \delta_{\mathrm{S}}{}^2(\tilde{H}_t, B_t).$$

Proof. The proof relies on two key properties of the S-divergence (Sra, 2016):

1. **Metric property:** δ_{S} is a metric on the cone of positive definite matrices, satisfying the triangle inequality.
2. **Contraction under translation:** For $X, Y \succ 0$ and $A \succeq 0$: ${\delta_{\mathrm{S}}}^2(A+X, A+Y) \leq {\delta_{\mathrm{S}}}^2(X, Y)$.

Step 1: Telescoping identity. Define the local gap $\Delta_t := \tilde{H}_t - (\tilde{H}_{t-1} + g_t g_t^\top) \succeq 0$ and cumulative gap $S_t := \sum_{s=1}^{t} \Delta_s$. By induction on t:

$$\tilde{H}_t = H_t + S_t, \quad B_t = H_t + S_{t-1}.$$

Base case ($t = 0$): $\tilde{H}_0 = H_0 = \epsilon I$ and $S_0 = 0$, so $\tilde{H}_0 = H_0 + S_0$. Inductive step: $\tilde{H}_t = \tilde{H}_{t-1} + g_t g_t^\top + \Delta_t = (H_{t-1} + S_{t-1}) + g_t g_t^\top + \Delta_t = H_t + S_t$.

Step 2: Rewrite local divergences. Using the telescoping identity:

$$\delta_{\mathrm{S}}(\tilde{H}_t, B_t) = \delta_{\mathrm{S}}(H_t + S_t, H_t + S_{t-1}).$$

Step 3: Apply triangle inequality. Since δ_{S} is a metric:

$$\delta_{\mathrm{S}}(H_T + S_T, H_T + S_0) \leq \sum_{t=1}^{T} \delta_{\mathrm{S}}(H_T + S_t, H_T + S_{t-1}).$$

Since $S_0 = 0$, the LHS equals $\delta_{\mathrm{S}}(\tilde{H}_T, H_T)$.

Step 4: Apply contraction property. Define the "remaining gradients" $R_t := H_T - H_t = \sum_{s=t+1}^{T} g_s g_s^\top \succeq 0$. For each term:

$$H_T + S_t = (H_t + R_t) + S_t,$$
$$H_T + S_{t-1} = (H_t + R_t) + S_{t-1}.$$

By the contraction property (adding $R_t \succeq 0$ reduces the divergence):

$$\begin{aligned} {\delta_{\mathrm{S}}}^2(H_T + S_t, H_T + S_{t-1}) &\leq {\delta_{\mathrm{S}}}^2(H_t + S_t, H_t + S_{t-1}) \\ &= {\delta_{\mathrm{S}}}^2(\tilde{H}_t, B_t). \end{aligned}$$

Taking square roots: $\delta_{\mathrm{S}}(H_T + S_t, H_T + S_{t-1}) \leq \delta_{\mathrm{S}}(\tilde{H}_t, B_t)$.

Step 5: Combine. Substituting into the triangle inequality:

$$\begin{aligned} \delta_{\mathrm{S}}(\tilde{H}_T, H_T) &\leq \sum_{t=1}^{T} \delta_{\mathrm{S}}(H_T + S_t, H_T + S_{t-1}) \\ &\leq \sum_{t=1}^{T} \delta_{\mathrm{S}}(\tilde{H}_t, B_t). \end{aligned}$$

Step 6: Cauchy-Schwarz for squared bound. For non-negative a_t:

$$\left(\sum_{t=1}^{T} a_t \right)^2 \leq T \sum_{t=1}^{T} a_t^2.$$

Applying this to $a_t = \delta_{\mathrm{S}}(\tilde{H}_t, B_t)$:

$$\begin{aligned} {\delta_{\mathrm{S}}}^2(\tilde{H}_T, H_T) &\leq \left(\sum_{t=1}^{T} \delta_{\mathrm{S}}(\tilde{H}_t, B_t) \right)^2 \\ &\leq T \sum_{t=1}^{T} {\delta_{\mathrm{S}}}^2(\tilde{H}_t, B_t). \end{aligned}$$

□

A.4 General Regret Framework

Lemma A.6 (A general regret upperbound (Hazan, 2016)). *For any sequence of $H_1, \dots, H_T \succ 0$ and $H_0 = 0$, the sequence of predictions given by the update rule $w_{t+1} = w_t - \eta H_t^{-1/2} g_t$ has the following regret upperbound:*

$$\mathcal{R}(T) \leq \underbrace{\frac{1}{2\eta} \sum_{t=1}^{T} (w_t - w^*)^\top (H_t^{1/2} - H_{t-1}^{1/2})(w_t - w^*)}_{\mathcal{R}_1(T)} + \underbrace{\frac{\eta}{2} \sum_{t=1}^{T} g_t^\top H_t^{-1/2} g_t}_{\mathcal{R}_2(T)}$$

Proof. The update rule $w_{t+1} = w_t - \eta H_t^{-1/2} g_t$ can be rewritten as follows:

$$\begin{aligned} w_{t+1} - w^* &= w_t - w^* - \eta H_t^{-1/2} g_t \\ H_t^{1/2}(w_{t+1} - w^*) &= H_t^{1/2}(w_t - w^*) - \eta g_t \end{aligned}$$

Multiplying the above two equations gives the following:

$$\begin{aligned} &(w_{t+1} - w^*)^\top H_t^{1/2}(w_{t+1} - w^*) \\ &= (w_t - w^*)^\top H_t^{1/2}(w_t - w^*) + \eta^2 g_t^\top H_t^{-1/2} g_t \\ &\quad - \eta(w_t - w^*)^\top g_t - \eta\langle \cancel{H_t^{-1/2} g_t}, \cancel{H_t^{1/2}}(w_t - w^*)\rangle \\ &= (w_t - w^*)^\top H_t^{1/2}(w_t - w^*) + \eta^2 g_t^\top H_t^{-1/2} g_t \\ &\quad - 2\eta(w_t - w^*)^\top g_t \end{aligned}$$

Adding $2\eta(w_t - w^*)^\top g_t$ and subtracting $(w_{t+1} - w^*)^\top H_t^{1/2}(w_{t+1} - w^*)$ on both sides gives:

$$\begin{aligned} 2\eta(w_t - w^*)^\top g_t &= (w_t - w^*)^\top H_t^{1/2}(w_t - w^*) \\ &\quad - (w_{t+1} - w^*)^\top H_t^{1/2}(w_{t+1} - w^*) \\ &\quad + \eta^2 g_t^\top H_t^{-1/2} g_t \\ (w_t - w^*)^\top g_t &= \frac{1}{2\eta}((w_t - w^*)^\top H_t^{1/2}(w_t - w^*) \\ &\quad - (w_{t+1} - w^*)^\top H_t^{1/2}(w_t - w^*)) \\ &\quad + \frac{\eta}{2} g_t^\top H_t^{-1/2} g_t \end{aligned}$$

summing up for all iterations $t \in \{1, \dots, T\}$ gives the following:

$$\sum_{t=1}^{T} (w_t - w^*)^\top g_t = \underbrace{\frac{1}{2\eta} \sum_{t=1}^{T} (w_t - w^*)^\top (H_t^{1/2} - H_{t-1}^{1/2})(w_t - w^*)}_{\mathcal{R}_1(T)} + \underbrace{\frac{\eta}{2} \sum_{t=1}^{T} g_t^\top H_t^{-1/2} g_t}_{\mathcal{R}_2(T)}.$$

Given that left hand side can be lower bounded as follows: $f_t(w_t) - f_t(w^*) \leq (w_t - w^*)^\top g_t$ (convexity of f_t) gives the result. □

Lemma A.7 (Spectral radius bound for inverse inflation). *For symmetric positive definite matrices $H, \tilde{H}$ with $\tilde{H} \succeq H$, define the* inverse inflation factor $\tilde{\kappa} = \rho(\tilde{H}^{-1} H) \leq 1$ *(spectral radius). Then:*

$$\|H^{1/4} \tilde{H}^{-1/2} H^{1/4}\|_2 \leq \sqrt{\tilde{\kappa}}.$$

Proof. Let $C = \tilde{H}^{-1/2} H^{1/2}$. We use the following facts:

Step 1 (Eigenvalue equivalence): For any matrices X, Y, the products XY and YX share the same nonzero eigenvalues. Taking $X = H^{1/4}$ and $Y = \tilde{H}^{-1/2} H^{1/4}$, we get that $H^{1/4} \tilde{H}^{-1/2} H^{1/4}$ and $C = \tilde{H}^{-1/2} H^{1/2}$ have the same eigenvalues $\mu_1, \ldots, \mu_d$.

Step 2 (Singular value dominates eigenvalue): For any matrix M, $|\lambda_{\max}(M)| \le \sigma_{\max}(M) = \|M\|_2$.

Step 3 (Spectral radius via symmetric form): Since $H^{1/2}$ and $\tilde{H}^{-1/2}$ are symmetric, $C^\top C = H^{1/2} \tilde{H}^{-1} H^{1/2}$. By the $XY \sim YX$ equivalence, $H^{1/2} \tilde{H}^{-1} H^{1/2}$ and $\tilde{H}^{-1} H$ share eigenvalues. Since the former is symmetric:

$$\begin{aligned} \|C^\top C\|_2 &= \|H^{1/2} \tilde{H}^{-1} H^{1/2}\|_2 = \rho(H^{1/2} \tilde{H}^{-1} H^{1/2}) \\ &= \rho(\tilde{H}^{-1} H) = \tilde{\kappa}. \end{aligned}$$

Combining: $\|H^{1/4} \tilde{H}^{-1/2} H^{1/4}\|_2 = \lambda_{\max}(C) \le \sigma_{\max}(C) = \sqrt{\tilde{\kappa}}$. □

Lemma A.8 (Inflation factor bounds via Jeffreys LogDet divergence)**.** *Let $\tilde{H}, H \succ 0$ be symmetric positive definite matrices. Let $\lambda_1, \ldots, \lambda_d$ be eigenvalues of $H^{-1}\tilde{H}$. Define:*

- $\kappa = \rho(H^{-1}\tilde{H}) = \max_i \lambda_i$ *(relative inflation factor)*
- $\tilde{\kappa} = \rho(\tilde{H}^{-1} H) = \max_i (1/\lambda_i) = 1/\min_i \lambda_i$ *(inverse inflation factor)*
- $S = \mathrm{D}_{\ell\mathrm{d}}(\tilde{H}, H) + \mathrm{D}_{\ell\mathrm{d}}(H, \tilde{H})$ *(Jeffreys LogDet divergence)*

Then both κ and $\tilde{\kappa}$ satisfy:

$$\kappa \le S + 2, \qquad \tilde{\kappa} \le S + 2.$$

More precisely, both satisfy $\kappa, \tilde{\kappa} \le \frac{(S+2)+\sqrt{S(S+4)}}{2}$. Furthermore, if $\kappa \ne 1/\tilde{\kappa}$ (i.e., not all eigenvalues are equal), the joint bound holds:

$$g(\kappa) + g(\tilde{\kappa}) \le S, \quad \text{where } g(x) = x + 1/x - 2.$$

Proof. **Step 1: Express the symmetrized divergence spectrally.**

$$\begin{aligned} \mathrm{D}_{\ell\mathrm{d}}(\tilde{H}, H) &= \mathrm{tr}(H^{-1}\tilde{H}) - \log\det(H^{-1}\tilde{H}) - d \\ &= \sum_{i=1}^{d} (\lambda_i - \log\lambda_i - 1), \\ \mathrm{D}_{\ell\mathrm{d}}(H, \tilde{H}) &= \mathrm{tr}(\tilde{H}^{-1} H) - \log\det(\tilde{H}^{-1} H) - d \\ &= \sum_{i=1}^{d} (1/\lambda_i + \log\lambda_i - 1). \end{aligned}$$

Summing: $S = \sum_{i=1}^{d} (\lambda_i + 1/\lambda_i - 2) = \sum_{i=1}^{d} g(\lambda_i)$, where $g(x) = x + 1/x - 2$.

Step 2: Key symmetry of $g(x)$. For $x > 0$: $g(x) \ge 0$ with equality iff $x = 1$. Crucially, $g(x) = g(1/x)$ for all $x > 0$ (symmetric under inversion).

Step 3: Bound κ. Since $g(\lambda_i) \ge 0$ for all i:

$$S = \sum_{i=1}^{d} g(\lambda_i) \ge g(\lambda_{\max}) = g(\kappa) = \kappa + 1/\kappa - 2.$$

Step 4: Bound $\tilde{\kappa}$. Similarly, using the symmetry $g(x) = g(1/x)$:

$$S \ge g(\lambda_{\min}) = g(1/\tilde{\kappa}) = g(\tilde{\kappa}) = \tilde{\kappa} + 1/\tilde{\kappa} - 2.$$

Step 5: Solve for both. From $g(x) \le S$, i.e., $x + 1/x \le S + 2$, multiplying by $x > 0$:

$$x^2 - (S+2)x + 1 \le 0.$$

The roots are $\frac{(S+2)\pm\sqrt{S(S+4)}}{2}$. Since both $\kappa \geq 1$ and $\tilde{\kappa} \geq 1$ (in the general case) lie between the roots:

$$\kappa, \tilde{\kappa} \leq \frac{(S+2)+\sqrt{S(S+4)}}{2} \leq S+2.$$

Joint bound: When $\lambda_{\max} \neq \lambda_{\min}$, the terms $g(\lambda_{\max})$ and $g(\lambda_{\min})$ correspond to distinct eigenvalues, so:

$$S = \sum_{i=1}^{d} g(\lambda_i) \geq g(\kappa) + g(\tilde{\kappa}).$$

Optimality: When $\tilde{H} = H$, $S = 0$ and the bounds give $\kappa, \tilde{\kappa} \leq 1$, which is tight. □

Theorem A.9 (Regret bound of general approximations)**.** *For sequence of approximations $\tilde{H}_t$ which satisfy the Loewner order conditions: (i) $\tilde{H}_t \succeq \tilde{H}_{t-1}$ (monotonically non-decreasing), and (ii) $\tilde{H}_t \succeq H_t$ (dominating the true second-moment matrix), with $\operatorname{tr}(\tilde{H}_T^{1/2}) = o(T)$, the updates $w_{t+1} \coloneqq w_t - \eta \tilde{H}_t^{-1/2} g_t$ with $\eta = \frac{D}{\sqrt{2}}$ yield:*

$$\tilde{\mathcal{R}}(T) \leq \frac{D}{\sqrt{2}} \left(\operatorname{tr}(\tilde{H}_T^{1/2}) + \sqrt{\tilde{\kappa}} \operatorname{tr}(H_T^{1/2}) \right),$$

where D is the diameter of the constraint set $\mathcal{W}$, $\tilde{\kappa} = \max_t \rho(\tilde{H}_t^{-1} H_t) \leq 1$ is the inverse inflation factor*, $S_t = \mathrm{D}_{\ell\mathrm{d}}(\tilde{H}_t, H_t) + \mathrm{D}_{\ell\mathrm{d}}(H_t, \tilde{H}_t)$ is the Jeffreys LogDet divergence at step t, and $\bar{S}_T = \max_{t \leq T} S_t$. By Lemma A.8, $\tilde{\kappa} \leq \frac{(\bar{S}_T+2)+\sqrt{\bar{S}_T(\bar{S}_T+4)}}{2}$.*

Proof. Using Lemma A.6, we get the decomposition:

$$\tilde{\mathcal{R}}(T) \leq \underbrace{\frac{1}{2\eta} \sum_{t=1}^{T} (w_t - w^*)^\top (\tilde{H}_t^{1/2} - \tilde{H}_{t-1}^{1/2})(w_t - w^*)}_{\tilde{\mathcal{R}}_1(T)} + \underbrace{\frac{\eta}{2} \sum_{t=1}^{T} g_t^\top \tilde{H}_t^{-1/2} g_t}_{\tilde{\mathcal{R}}_2(T)}$$

Bounding $\tilde{\mathcal{R}}_2(T)$: For each t, define $\tilde{\kappa}_t = \rho(\tilde{H}_t^{-1} H_t) \leq 1$ (since $\tilde{H}_t \succeq H_t$). Using a Hölder-type argument:

$$\begin{aligned} g_t^\top \tilde{H}_t^{-1/2} g_t &= \operatorname{tr}(H_t^{-1/4} (H_t^{1/4} \tilde{H}_t^{-1/2} H_t^{1/4}) H_t^{-1/4} g_t g_t^\top) \\ &\leq \| H_t^{1/4} \tilde{H}_t^{-1/2} H_t^{1/4} \|_2 \cdot g_t^\top H_t^{-1/2} g_t \\ &\leq \sqrt{\tilde{\kappa}_t} \cdot g_t^\top H_t^{-1/2} g_t \end{aligned}$$

where the last inequality applies Lemma A.7. Let $\tilde{\kappa} = \max_t \tilde{\kappa}_t \leq 1$. Taking the max out and applying Lemma 3.1:

$$\sum_{t=1}^{T} g_t^\top \tilde{H}_t^{-1/2} g_t \leq \sqrt{\tilde{\kappa}} \sum_{t=1}^{T} g_t^\top H_t^{-1/2} g_t \leq \sqrt{\tilde{\kappa}} \cdot 2 \operatorname{tr}(H_T^{1/2})$$

Therefore:

$$\begin{aligned} \tilde{\mathcal{R}}_2(T) &= \frac{\eta}{2} \sum_{t=1}^{T} g_t^\top \tilde{H}_t^{-1/2} g_t \\ &\leq \frac{\eta}{2} \cdot \sqrt{\tilde{\kappa}} \cdot 2 \operatorname{tr}(H_T^{1/2}) = \frac{D}{\sqrt{2}} \sqrt{\tilde{\kappa}} \operatorname{tr}(H_T^{1/2}) \end{aligned}$$

When $\tilde{H}_t = H_t$, we have $\tilde{\kappa} = 1$, recovering the FULL-MATRIX ADAGRAD bound $\frac{D}{\sqrt{2}} \operatorname{tr}(H_T^{1/2})$.

Bounding $\tilde{\mathcal{R}}_1(T)$: Using condition (i) $\tilde{H}_t \succeq \tilde{H}_{t-1}$ and Lemma A.1(vi,vii,viii):

$$\tilde{\mathcal{R}}_1(T) \le \frac{D^2}{2\eta} \sum_{t=1}^{T} \operatorname{tr}(\tilde{H}_t^{1/2} - \tilde{H}_{t-1}^{1/2}) = \frac{D^2}{2\eta} \operatorname{tr}(\tilde{H}_T^{1/2})$$

Setting $\eta = D/\sqrt{2}$ and combining:

$$\begin{aligned}\tilde{\mathcal{R}}(T) &\le \frac{D}{\sqrt{2}} \operatorname{tr}(\tilde{H}_T^{1/2}) + \frac{D}{\sqrt{2}} \sqrt{\tilde{\kappa}} \operatorname{tr}(H_T^{1/2}) \\ &= \frac{D}{\sqrt{2}} \left(\operatorname{tr}(\tilde{H}_T^{1/2}) + \sqrt{\tilde{\kappa}} \operatorname{tr}(H_T^{1/2}) \right).\end{aligned}$$

When $\tilde{H}_t = H_t$ (exact FULL-MATRIX ADAGRAD), $\tilde{\kappa} = 1$, recovering $\frac{D}{\sqrt{2}}(1+1)\operatorname{tr}(H_T^{1/2}) = \sqrt{2} D \operatorname{tr}(H_T^{1/2})$.

Connection to LogDet divergence: Let $S_t = \mathrm{D}_{\ell\mathrm{d}}(\tilde{H}_t, H_t) + \mathrm{D}_{\ell\mathrm{d}}(H_t, \tilde{H}_t)$ be the Jeffreys LogDet divergence at step t, and $\bar{S}_T = \max_{t \le T} S_t$. By Lemma A.8 applied at each step t, we have $\tilde{\kappa}_t \le \frac{(S_t+2)+\sqrt{S_t(S_t+4)}}{2}$. Taking the maximum over t:

$$\tilde{\kappa} = \max_t \tilde{\kappa}_t \le \frac{(\bar{S}_T + 2) + \sqrt{\bar{S}_T(\bar{S}_T + 4)}}{2}.$$

When $\bar{S}_T = 0$ (perfect approximation at all times), this gives $\tilde{\kappa} \le 1$, which is tight since $\tilde{\kappa} = 1$ when $\tilde{H}_t = H_t$. When $\bar{S}_T$ is small (good approximation uniformly over time), $\tilde{\kappa}$ is close to 1, yielding $\sqrt{\tilde{\kappa}} \approx 1$ and a bound close to the optimal FULL-MATRIX ADAGRAD bound. □

Lemma A.10 (Sufficient condition for Loewner order)**.** *If $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ for all t, then both conditions (i) $\tilde{H}_t \succeq \tilde{H}_{t-1}$ and (ii) $\tilde{H}_t \succeq H_t$ hold.*

Proof. Condition (i) follows immediately since $g_t g_t^\top \succeq 0$. For condition (ii), summing from $s = 1$ to t with $\tilde{H}_0 = 0$: $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top \succeq \ldots \succeq \sum_{s=1}^{t} g_s g_s^\top = H_t$. □

A.5 FROM OCO REGRET TO NON-CONVEX STATIONARITY

While our analysis is in the OCO framework, the empirical workloads in Section 4 are non-convex deep-learning objectives. The reductions of Cutkosky et al. (2023); Ahn & Cutkosky (2024) show that any algorithm with OCO regret $\mathcal{R}(T)$ can be converted into a stochastic non-convex optimizer that reaches an approximate stationary point at rate $\mathcal{R}(T)/T$. Our coupled preconditioners attain $\mathcal{R}(T) = \mathcal{O}(\sqrt{T})$ (Corollary 3.4), which yields the $\mathcal{O}(1/\sqrt{T})$ rate to stationarity — matching the optimal first-order rate for stochastic non-convex optimization. The Loewner order constraint and coupled Kronecker structure pass through the reduction unchanged because they only constrain the preconditioner construction, not the reduction itself; the only structural requirement the reduction imposes is that $\tilde{H}_t$ be monotonically non-decreasing and dominate H_t, which is exactly what our Loewner constraint guarantees (Lemma A.10). Additionally, Cesa-Bianchi et al. (2004) show that an OCO regret upper bound $\mathcal{R}(T)$ translates to a convergence upper bound $\mathcal{R}(T)/T$ in stochastic convex optimization, so the same $\mathcal{O}(1/\sqrt{T})$ rate applies in that setting.

A.6 LOEWNER ORDER PROOFS FOR COUPLED PRECONDITIONERS

Lemma A.11 (Alternating minimization for coupled Kronecker factors)**.** *Let $\tilde{H}_{t-1} = L_{t-1} \otimes R_{t-1}$ where $G_t \in \mathbb{R}^{m \times n}$, $L_t \in \mathbb{R}^{m \times m}$, $R_t \in \mathbb{R}^{n \times n}$. Solving the subproblem $\min_{L,R} \mathrm{D}_{\ell\mathrm{d}}(\tilde{H}_{t-1} + g_t g_t^\top, L \otimes R)$ via alternating minimization yields:*

$$\begin{aligned}L_t &= \arg\min_L \mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top, L \otimes R_{t-1}) \\ &= L_{t-1} + G_t R_{t-1}^{-1} G_t^\top / n, \\ R_t &= \arg\min_R \mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top, L_t \otimes R) \\ &= \frac{\operatorname{tr}(L_{t-1} L_t^{-1})}{m} R_{t-1} + \frac{1}{m} G_t^\top L_t^{-1} G_t,\end{aligned}$$

where $g_t = \overline{\mathrm{vec}}(G_t)$.

Proof. **Solving for L:** Using transformation invariance with $M = I_m \otimes R_{t-1}^{-1/2}$:

$$\begin{aligned}&\mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top, L \otimes R_{t-1})\\ &= \mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes I_n + \bar{g}_t \bar{g}_t^\top, L \otimes I_n),\end{aligned}$$

where $\bar{g}_t = (I_m \otimes R_{t-1}^{-1/2}) g_t = \overline{\mathrm{vec}}(G_t R_{t-1}^{-1/2})$. Since $L \otimes I_n$ is block-diagonal (up to permutation), the LogDet divergence decomposes over n blocks. Let $A^{(i)} \in \mathbb{R}^{m \times m}$ denote the i-th diagonal block. Taking the gradient with respect to L^{-1} and setting to zero: $L^* = \frac{1}{n}\sum_{i=1}^n A^{(i)} = L_{t-1} + G_t R_{t-1}^{-1} G_t^\top / n$.

Solving for R: Expanding the LogDet divergence:

$$\begin{aligned}&\mathrm{D}_{\ell\mathrm{d}}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top, L_t \otimes R)\\ &= \mathrm{tr}(L_{t-1} L_t^{-1}) \,\mathrm{tr}(R_{t-1} R^{-1}) + \mathrm{tr}(G_t^\top L_t^{-1} G_t R^{-1})\\ &\quad + n \log\det(L_t) + m \log\det(R) - \text{const.}\end{aligned}$$

Taking the gradient with respect to R and setting to zero:

$$\begin{aligned}&- \mathrm{tr}(L_{t-1} L_t^{-1}) R^{-1} R_{t-1} R^{-1}\\ &\quad - R^{-1} G_t^\top L_t^{-1} G_t R^{-1} + m R^{-1} = 0\\ &mR = \mathrm{tr}(L_{t-1} L_t^{-1}) R_{t-1} + G_t^\top L_t^{-1} G_t\end{aligned}$$

Therefore: $R^* = \frac{\mathrm{tr}(L_{t-1} L_t^{-1})}{m} R_{t-1} + \frac{1}{m} G_t^\top L_t^{-1} G_t$. □

Lemma A.12 (Loewner order for partially coupled Kronecker factors)**.** *Setting R_t via the update rule:*

$$R_t \coloneqq \|L_t^{-1/2} L_{t-1} L_t^{-1/2}\|_2 R_{t-1} + r G_t^\top L_t^{-1} G_t \tag{11}$$

satisfies $L_t \otimes R_t \succeq L_{t-1} \otimes R_{t-1} + g_t g_t^\top$, where $r = \mathrm{rank}(G_t)$.

Proof. The right hand side of the constraint can be upperbounded as follows:

$$\begin{aligned}&L_{t-1} \otimes R_{t-1} + g_t g_t^\top\\ &= (L_t^{1/2} \otimes I_n)(L_t^{-1/2} \otimes I_n)((L_{t-1} \otimes R_{t-1}) + g_t g_t^\top)\\ &\quad \times (L_t^{-1/2} \otimes I_n)(L_t^{1/2} \otimes I_n)\\ &= (L_t^{1/2} \otimes I_n)((L_t^{-1/2} L_{t-1} L_t^{-1/2}) \otimes R_{t-1}\\ &\quad + (L_t^{-1/2} \otimes I_n) g_t g_t^\top (L_t^{-1/2} \otimes I_n))(L_t^{1/2} \otimes I_n)\end{aligned}$$

We now use Lemma A.4 as follows. Let $M = L_t^{1/2} \otimes I_n$ and $A = L_t^{-1/2} L_{t-1} L_t^{-1/2}$:

$$\begin{aligned}&M(A \otimes R_{t-1} + (L_t^{-1/2} \otimes I_n) g_t g_t^\top (L_t^{-1/2} \otimes I_n))M\\ &\preceq M(A \otimes R_{t-1} + I_m \otimes r G_t^\top L_t^{-1} G_t)M\\ &\preceq M(\|A\|_2 \cdot I_m \otimes R_{t-1} + I_m \otimes r G_t^\top L_t^{-1} G_t)M\\ &= M(I_m \otimes (\|A\|_2 \cdot R_{t-1} + r G_t^\top L_t^{-1} G_t))M\\ &= L_t \otimes (\|L_t^{-1/2} L_{t-1} L_t^{-1/2}\|_2 \cdot R_{t-1} + r G_t^\top L_t^{-1} G_t)\\ &= L_t \otimes R_t\end{aligned}$$

□

Lemma A.13 (Loewner order for fully coupled Kronecker factors)**.** *For the fully coupled preconditioner $\tilde{H}_t = L_t \otimes R_t$ with updates:*

$$\begin{aligned}L_t &\coloneqq L_{t-1} + \tilde{r}_R \cdot G_t R_{t-1}^{-1} G_t^\top,\\ R_t &\coloneqq R_{t-1} + \tilde{r}_L \cdot G_t^\top L_{t-1}^{-1} G_t,\end{aligned}$$

where $\tilde{r}_R = \max_t \mathrm{rank}(G_t R_{t-1}^{-1/2})$ and $\tilde{r}_L = \max_t \mathrm{rank}(L_{t-1}^{-1/2} G_t)$, the Loewner order $L_t \otimes R_t \succeq L_{t-1} \otimes R_{t-1} + g_t g_t^\top$ is satisfied. Since multiplying by invertible matrices preserves rank, $\tilde{r}_R = \tilde{r}_L = \tilde{r} \coloneqq \max_t \mathrm{rank}(G_t)$.

Proof. We split the right hand side into two halves:

$$L_{t-1} \otimes R_{t-1} + g_t g_t^\top$$
$$= \tfrac{1}{2}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top) + \tfrac{1}{2}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top)$$

First half: Multiply and divide by $(I_m \otimes R_{t-1}^{1/2})$:

$$\tfrac{1}{2}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top)$$
$$= \tfrac{1}{2}(I_m \otimes R_{t-1}^{1/2})(L_{t-1} \otimes I_n$$
$$+ (I_m \otimes R_{t-1}^{-1/2}) g_t g_t^\top (I_m \otimes R_{t-1}^{-1/2}))(I_m \otimes R_{t-1}^{1/2})$$

Using Lemma A.4 with $\bar{G} = G_t R_{t-1}^{-1/2}$ of rank $\tilde{r}_R$, since $(I_m \otimes R_{t-1}^{-1/2}) g_t g_t^\top (I_m \otimes R_{t-1}^{-1/2}) = \overline{\mathrm{vec}}(\bar{G})\,\overline{\mathrm{vec}}(\bar{G})^\top \preceq \tilde{r}_R \cdot \bar{G}\bar{G}^\top \otimes I_n = \tilde{r}_R \cdot G_t R_{t-1}^{-1} G_t^\top \otimes I_n$:

$$\tfrac{1}{2}(I_m \otimes R_{t-1}^{1/2})(L_{t-1} \otimes I_n$$
$$+ (I_m \otimes R_{t-1}^{-1/2}) g_t g_t^\top (I_m \otimes R_{t-1}^{-1/2}))(I_m \otimes R_{t-1}^{1/2})$$
$$\preceq \tfrac{1}{2}(I_m \otimes R_{t-1}^{1/2})((L_{t-1} + \tilde{r}_R \cdot G_t R_{t-1}^{-1} G_t^\top) \otimes I_n)$$
$$\times (I_m \otimes R_{t-1}^{1/2}) = \tfrac{1}{2} L_t \otimes R_{t-1}$$

Second half: Multiply and divide by $(L_{t-1}^{1/2} \otimes I_n)$:

$$\tfrac{1}{2}(L_{t-1} \otimes R_{t-1} + g_t g_t^\top)$$
$$= \tfrac{1}{2}(L_{t-1}^{1/2} \otimes I_n)(I_m \otimes R_{t-1}$$
$$+ (L_{t-1}^{-1/2} \otimes I_n) g_t g_t^\top (L_{t-1}^{-1/2} \otimes I_n))(L_{t-1}^{1/2} \otimes I_n)$$

Using Lemma A.4 with $\bar{G} = L_{t-1}^{-1/2} G_t$ of rank $\tilde{r}_L$, since $(L_{t-1}^{-1/2} \otimes I_n) g_t g_t^\top (L_{t-1}^{-1/2} \otimes I_n) = \overline{\mathrm{vec}}(\bar{G})\,\overline{\mathrm{vec}}(\bar{G})^\top \preceq \tilde{r}_L \cdot I_m \otimes \bar{G}^\top \bar{G} = \tilde{r}_L \cdot I_m \otimes G_t^\top L_{t-1}^{-1} G_t$:

$$\tfrac{1}{2}(L_{t-1}^{1/2} \otimes I_n)(I_m \otimes R_{t-1}$$
$$+ (L_{t-1}^{-1/2} \otimes I_n) g_t g_t^\top (L_{t-1}^{-1/2} \otimes I_n))(L_{t-1}^{1/2} \otimes I_n)$$
$$\preceq \tfrac{1}{2}(L_{t-1}^{1/2} \otimes I_n)(I_m \otimes (R_{t-1} + \tilde{r}_L \cdot G_t^\top L_{t-1}^{-1} G_t))$$
$$\times (L_{t-1}^{1/2} \otimes I_n) = \tfrac{1}{2} L_{t-1} \otimes R_t$$

Adding both halves: Since $L_t \succeq L_{t-1}$ (as $G_t R_{t-1}^{-1} G_t^\top \succeq 0$):

$$\begin{aligned} L_{t-1} \otimes R_{t-1} + g_t g_t^\top &\preceq \frac{1}{2} L_t \otimes R_{t-1} + \frac{1}{2} L_{t-1} \otimes R_t \\ &\preceq \frac{1}{2} L_t \otimes R_{t-1} + \frac{1}{2} L_t \otimes R_t \\ &= L_t \otimes \frac{1}{2}(R_{t-1} + R_t) \\ &\preceq L_t \otimes R_t \end{aligned}$$

The last inequality uses $R_t \succeq R_{t-1}$. □

A.7 Regret Bounds for Coupled Preconditioners

Theorem A.14 (Regret bound of partially coupled preconditioner)**.** *Using the approximation $\tilde{H}_t = L_t \otimes R_t$ with the update rule:*

$$\begin{aligned} L_t &\coloneqq L_{t-1} + G_t G_t^\top / n \\ R_t &\coloneqq c_t R_{t-1} + \tilde{r} G_t^\top L_t^{-1} G_t, \quad c_t = \|L_t^{-1/2} L_{t-1} L_t^{-1/2}\|_2 \\ W_{t+1} &\coloneqq W_t - \eta L_t^{-1/2} G_t R_t^{-1/2}, \quad \eta = \frac{D}{\sqrt{2}} \end{aligned}$$

where $G_t = \nabla f_t(W_t)$ is the gradient, $\tilde{r} = \max_t \operatorname{rank}(L_t^{-1/2} G_t) = \max_t \operatorname{rank}(G_t)$ (since $L_t^{-1/2}$ is invertible), and $f_t : \mathbb{R}^{m\times n} \to \mathbb{R}$, $t \in \{1, \dots, T\}$ are sequence of functions to be minimized in an online fashion, then the regret with respect to minima $W^* = \min_W \sum_t f_t(W)$ can be upperbounded:

$$\sum_{t=1}^{T} f_t(W_t) - f(W^*) \leq \sqrt{2} D \operatorname{tr}(L_T^{1/2}) \operatorname{tr}(R_T^{1/2})$$

Applied to the regularized preconditioner $\bar{L}_t = L_t + \delta_t I$, $\bar{R}_t = R_t + \delta_t I$ with $\delta_t = \sum_{s\leq t} c/\sqrt{s}$, the same identity gives $\sum_t f_t(W_t) - f(W^*) \leq \sqrt{2} D \operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) = \mathcal{O}(Dmng\sqrt{\tilde{r}T})$ under bounded gradients $G_t G_t^\top \preceq g^2 I$ (Theorem A.18).

Proof. Using Lemma A.12 we know that for $\tilde{H}_t = L_t \otimes R_t$, the constraint $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ is satisfied. Given that the update being made can be viewed as follows (where $g_t = \overline{\operatorname{vec}}(G_t)$ and $w_t = \overline{\operatorname{vec}}(W_t)$, using row-major vectorization):

$$w_{t+1} \coloneqq w_t - \eta (L_t^{-1/2} \otimes R_t^{-1/2}) g_t = w_t - \eta \tilde{H}_t^{-1/2} g_t$$

Theorem A.9 can be used to upperbound regret as follows:

$$\begin{aligned}\tilde{\mathcal{R}}(T) &= \sum_{t=1}^{T} f_t(W_t) - f(W^*) \leq \sqrt{2} D \operatorname{tr}(\tilde{H}_T^{1/2}) \\ &= \sqrt{2} D \operatorname{tr}(L_T^{1/2} \otimes R_T^{1/2}) = \sqrt{2} D \operatorname{tr}(L_T^{1/2}) \operatorname{tr}(R_T^{1/2})\end{aligned}$$

Substituting $\bar{L}_T, \bar{R}_T$ for L_T, R_T above and applying Lemma A.17 gives $\operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) = \mathcal{O}(mn\, g\sqrt{\tilde{r}T})$ at $c = g\sqrt{\tilde{r}}$. □

Theorem A.15 (Regret bound of fully coupled preconditioner)**.** *Using the approximation $\tilde{H}_t = L_t \otimes R_t$ with the update rule:*

$$\begin{aligned}L_t &\coloneqq L_{t-1} + \tilde{r}_R \cdot G_t R_{t-1}^{-1} G_t^\top \\ R_t &\coloneqq R_{t-1} + \tilde{r}_L \cdot G_t^\top L_{t-1}^{-1} G_t \\ W_{t+1} &\coloneqq W_t - \eta L_t^{-1/2} G_t R_t^{-1/2}, \quad \eta = \frac{D}{\sqrt{2}}\end{aligned}$$

where $\tilde{r}_R = \max_t \operatorname{rank}(G_t R_{t-1}^{-1/2})$, $\tilde{r}_L = \max_t \operatorname{rank}(L_{t-1}^{-1/2} G_t)$, $G_t = \nabla f_t(W_t)$ is the gradient, and $f_t : \mathbb{R}^{m\times n} \to \mathbb{R}$, $t \in \{1, \dots, T\}$ are functions to be minimized online, the regret with respect to minima $W^* = \arg\min_W \sum_t f_t(W)$ can be upperbounded:

$$\sum_{t=1}^{T} f_t(W_t) - f(W^*) \leq \sqrt{2} D \operatorname{tr}(L_T^{1/2}) \operatorname{tr}(R_T^{1/2})$$

Applied to the regularized preconditioner $\bar{L}_t = L_t + \delta_t I$, $\bar{R}_t = R_t + \delta_t I$ with $\delta_t = \sum_{s\leq t} c/\sqrt{s}$, the same identity yields $\sum_t f_t(W_t) - f(W^*) \leq \sqrt{2} D \operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) = \mathcal{O}(Dmng\sqrt{\tilde{r}T})$ under bounded gradients $G_t G_t^\top \preceq g^2 I$ (Theorem A.18).

Proof. Using Lemma A.13 we know that for $\tilde{H}_t = L_t \otimes R_t$, the constraint $\tilde{H}_t \succeq \tilde{H}_{t-1} + g_t g_t^\top$ is satisfied. The regret bound follows analogously to Theorem A.14:

$$\tilde{\mathcal{R}}(T) \leq \sqrt{2} D \operatorname{tr}(\tilde{H}_T^{1/2}) = \sqrt{2} D \operatorname{tr}(L_T^{1/2}) \operatorname{tr}(R_T^{1/2})$$

Substituting $\bar{L}_T, \bar{R}_T$ for L_T, R_T above and applying Lemma A.17 yields the optimal $\mathcal{O}(Dmng\sqrt{\tilde{r}T})$ regret rate at $c = g\sqrt{\tilde{r}}$. □

Regularized variant. For padded preconditioners $\bar{L}_t = L_t + \delta_t I$, $\bar{R}_t = R_t + \delta_t I$ with $\delta_t = \sum_{s\leq t} c/\sqrt{s}$, the Loewner constraint $\bar{L}_t \otimes \bar{R}_t \succeq \bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top$ holds by Lemma A.16. Substituting $\bar{L}_T, \bar{R}_T$ for L_T, R_T in the displayed bound gives $\sum_t f_t(W_t) - f(W^*) \leq \sqrt{2} D \operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2})$, controlled by Lemma A.17.

A.8 Scaling Analysis and Optimal Regret Proofs

We provide scalar intuition, proof sketches, and full proofs for the scaling analysis results in Section 3.4.

A.8.1 Scalar Intuition

Fully coupled case. Consider the scalar recurrence $\ell_t = \ell_{t-1} + g^2/r_{t-1}$ and $r_t = r_{t-1} + g^2/\ell_{t-1}$ with $\ell_0 = r_0$. By symmetry, $\ell_t = r_t$ for all t. The product $P_t = \ell_t \cdot r_t$ satisfies:

$$P_t = P_{t-1} + 2g^2 + g^4/P_{t-1} \geq P_{t-1} + 2g^2,$$

yielding $P_t = \Theta(g^2 t)$ and thus $\ell_t = r_t = \Theta(g\sqrt{t})$. The coupling creates a *self-regulating feedback*: when one factor grows, its inverse shrinks the increment to the other factor.

Partially coupled case. For the partially coupled case $\ell_t = \ell_{t-1} + g^2$ and $r_t = c_t r_{t-1} + g^2/\ell_{t-1}$ with shrinkage $c_t \leq 1$, we have $\ell_t = \Theta(g^2 t)$. Without shrinkage ($c_t = 1$), the increment $g^2/\ell_{t-1} \approx 1/t$ sums to $r_T = \Theta(\log T)$ via the harmonic series. With optimal shrinkage $c_t = 1 - \epsilon/t$ (arising from $c_t = \|\ell_t^{-1}\ell_{t-1}\|$), the contraction and increment balance exactly: $r_t = c_t r_{t-1} + 1/t$ yields $r_t = \Theta(1)$, eliminating the $\log T$ factor.

Matrix generalization. Under the assumption that gradients have bounded minimum singular value $G_t G_t^\top \succeq \epsilon g^2 I_m$ ("square-shaped" gradients), the scalar intuition extends to matrices. This assumption ensures uniform eigenvalue growth and bounded condition numbers $\kappa(L_t), \kappa(R_t) = \mathcal{O}(1/\epsilon)$.

A.8.2 Proof Sketches

For an alternative analysis of the regularized preconditioners $\bar{L}_t, \bar{R}_t$ under the bounded-gradient assumption $G_t G_t^\top \preceq g^2 I$, see Lemmas A.16, A.17 and Theorem A.18 below.

A.8.3 Full Proofs

We now provide full proofs. Throughout, we assume the "square-shaped" gradient condition: there exist constants $g > 0$ and $\epsilon \in (0, 1]$ such that $\epsilon g^2 I_m \preceq G_t G_t^\top \preceq g^2 I_m$ for all t.

Notation. Let $\lambda_t = \lambda_{\min}(L_t)$, $\Lambda_t = \lambda_{\max}(L_t)$, $\mu_t = \lambda_{\min}(R_t)$, $M_t = \lambda_{\max}(R_t)$, and $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ denote the condition number.

A.8.4 Regularized preconditioners

We now develop full proofs for the regularized coupled Kronecker preconditioners under the bounded-gradient assumption $G_t G_t^\top \preceq g^2 I$.

Padded preconditioners. For any constant $c > 0$, define

$$\delta_t \coloneqq \sum_{s=1}^{t} \frac{c}{\sqrt{s}}, \qquad \bar{L}_t \coloneqq L_t + \delta_t I_m, \qquad \bar{R}_t \coloneqq R_t + \delta_t I_n,$$

and update the parameters via $W_{t+1} \coloneqq W_t - \eta\, \bar{L}_t^{-1/2} G_t \bar{R}_t^{-1/2}$. The padded updates are equivalent to standard ϵ-regularization applied before each inverse square root and obey

$$\bar{L}_t = \bar{L}_{t-1} + \tilde{r}\, G_t \bar{R}_{t-1}^{-1} G_t^\top + \tfrac{c}{\sqrt{t}} I_m, \qquad \bar{R}_t = \bar{R}_{t-1} + \tilde{r}\, G_t^\top \bar{L}_{t-1}^{-1} G_t + \tfrac{c}{\sqrt{t}} I_n.$$

Here $\tilde{r} = \max_t \operatorname{rank}(G_t)$ as in Lemma A.4, and the partially coupled case corresponds to fixing $\bar{R}_{t-1} = I_n$ inside the $\bar{L}_t$-update (equivalent statements hold by Lemma A.12).

Lemma A.16 (Loewner order for regularized coupled preconditioner)**.** *For the padded preconditioners above,*

$$\bar{L}_t \otimes \bar{R}_t \;\succeq\; \bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top, \qquad g_t = \overline{\mathrm{vec}}(G_t).$$

Proof. Write $\Delta_L = \tilde{r}\, G_t \bar{R}_{t-1}^{-1} G_t^\top$ and $\Delta_R = \tilde{r}\, G_t^\top \bar{L}_{t-1}^{-1} G_t$. Following Lemma A.13's half-splitting strategy, transform the first half of $g_t g_t^\top$ by $(I_m \otimes \bar{R}_{t-1}^{1/2})$. Applying Lemma A.4 to $\hat{G} = G_t \bar{R}_{t-1}^{-1/2}$ of rank $\tilde{r}$ gives

$$\tfrac{1}{2}\big(\bar{L}_{t-1} \otimes I_n + \hat{g}\hat{g}^\top\big) \;\preceq\; \tfrac{1}{2}\big(\bar{L}_{t-1} + \Delta_L\big) \otimes I_n,$$

and sandwiching by $(I_m \otimes \bar{R}_{t-1}^{1/2})$ on both sides yields

$$\text{(H1)} \quad \tfrac{1}{2}\big(\bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top\big) \;\preceq\; \tfrac{1}{2}(\bar{L}_{t-1} + \Delta_L) \otimes \bar{R}_{t-1} \;\preceq\; \tfrac{1}{2}\,\bar{L}_t \otimes \bar{R}_{t-1},$$

where the last step uses $\bar{L}_t \succeq \bar{L}_{t-1} + \Delta_L$ since $(c/\sqrt{t})I_m \succeq 0$. Applying the symmetric construction with $\hat{G} = \bar{L}_{t-1}^{-1/2} G_t$ of rank $\tilde{r}$ and transformation $(\bar{L}_{t-1}^{1/2} \otimes I_n)$ gives

$$\text{(H2)} \quad \tfrac{1}{2}\big(\bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top\big) \;\preceq\; \tfrac{1}{2}\,\bar{L}_{t-1} \otimes \bar{R}_t.$$

Adding (H1) and (H2):

$$\bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top \;\preceq\; \tfrac{1}{2}\,\bar{L}_t \otimes \bar{R}_{t-1} + \tfrac{1}{2}\,\bar{L}_{t-1} \otimes \bar{R}_t.$$

Since $\bar{L}_t \succeq \bar{L}_{t-1}$ and $\bar{R}_t \succeq \bar{R}_{t-1}$ (both increments are PSD), each summand is bounded above by $\frac{1}{2}\bar{L}_t \otimes \bar{R}_t$, giving the claim. □

Lemma A.17 (Scaling of regularized preconditioners)**.** *Assume $G_t G_t^\top \preceq g^2 I_m$ for all t. Let*

$$\bar{\Lambda}_t = \lambda_{\max}(\bar{L}_t), \qquad \bar{\mu}_t = \lambda_{\min}(\bar{R}_t), \qquad A \;=\; \frac{\tilde{r} g^2}{c} + c.$$

*Then **Step 1 (eigenvalue floor):** $\bar{\mu}_t \geq c\sqrt{t}$ for $t \geq 4$, and symmetrically $\lambda_{\min}(\bar{L}_t) \geq c\sqrt{t}$. **Step 2 (max eigenvalue):** $\bar{\Lambda}_T \leq \mathcal{O}(1) + 2A\sqrt{T}$ and $\lambda_{\max}(\bar{R}_T) \leq \mathcal{O}(1) + 2A\sqrt{T}$. **Step 3 (trace product):** $\operatorname{tr}(\bar{L}_T^{1/2})\, \operatorname{tr}(\bar{R}_T^{1/2}) \leq mn\big(\mathcal{O}(1) + 2A\sqrt{T}\big) = \mathcal{O}(mn\, A\sqrt{T})$.*

Proof. **Step 1.** The Lemma A.16 update gives $\bar{R}_t = \bar{R}_{t-1} + \tilde{r}\, G_t^\top \bar{L}_{t-1}^{-1} G_t + (c/\sqrt{t})I_n$, where the first increment is PSD, so $\bar{\mu}_t \geq \bar{\mu}_{t-1} + c/\sqrt{t}$. Summing: $\bar{\mu}_t \geq \bar{\mu}_0 + c\sum_{s=1}^{t} s^{-1/2} \geq 1 + 2c(\sqrt{t} - 1) \geq c\sqrt{t}$ for $t \geq 4$. The bound on $\lambda_{\min}(\bar{L}_t)$ follows by the identical PSD argument applied to the $\bar{L}_t$ update.

Step 2. For any unit vector v, $v^\top (G_t \bar{R}_{t-1}^{-1} G_t^\top) v = (G_t^\top v)^\top \bar{R}_{t-1}^{-1} (G_t^\top v) \leq \frac{1}{\bar{\mu}_{t-1}} \|G_t^\top v\|^2 \leq \frac{g^2}{\bar{\mu}_{t-1}}$. Hence $G_t \bar{R}_{t-1}^{-1} G_t^\top \preceq (g^2/\bar{\mu}_{t-1}) I_m$, and the $\bar{L}_t$ update yields

$$\bar{\Lambda}_t \;\leq\; \bar{\Lambda}_{t-1} + \frac{\tilde{r} g^2}{\bar{\mu}_{t-1}} + \frac{c}{\sqrt{t}} \;\leq\; \bar{\Lambda}_{t-1} + \frac{\tilde{r} g^2}{c\sqrt{t-1}} + \frac{c}{\sqrt{t}} \;\leq\; \bar{\Lambda}_{t-1} + \frac{A}{\sqrt{t-1}}$$

for $t \geq 5$, after substituting Step 1 and using $A = \tilde{r} g^2/c + c$. Summing $\sum_{s=1}^{T-1} s^{-1/2} \leq 2\sqrt{T}$:

$$\bar{\Lambda}_T \;\leq\; \mathcal{O}(1) + A \cdot 2\sqrt{T} \;=\; \mathcal{O}(1) + 2A\sqrt{T}.$$

The bound on $\lambda_{\max}(\bar{R}_T)$ follows by the identical argument with the roles of L and R swapped, again using only $G_t G_t^\top \preceq g^2 I$ (which gives $G_t^\top G_t \preceq g^2 I$ as well, since both have the same nonzero eigenvalues).

Step 3. Each of the m eigenvalues of $\bar{L}_T^{1/2}$ is at most $\sqrt{\bar{\Lambda}_T}$, so $\operatorname{tr}(\bar{L}_T^{1/2}) \leq m\sqrt{\bar{\Lambda}_T}$. Likewise $\operatorname{tr}(\bar{R}_T^{1/2}) \leq n\sqrt{\bar{M}_T}$ with $\bar{M}_T = \lambda_{\max}(\bar{R}_T)$. Substituting Step 2 gives the claim. □

Theorem A.18 (Regularized scaling for fully and partially coupled preconditioners — full proof)**.** *Under the bounded-gradient assumption $G_t G_t^\top \preceq g^2 I$, both regularized variants satisfy*

$$\sum_{t=1}^{T} f_t(W_t) - f(W^*) \;\leq\; \sqrt{2}\, D\, \operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) \;=\; \mathcal{O}\big(D\, mn\, g\sqrt{\tilde{r}T}\big).$$

Specifically:

(a) Fully coupled regularized variant *(Lemma A.16, Lemma A.17) with* $\delta_t = \sum_{s=1}^{t} c/\sqrt{s}$ *and* $c = g\sqrt{\tilde{r}}$ *gives* $\operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) = \mathcal{O}(mn\, g\sqrt{\tilde{r}T})$.

(b) Partially coupled regularized variant *(Lemma A.19, Lemma A.20) with* $\delta_t = \rho t$ *and* $\rho = g^2$ *gives* $\operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) = \mathcal{O}(mn\, g\sqrt{\tilde{r}T})$.

Proof. For each variant, two ingredients are needed: (i) the Loewner constraint $\bar{L}_t \otimes \bar{R}_t \succeq \bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top$, which feeds into Theorem A.9 via the regret bounds of Theorems A.14 and A.15 (with $\bar{L}_T, \bar{R}_T$ in place of L_T, R_T); and (ii) the trace product bound on $\operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2})$.

Case (a): Lemma A.16 supplies (i); Lemma A.17 with $c = g\sqrt{\tilde{r}}$ supplies (ii).

Case (b): Lemma A.19 supplies (i); Lemma A.20 with $\rho = g^2$ supplies (ii).

In both cases, combining (i) with Theorems A.14 or A.15 yields the displayed regret bound, with the trace product on the right-hand side controlled by (ii). □

Dependency structure. The proof of Lemma A.17 forms an acyclic chain: Step 1 depends only on the regularization; Step 2 depends on Step 1 and the gradient bound $G_t G_t^\top \preceq g^2 I$; Step 3 depends on Step 2 and Lemma A.16.

Partially coupled regularized variant. The partially coupled construction uses a different regularization schedule from the fully coupled case (cf. Theorem A.18, where $\delta_t = \sum_{s=1}^{t} c/\sqrt{s}$). We use a distinct constant $\rho > 0$ here and define

$$\bar{L}_t = \bar{L}_{t-1} + G_t G_t^\top + \rho\, I_m, \qquad \bar{R}_t = \bar{c}_t \bar{R}_{t-1} + \tilde{r}\, G_t^\top \bar{L}_t^{-1} G_t,$$

where $\bar{c}_t = \|\bar{L}_t^{-1/2} \bar{L}_{t-1} \bar{L}_t^{-1/2}\|$. Equivalently, $\bar{L}_t = L_t + \rho t\, I$ where $L_t = L_{t-1} + G_t G_t^\top$ is the standard L accumulator: the per-step ρI provides both invertibility and the increment that drives shrinkage on $\bar{R}_t$. Note ρ is unrelated to the schedule constant c in Theorem A.18 – their roles and optimal values differ.

Lemma A.19 (Loewner order for partially coupled regularized preconditioner)**.** *The padded partially coupled preconditioners satisfy* $\bar{L}_t \otimes \bar{R}_t \succeq \bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top$.

Proof. Apply Lemma A.12 with $(\bar{L}_{t-1}, \bar{L}_t, \bar{R}_{t-1})$ in place of (L_{t-1}, L_t, R_{t-1}): since $\bar{L}_t \succeq \bar{L}_{t-1}$ (the increment $G_t G_t^\top + \rho I \succeq 0$), the lemma's hypotheses hold and yield $\bar{L}_t \otimes (\bar{c}_t \bar{R}_{t-1} + \tilde{r}\, G_t^\top \bar{L}_t^{-1} G_t) \succeq \bar{L}_{t-1} \otimes \bar{R}_{t-1} + g_t g_t^\top$. □

Lemma A.20 (Scaling of partially coupled regularized preconditioners)**.** *Assume* $G_t G_t^\top \preceq g^2 I_m$ *for all* t *(no lower bound). Set* $\eta = \rho/(g^2 + \rho)$*. Then:* ***Step 1 (*** L ***bounds):*** *With initialization* $\bar{L}_0 = 0$*:* $\lambda_{\min}(\bar{L}_t) \geq \rho t$ *and* $\bar{\Lambda}_t := \lambda_{\max}(\bar{L}_t) \leq (g^2 + \rho)t$*, so* $\operatorname{tr}(\bar{L}_T^{1/2}) \leq m\sqrt{\bar{\Lambda}_T} = \mathcal{O}(m\sqrt{(g^2 + \rho)T})$*.* ***Step 2 (shrinkage from regularization):*** $\bar{c}_t \leq 1 - \eta/t$ *without any assumption on* $G_t G_t^\top$ *beyond the upper bound.* ***Step 3 (bounded*** $\lambda_{\max}(\bar{R}_T)$***):*** $\lambda_{\max}(\bar{R}_T) \leq \tilde{r} g^2/(\rho\eta) = \tilde{r} g^2 (g^2 + \rho)/\rho^2 = \mathcal{O}(1)$ *in* T*, so* $\operatorname{tr}(\bar{R}_T^{1/2}) \leq n\sqrt{\lambda_{\max}(\bar{R}_T)} = \mathcal{O}(ng\sqrt{\tilde{r}(g^2 + \rho)/\rho})$*. Combining with Theorem A.15:*

$$\mathcal{R}(T) \leq \sqrt{2}\, D\, \operatorname{tr}(\bar{L}_T^{1/2}) \operatorname{tr}(\bar{R}_T^{1/2}) = \mathcal{O}\Big(D\, mn\, g \cdot \tfrac{g^2 + \rho}{\rho} \sqrt{\tilde{r}T}\Big).$$

The choice $\rho = g^2$ *gives* $\mathcal{R}(T) = \mathcal{O}\big(D\, mn\, g\sqrt{\tilde{r}T}\big)$.

Proof. **Step 1.** Initialize $\bar{L}_0 = 0$. The PSD increment $G_t G_t^\top \succeq 0$ plus the ρI regularization gives $\lambda_{\min}(\bar{L}_t) \geq \lambda_{\min}(\bar{L}_{t-1}) + \rho$, hence $\lambda_{\min}(\bar{L}_t) \geq \rho t$. Using $G_t G_t^\top \preceq g^2 I$ for the upper bound: $\bar{\Lambda}_t \leq \bar{\Lambda}_{t-1} + g^2 + \rho \leq (g^2 + \rho)t$. Each eigenvalue of $\bar{L}_T^{1/2}$ is at most $\sqrt{\bar{\Lambda}_T}$, so $\operatorname{tr}(\bar{L}_T^{1/2}) \leq m\sqrt{\bar{\Lambda}_T} = \mathcal{O}(m\sqrt{(g^2 + \rho)T})$.

Step 2. The shrinkage factor satisfies

$$\bar{c}_t = 1 - \lambda_{\min}\big(\bar{L}_t^{-1/2}(\bar{L}_t - \bar{L}_{t-1})\bar{L}_t^{-1/2}\big) = 1 - \lambda_{\min}\big(\bar{L}_t^{-1/2}(G_t G_t^\top + \rho I)\bar{L}_t^{-1/2}\big).$$

Since $G_t G_t^\top \succeq 0$, $\bar{L}_t^{-1/2}(G_t G_t^\top + \rho I)\bar{L}_t^{-1/2} \succeq \rho \bar{L}_t^{-1}$, so $\lambda_{\min} \geq \rho/\bar{\Lambda}_t \geq \rho/((g^2+\rho)t) = \eta/t$, giving $\bar{c}_t \leq 1 - \eta/t$. The shrinkage comes entirely from the regularization, regardless of the spectrum of $G_t G_t^\top$.

Step 3. For any unit $v \in \mathbb{R}^n$, $v^\top G_t^\top \bar{L}_t^{-1} G_t v = (G_t v)^\top \bar{L}_t^{-1}(G_t v) \leq \|G_t v\|^2/\lambda_{\min}(\bar{L}_t) \leq g^2/\lambda_{\min}(\bar{L}_t) \leq g^2/(\rho t)$, using $G_t G_t^\top \preceq g^2 I_m$ (equivalently $G_t^\top G_t \preceq g^2 I_n$, since $G_t G_t^\top$ and $G_t^\top G_t$ share the same nonzero eigenvalues, so $\lambda_{\max}(G_t^\top G_t) = \lambda_{\max}(G_t G_t^\top) \leq g^2$) and Step 1. Therefore $G_t^\top \bar{L}_t^{-1} G_t \preceq (g^2/(\rho t)) I_n$, and the $\bar{R}_t$ update yields the operator inequality

$$\bar{R}_t \;\preceq\; \bar{c}_t \bar{R}_{t-1} + \tfrac{\tilde{r} g^2}{\rho t}\, I_n.$$

Taking $\lambda_{\max}$ and using $\bar{c}_t \leq 1 - \eta/t$:

$$\lambda_{\max}(\bar{R}_t) \;\leq\; (1 - \eta/t)\, \lambda_{\max}(\bar{R}_{t-1}) + \tfrac{\tilde{r} g^2}{\rho t}.$$

We claim by induction that $\lambda_{\max}(\bar{R}_t) \leq K' := \tilde{r} g^2/(\rho\eta)$ for all $t \geq 0$. *Base case* ($t = 0$): with the standard initialization $\bar{R}_0 = 0$, $\lambda_{\max}(\bar{R}_0) = 0 \leq K'$. *Inductive step* ($t \geq 1$): assuming $\lambda_{\max}(\bar{R}_{t-1}) \leq K'$,

$$\lambda_{\max}(\bar{R}_t) \;\leq\; (1 - \eta/t)\, \lambda_{\max}(\bar{R}_{t-1}) + \tfrac{\tilde{r} g^2}{\rho t} \;\leq\; (1 - \eta/t)\, K' + \tfrac{\tilde{r} g^2}{\rho t} \;=\; K' - \tfrac{K'\eta - \tilde{r} g^2/\rho}{t} \;=\; K',$$

where the final equality uses the choice $K' = \tilde{r} g^2/(\rho\eta)$ so that $K'\eta - \tilde{r} g^2/\rho = 0$. Since each eigenvalue of $\bar{R}_T^{1/2}$ is at most $\sqrt{\lambda_{\max}(\bar{R}_T)}$, $\mathrm{tr}(\bar{R}_T^{1/2}) \leq n\sqrt{K'} = \mathcal{O}(n\sqrt{\tilde{r}(g^2+\rho)}/\rho)$. □

Discussion. Lemma A.20 requires only the upper bound $G_t G_t^\top \preceq g^2 I$. The shrinkage that previously came from a lower-bound assumption on $G_t G_t^\top$ is now supplied by the regularization itself: each step's ρI contribution to the $\bar{L}_t$ increment forces $\bar{c}_t \leq 1 - \eta/t$, exactly the rate needed to keep $\lambda_{\max}(\bar{R}_T)$ bounded in T. The regularization schedule is variant-specific: fully coupled (Theorem A.18) uses $\delta_t = \sum c/\sqrt{s} = \mathcal{O}(\sqrt{t})$ with $c = g\sqrt{\tilde{r}}$ to match its symmetric $\sqrt{T}$ growth; partially coupled uses $\delta_t = \rho t$ with $\rho = g^2$ to match L's linear accumulation and induce $\Theta(1/t)$ shrinkage on $\bar{R}_t$. The constants c and ρ are distinct (schedule-coefficient vs. per-step increment) with different optimal values.

A.9 Frobenius Norm Approximation

This section derives the Shampoo left factor $L_t = \frac{1}{n}\sum_{s=1}^t G_s G_s^\top$ as the solution to a matrix nearness problem in Frobenius norm.

Lemma A.21 (minimizing squared Frobenius norm between H_t and $I_n \otimes L$)**.** *When searching over all block-diagonal approximations $\{I_n \otimes L : L \succeq 0\}$, the optimal L approximating H_t is:*

$$L_t^* = \arg\min_L \left\| \sum_{s=1}^t g_s g_s^\top - I_n \otimes L \right\|_F^2 = \frac{1}{n}\sum_{s=1}^t G_s G_s^\top.$$

Proof. Since $I_n \otimes L$ is block-diagonal with n copies of $L \in \mathbb{R}^{m\times m}$, the Frobenius norm decomposes over diagonal blocks. The optimal L averages the diagonal blocks of H_t, giving $L^* = \frac{1}{n}\sum_{s=1}^t G_s G_s^\top$. □

A.10 Experimental Details and Hyperparameters

Benchmarks. We evaluate on 7 AlgoPerf benchmarks (Dahl et al., 2023):

- ImageNet ViT and ImageNet ResNet-50: Vision transformer (Dosovitskiy et al., 2020) and ResNet (He et al., 2016) on ImageNet (Deng et al., 2009)
- OGBG GNN: GNN (Scarselli et al., 2009) on OGBG-MOLPCBA (Hu et al., 2020)
- WMT Transformer: Transformer (Vaswani et al., 2017) on WMT 2017 De→En (Bojar et al., 2017)
- Criteo 1TB DLRMsmall: DLRM (Naumov et al., 2019) on Criteo (Lab, 2014)
- LibriSpeech Conformer and LibriSpeech DeepSpeech: Conformer (Gulati et al., 2020) and DeepSpeech (Amodei et al., 2016) on LibriSpeech (Panayotov et al., 2015)

Table 3: Table of different models, datasets, performance metrics, and steps budget in AlgoPerf which was used in Section 4.2. In Section 4.1, we used 0.75 fraction of steps budget.

Task	Dataset	Model	Loss	Metric	Validation Target	Test Target	Max Runtime (s)	Steps Budget
Clickthrough rate prediction	Criteo 1TB	DLRMsmall	CE	CE	0.123735	0.126041	7,703	10,666
Image classification	ImageNet	ResNet-50	CE	ER	0.22569	0.3440	63,008	186,666
Image classification	ImageNet	ViT	CE	ER	0.22691	0.3481	77,520	186,666
Speech recognition	LibriSpeech	Conformer	CTC	WER	0.085884	0.052981	61,068	80,000
Speech recognition	LibriSpeech	DeepSpeech	CTC	WER	0.119936	0.074143	55,506	48,000
Molecular property prediction	OGBG	GNN	CE	mAP	0.28098	0.268729	18,477	80,000
Translation	WMT	Transformer	CE	BLEU	30.8491	30.7219	48,151	133,333

Practical implementation. For experiments, we use exponential moving averages for second moments and precondition the momentum vector:

$$
\begin{aligned}
L_t &\coloneqq \beta_2 L_{t-1} + (1-\beta_2) G_t G_t^\top \\
R_t &\coloneqq (\beta_2 c_t R_{t-1} + (1-\beta_2) G_t^\top L_t^{-1} G_t)/m \\
\mu_t &\coloneqq \beta_1 \mu_{t-1} + (1-\beta_1) G_t \\
W_{t+1} &\coloneqq W_t - \eta L_t^{-1/2} \mu_t R_t^{-1/2}
\end{aligned}
$$

We set $c_t = 1$ as it gives similar results to $c_t = \|L_t^{-1/2} L_{t-1} L_t^{-1/2}\|_2$ while avoiding additional matrix multiplications. Preconditioners are computed once every few iterations. We use grafting (Agarwal et al., 2020) and blocking (Anil et al., 2020) for fair comparison. Experiments use JAX (Bradbury et al., 2018) with 8 TPUv5e chips per run.

Hyperparameter search. In Table 2 we outline the hyperparameter search space for all the benchmarks in Section 4.1.

Table 2: Hyperparameter search space used in Section 4.1.

Parameter	Min	Max	Scaling/Feasible Points
`learning_rate`	10^{-4}	10^{-2}	log
$1-\beta_1$	10^{-2}	0.15	log
β_2	-	-	0.9, 0.99, 0.999
`warmup_factor`	-	-	0.05
`weight_decay`	5×10^{-3}	1.0	log
`label_smoothing`	-	-	0.1, 0.2
`dropout_rate`	-	-	0.1

Table 4: Hyperparameters of SHAMPOO, CASPR and the partially coupled preconditioner in Section 4.1.

Shampoo parameter	value
`block_size`	1024
`graft_epsilon`	1e-8
`matrix_epsilon`	1e-6
`preconditioning_compute_steps`	20
`start_preconditioning_step`	101
`exponent_override`	0
`nesterov`	True
`relative_epsilon`	True
`inverse_type`	coupled newton (Anil et al., 2020)
`error_tolerance`	0.1
`clip_by_scaled_gradient_norm`	False