KL-Shampoo: Understanding and Improving Shampoo

Wu Lin, Scott C. Lowe, Felix Dangel,
Runa Eschenhagen, Zikun Xu, Roger B. Grosse

UNIVERSITY OF TORONTO · VECTOR INSTITUTE · UNIVERSITY OF CAMBRIDGE · Microsoft

We present a **principled preconditioner estimation scheme** via **Kullback–Leibler covariance minimization** to make Shampoo converge faster and more memory efficient.

arXiv

Idealized KL-Shampoo: KL-based Estimation Scheme for Shampoo's Preconditioner

Shampoo, winner of the *AlgoPerf: Training Algorithms* benchmark, outperforms Adam. SOAP is an efficient variant using Adam and QR.

Limitations of Original Shampoo and SOAP:

- Shampoo: step-size grafting for stabilization (heuristic, high memory)
- Shampoo: employment of eigendecomposition (slow runtime)
- SOAP: extra Adam's second moment (heuristic, high memory)

KL Minimization for Covariance Estimation

$$\min_{\mathbf{S}} \underbrace{\text{LogDet}\left(\mathbb{E}[\mathbf{g}\mathbf{g}^\top], \mathbf{S}\right)}_{\text{Symmetric Positive-Definite (SPD) Matrix}} \equiv \underbrace{\text{KL}\left(\mathcal{N}(\mathbf{0}, \mathbb{E}[\mathbf{g}\mathbf{g}^\top]), \mathcal{N}(\mathbf{0}, \mathbf{S})\right)}_{\text{Zero-mean Gaussian}}$$

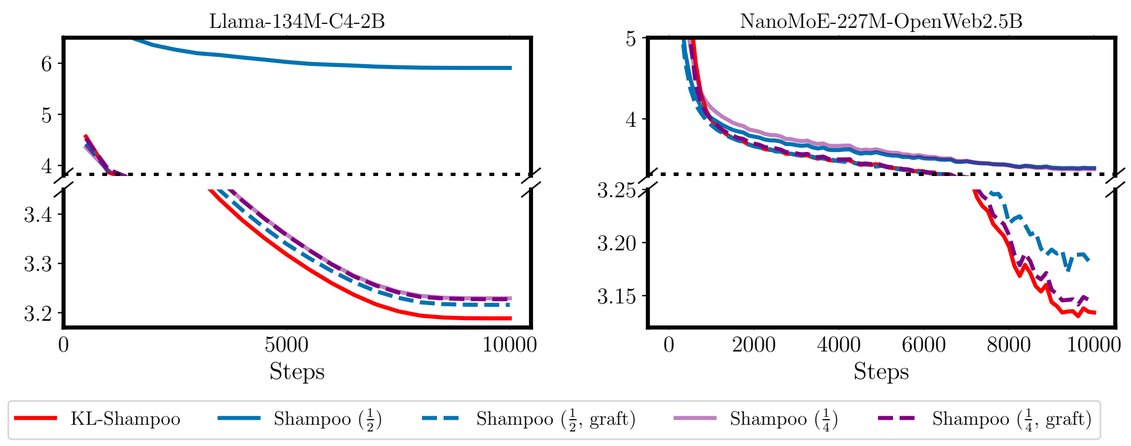
Goal: A principled way to approximate flattened gradient 2nd moment $\mathbb{E}[\mathbf{g}\mathbf{g}^\top]$ with a Kronecker-factored $\mathbf{S} = \mathbf{S}_1 \otimes \mathbf{S}_2$ via KL minimization for Covariances. (unflatten: $\mathbf{G} := \text{mat}(\mathbf{g})$)

$$\text{Original Shampoo: } \begin{cases} \mathbf{S}_1 = \mathbb{E}[\mathbf{G}\mathbf{G}^\top] \\ \mathbf{S}_2 = \mathbb{E}[\mathbf{G}^\top\mathbf{G}] \end{cases}$$

vs.

$$\text{KL-Shampoo: } \begin{cases} \mathbf{S}_1 = \frac{1}{d_2}\mathbb{E}[\mathbf{G}\,\mathbf{S}_2^{-1}\,\mathbf{G}^\top] \\ \mathbf{S}_2 = \frac{1}{d_1}\mathbb{E}[\mathbf{G}^\top\mathbf{S}_1^{-1}\,\mathbf{G}] \end{cases}$$

No Need for Step-size Grafting



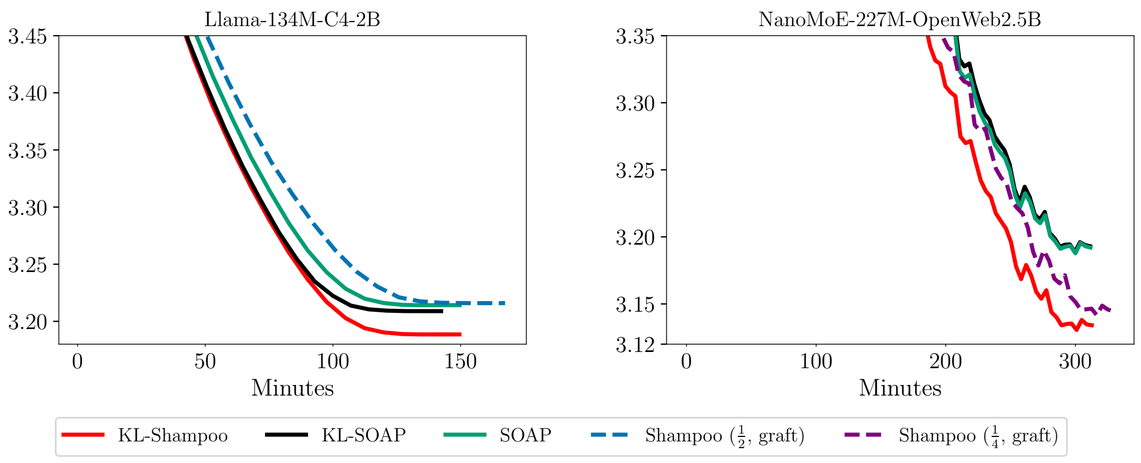
KL-Shampoo, using faster QR decomposition and without grafting, outperforms Shampoo with eigendecomposition and grafting.

Converge Faster and Use Less Memory

KL-Shampoo (no grafting): $\boldsymbol{\lambda}_1 \otimes \boldsymbol{\lambda}_2$ (eigenvalues, memory: $d_1 + d_2$)

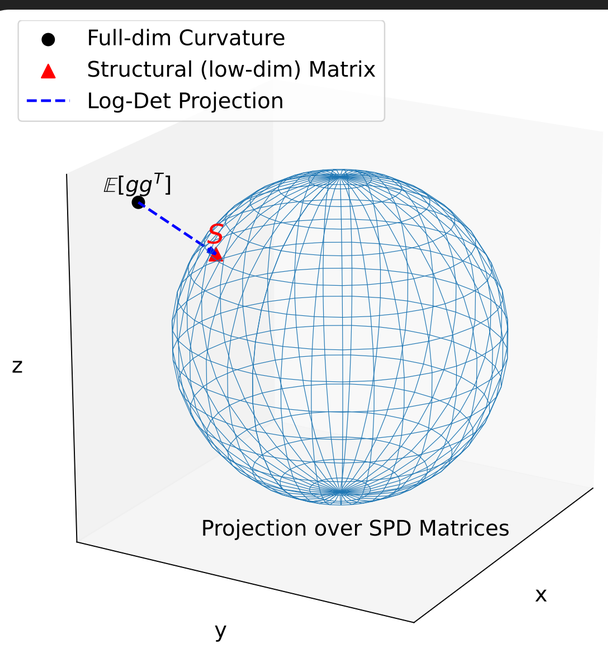
vs.

SOAP: $\mathbf{d}$ (Adam's 2nd moment as eigenvalues, memory: $d_1 d_2$)



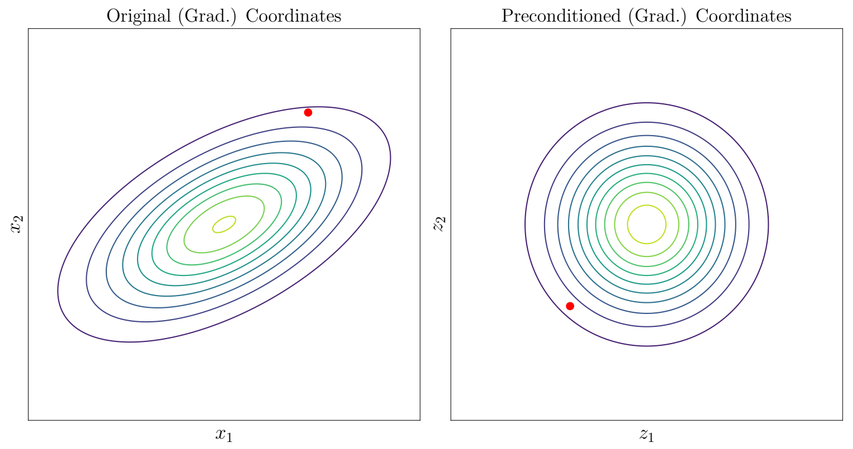
KL-Shampoo outperforms SOAP, Shampoo with grafting, and even KL-SOAP, while matching SOAP's per-iteration runtime and using less memory.

Understanding KL-Shampoo



View 1: Structural RMSProp via divergence-based projection over SPD matrices (No Gaussian assumption)

View 2: Maximum-likelihood scheme for zero-mean (matrix) Gaussian whitening (Covariance $\mathbf{S} = \mathbf{S}_1 \otimes \mathbf{S}_2$)



$$\text{KL: } \underbrace{\frac{1}{N}\sum_i^N \log \mathcal{N}(\mathbf{g}_i; 0, \mathbf{S})}_{\text{max log-likelihood}} \iff \underbrace{\text{KL}\left(\frac{1}{N}\sum_i^N \mathbf{g}_i\mathbf{g}_i^T, \mathbf{S}\right)}_{\text{min KL}}$$

$$\text{Orig: } \underbrace{\mathbb{E}[\mathbf{G}\mathbf{G}^T] = \frac{\mathbf{S}_1}{\text{Tr}(\mathbf{S}_2)}}_{\text{Row-wise moment matching}}, \quad \underbrace{\mathbb{E}[\mathbf{G}^T\mathbf{G}] = \frac{\mathbf{S}_2}{\text{Tr}(\mathbf{S}_1)}}_{\text{Column-wise moment matching}}$$