Novelty Assessment Report

Paper: Understanding and improving Shampoo and SOAP via Kullback-Leibler Minimization
PDF URL: https://openreview.net/pdf?id=pQQuC1nIQq
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Shampoo and its efficient, Adam-stabilized variant SOAP, employ structured second-moment estimation and have received growing attention for their effectiveness. In practice, Shampoo requires step-size grafting with Adam to achieve competitive performance. SOAP mitigates this by applying Adam in Shampoo's eigenbasis and further reducing per-iteration runtime. However, reliance on Adam introduces additional memory overhead in both methods. Prior theoretical interpretations have primarily examined their estimation schemes using the Frobenius norm. Motivated by the natural correspondence between the second moment and a covariance matrix, we reinterpret the estimation procedures in Shampoo and SOAP as instances of covariance estimation through the lens of Kullback–Leibler (KL) divergence minimization. This perspective reveals a previously overlooked theoretical limitation and motivates principled improvements to their design. Building on the KL perspective, we propose practical estimation schemes---KL-Shampoo and KL-SOAP---that match or exceed the performance of Shampoo and SOAP for pre-training a range of neural network models while maintaining SOAP-level per-iteration runtime. Notably, KL-Shampoo does not rely on Adam to achieve superior performance, thereby avoiding the associated memory overhead. Surprisingly, KL-Shampoo consistently outperforms the other methods in our experiments.

Core Task Landscape

This paper addresses: **Structured Second-Moment Estimation for Neural Network Optimization**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Structured Preconditioner Design and Approximation**
- **Adaptive Moment Methods and Gradient Statistics**
- **Covariance and Correlation Structure Learning**
- **Theoretical Analysis and Optimization Dynamics**
- **Domain-Specific Applications of Second-Order Methods**

Complete Taxonomy Tree

- Structured Second-Moment Estimation for Neural Network Optimization Survey Taxonomy
- Structured Preconditioner Design and Approximation
  - Kronecker-Factored and Block-Diagonal Approximations ★ (4 papers)
  - [0] Understanding and improving Shampoo and SOAP via Kullback-Leibler Minimization (Anon et al., 2026) View paper
  - [2] Scalable second order optimization for deep learning (Anil, 2020) View paper
  - [5] Tensor normal training for deep learning models (Yi Ren, 2021) View paper
  - [10] A kronecker-factored approximate fisher matrix for convolution layers (Grosse, 2016) View paper
  - Low-Rank and Eigenspace Methods (3 papers)
  - [29] Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient (Mishkin, 2018) View paper
  - [36] Gradient descent on neurons and its link to approximate second-order optimization (Benzing, 2022) View paper
  - [39] PLUMAGE: Probabilistic Low rank Unbiased Min Variance Gradient Estimator for Efficient Large Model Training (Haroush, 2025) View paper
  - Diagonal and Structured Diagonal Preconditioners (3 papers)
  - [9] Improving Adaptive Moment Optimization via Preconditioner Diagonalization (Nguyen Son, 2025) View paper
  - [18] Can We Remove the Square-Root in Adaptive Gradient Methods? A Second-Order Perspective (Lin Wu, 2024) View paper
  - [48] Doubly adaptive scaled algorithm for machine learning using second-order information (Jahani, 2021) View paper
- Adaptive Moment Methods and Gradient Statistics
  - Exponential Moving Average-Based Optimizers (5 papers)
  - [28] Î»-FAdaMax: A novel fractional-order gradient descent method with decaying second moment for neural network training (Guangyao Chen, 2025) View paper
  - [30] Deep learning LSTM for predicting thermally induced geometric errors using rotary axes' powers as input parameters (Huy Vu Ngoc, 2022) View paper
  - [32] Parameter Tuning Using Adaptive Moment Estimation in Deep Learning Neural Networks (E. Okewu, 2020) View paper
  - [35] Stock Prediction Based on Adaptive Gradient Descent Deep Learning (Bo Li, 2021) View paper
  - [37] Explainable Multi-Module Semantic Guided Attention Network for Accurate Medical Image Segmentation (R. Inbaraj, 2025) View paper
  - Hessian-Based Curvature Estimation (2 papers)
  - [4] Adahessian: An adaptive second order optimizer for machine learning (Gholami, 2021) View paper

    - [42] Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians (Vardan Papyan, 2019) View paper
    - Bayesian and Posterior Approximation Methods (3 papers)
    - [7] Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods? (Meinert, 2024) View paper
    - [26] Bayesian Deep Learning Detection of Anomalies and Failure: Application To Medical Images (Giuseppina Carannante, 2023) View paper
    - [34] L2M: Practical posterior Laplace approximation with optimization-driven second moment estimation (Perone, 2021) View paper
- Covariance and Correlation Structure Learning
    - Neural Network-Based Covariance Estimation (6 papers)
    - [1] Self-Supervised Learning for Covariance Estimation (Tzvi Diskin, 2024) View paper
    - [3] Constraining the Hubble constant with a simulated full covariance matrix using neural networks (Niu Jing, 2026) View paper
    - [20] Structured Covariance Gaussian Networks for Orion Crew Module Aerodynamic Uncertainty Quantification (Tenavi Nakamura-Zimmerer, 2023) View paper
    - [21] Structured uncertainty prediction networks (Garoe Dorta, 2018) View paper
    - [44] Neural Networks for Parameter Estimation in Geometrically Anisotropic Geostatistical Models (Alegria Alfredo, 2024) View paper
    - [46] Fast covariance parameter estimation of spatial Gaussian process models using neural networks (Florian Gerber, 2020) View paper
    - Graph Neural Networks with Covariance Structures (5 papers)
    - [23] coVariance Neural Networks (Sihag, 2022) View paper
    - [24] CCP-GNN: Competitive Covariance Pooling for Improving Graph Neural Networks (Pengfei Zhu, 2024) View paper
    - [43] Fair CoVariance Neural Networks (Andrea Cavallo, 2025) View paper
    - [49] Spatiotemporal Covariance Neural Networks (Cavallo, 2024) View paper
    - [50] Transferability of coVariance Neural Networks (Saurabh Sihag, 2024) View paper
    - Covariance Structure in Model Weights and Features (2 papers)
    - [8] Learning and inference with correlated neural variability (Yang Qi, 2025) View paper
    - [13] Understanding the covariance structure of convolutional filters (Trockman, 2022) View paper
    - High-Dimensional and Statistical Covariance Estimation (1 papers)
    - [19] Highâ□□Dimensional Covariance Estimation From a Small Number of Samples (Matthias Morzfeld, 2024) View paper
- Theoretical Analysis and Optimization Dynamics
    - Generalization and Information-Theoretic Perspectives (3 papers)
    - [6] A Second-Order Perspective on Model Compositionality and Incremental Learning (Porrello, 2024) View paper
    - [11] Gradient-based feature learning under structured data (Mousavi-Hosseini, 2023) View paper
    - [15] Whitening and second order optimization both make information in the dataset unusable during training, and can reduce or prevent generalization (Wadia, 2021) View paper
    - Convergence and Optimization Landscape Analysis (2 papers)
    - [33] An empirical study of stochastic gradient descent with structured covariance noise (Yeâ□□Ming Wen, 2020) View paper
    - [41] An empirical study of large-batch stochastic gradient descent with structured covariance noise (Wen Ye-ming, 2019) View paper
    - Survey and Meta-Analysis of Second-Order Methods (3 papers)
    - [17] Review of second-order optimization techniques in artificial neural networks backpropagation (H. Tan, 2019) View paper
    - [45] A survey of deep learning optimizers--first and second order methods (Kashyap, 2022) View paper
    - [47] Second Order Neural Network Optimization: Meta Analysis (Jeshwanth Challagundla, 2024) View paper
- Domain-Specific Applications of Second-Order Methods
    - Bioinformatics and Computational Biology (2 papers)
    - [16] ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks (Yang Li, 2019) View paper
    - [22] Mixed Precision Fermi-Operator Expansion on Tensor Cores from a Machine Learning Perspective. (Joshua Finkelstein, 2021) View paper
    - Computer Vision and Image Processing (2 papers)
    - [27] Learning Second Order Local Anomaly for General Face Forgery Detection (Jianwei Fei, 2022) View paper
    - [31] VAEs with structured image covariance applied to compressed sensing MRI (Margaret Duff, 2023) View paper
    - Signal Processing and Time-Series Analysis (2 papers)
    - [12] Multi-classification for EEG motor imagery signals using data evaluation-based auto-selected regularized FBCSP and convolutional neural network (Yang An, 2023) View paper
    - [14] Physics-Informed Deep Learning for Traffic State Estimation: A Hybrid Paradigm Informed By Second-Order Traffic Models (Di, 2021) View paper
    - Specialized Computational Tasks (3 papers)
    - [25] Smooth Model Compression without Fine-Tuning (Runkel, 2025) View paper
    - [38] Efficient federated graph aggregation for privacy-preserving GNN-based session recommendation (Jing Lou, 2025) View paper
    - [40] Deep learning-based moment closure for multi-phase computation of semiclassical limit of the Schr\" odinger equation (Jin Woo Jang, 2025) View paper

Narrative

Core task: structured second-moment estimation for neural network optimization. The field organizes around several complementary branches that address how to efficiently capture and exploit curvature information during training. Structured Preconditioner Design and Approximation focuses on computationally tractable approximations to the full Hessian or Fisher information matrix, often using Kronecker factorizations or block-diagonal structures to reduce memory and computation while preserving useful geometric information. Adaptive Moment Methods and Gradient Statistics encompasses first- and second-moment estimators that adapt learning rates based on gradient history, bridging classical stochastic methods with modern variance-reduction techniques. Covariance and Correlation Structure Learning examines how to model dependencies among parameters or activations, sometimes drawing on statistical estimation of high-dimensional covariance matrices. Theoretical Analysis and Optimization Dynamics investigates convergence guarantees, curvature properties, and the interplay between batch size and noise structure. Domain-Specific Applications of Second-Order Methods tailors these ideas to specialized settings such as computer vision, natural language processing, or scientific computing, where problem structure can be further exploited.

A particularly active line of work revolves around Kronecker-factored and block-diagonal approximations, which balance scalability with the benefits of second-order information. Scalable Second Order[2] and Tensor Normal Training[5] exemplify efforts to decompose large curvature matrices into manageable factors, while Kronecker Fisher Matrix[10] laid foundational ideas for factorizing the Fisher information. Shampoo SOAP KL[0] sits squarely within this branch, proposing a structured preconditioner that leverages Kronecker products and block structures to achieve efficient updates. Compared to Tensor Normal Training[5], which emphasizes tensor-based reparameterizations, Shampoo SOAP KL[0] focuses more directly on preconditioning via second-moment approximations. Meanwhile, works like Hubble Covariance Networks[3] and Self-Supervised Covariance[1] explore covariance structure in different contexts, highlighting ongoing questions about how best to estimate and regularize second-moment information across diverse architectures and training regimes.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Scalable second order optimization for deep learning

Authors: Anil, Rohan, Gupta, Vineet, Koren, et al. (10 authors total) | **Year/Venue**: 2020 | **URL**: View paper

Abstract

Optimization in machine learning, both theoretical and applied, is presently dominated by first-order gradient methods such as stochastic gradient descent. Second-order optimization methods, that involve second derivatives and/or second order statistics of the data, are far less prevalent despite strong theoretical properties, due to their prohibitive computation, memory and communication costs. In an attempt to bridge this gap between theoretical and practical optimization, we present a scalabl...

Relationship Analysis

Both papers belong to the Kronecker-Factored and Block-Diagonal Approximations category, focusing on structured second-moment estimation using Kronecker products to reduce computational costs in neural network optimization. The original paper reinterprets Shampoo and SOAP through KL divergence minimization and proposes improved estimation schemes (KL-Shampoo and KL-SOAP) that avoid Adam's memory overhead while maintaining competitive performance. The candidate paper presents a scalable implementation of full-matrix Adagrad with Kronecker-factored approximations, emphasizing practical deployment on heterogeneous hardware architectures and demonstrating wall-clock time improvements on large-scale tasks, but does not explore the KL divergence perspective or the specific improvements to Shampoo/SOAP estimation rules.

---

2. Tensor normal training for deep learning models

Authors: Yi Ren, Donald Goldfarb, D. Goldfarb | **Year/Venue**: 2021 | **URL**: View paper

Abstract

Despite the predominant use of first-order methods for training deep learning models, second-order methods, and in particular, natural gradient methods, remain of interest because of their potential for accelerating training through the use of curvature information. Several methods with non-diagonal preconditioning matrices, including KFAC, Shampoo, and K-BFGS, have been proposed and shown to be effective. Based on the so-called tensor normal (TN) distribution, we propose and analyze a brand new...

Relationship Analysis

Both papers belong to the Kronecker-Factored and Block-Diagonal Approximations category, using Kronecker product structures to approximate second-moment matrices for efficient neural network optimization. They overlap in employing structured preconditioners based on Kronecker factorizations of gradient statistics, but differ fundamentally in their theoretical motivation and estimation approach: the original paper reinterprets Shampoo/SOAP through KL divergence minimization for covariance estimation and proposes improved KL-based schemes, while the candidate paper (TNT) assumes gradients follow a tensor-normal distribution and derives preconditioners as maximum likelihood estimators of the Fisher matrix under this distributional assumption.

---

3. A kronecker-factored approximate fisher matrix for convolution layers

Authors: Grosse, Roger, Martens, James, R. Grosse, et al. (6 authors total) | **Year/Venue**: 2016 | **URL**: View paper

Abstract

Second-order optimization methods such as natural gradient descent have the potential to speed up training of neural networks by correcting for the curvature of the loss function. Unfortunately, the exact natural gradient is impractical to compute for large models, and most approximations either require an expensive iterative procedure or make crude approximations to the curvature. We present Kronecker Factors for Convolution (KFC), a tractable approximation to the Fisher matrix for convolutiona...

Relationship Analysis

Both papers belong to the Kronecker-Factored and Block-Diagonal Approximations category, using Kronecker product structures to approximate second-moment matrices for efficient neural network optimization. The original paper reinterprets Shampoo and SOAP through KL divergence minimization and proposes improved estimation schemes (KL-Shampoo and KL-SOAP) that avoid Adam's memory overhead while maintaining competitive performance. The candidate paper introduces KFC (Kronecker Factors for Convolution), which extends Kronecker-factored approximations specifically to convolutional layers by modeling spatial homogeneity and independence assumptions, whereas the original paper focuses on improving general Shampoo/SOAP estimation rules across various architectures including transformers and recurrent models.

Contributions Analysis

Overall novelty summary. The paper proposes KL-Shampoo and KL-SOAP, reinterpreting Shampoo and SOAP's second-moment estimation through Kullback–Leibler divergence minimization rather than Frobenius norm. It resides in the Kronecker-Factored and Block-Diagonal Approximations leaf, which contains four papers total including this work. This leaf sits within the broader Structured Preconditioner Design and Approximation branch, indicating a moderately populated research direction focused on computationally tractable curvature approximations. The sibling papers address related Kronecker factorizations and block-diagonal structures, suggesting the paper enters an active but not overcrowded subfield.

The taxonomy reveals neighboring leaves addressing Low-Rank and Eigenspace Methods and Diagonal and Structured Diagonal Preconditioners, both offering alternative approximation strategies. The Adaptive Moment Methods branch, particularly Exponential Moving Average-Based Optimizers, provides context for Adam-based techniques that Shampoo and SOAP incorporate. The paper's KL divergence lens bridges structured preconditioning with covariance estimation principles found in the Covariance and Correlation Structure Learning branch, though it remains firmly within optimization rather than statistical modeling. This positioning suggests the work synthesizes ideas across multiple taxonomy branches while maintaining focus on preconditioner design.

Among 26 candidates examined across three contributions, none clearly refute the proposed methods. The KL divergence perspective examined 10 candidates with zero refutable overlaps, suggesting this theoretical lens is relatively unexplored in prior Shampoo literature. The KL-Shampoo and KL-SOAP methods similarly faced 10 candidates without clear prior instantiation. The memory-efficient variant without Adam grafting examined 6 candidates, also without refutation. These statistics indicate that within the limited search

scope, the specific combination of KL-based estimation and memory-efficient design appears novel, though the search does not cover the entire optimization literature.

The analysis suggests the paper introduces a fresh theoretical perspective and practical variants within an established research direction. The limited search scope means we cannot rule out related work in broader optimization or information geometry communities. The taxonomy placement and sibling papers indicate the work builds on well-known Shampoo foundations while proposing a distinct estimation principle. The absence of refuting candidates among 26 examined supports novelty claims, though exhaustive verification would require deeper literature coverage beyond top-K semantic matches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: KL divergence perspective for Shampoo and SOAP estimation

Description: The authors introduce a novel theoretical framework that reinterprets the second-moment estimation schemes in Shampoo and SOAP optimizers as covariance estimation problems solved via KL divergence minimization. This perspective reveals a previously overlooked theoretical limitation in these methods and provides a principled foundation for improvements.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Comparing KL Divergence and MSE for Covariance Estimation in Target Detection

URL: View paper

Brief Assessment

KL Divergence MSE[60] focuses on covariance estimation for target detection applications, not on adaptive optimizers like Shampoo and SOAP. The candidate paper's abstract mentions 'covariance estimation' and 'mean squared error (mse) metric' in a target detection context, which is fundamentally different from the original paper's focus on second-moment estimation in neural network optimization.

---

2. Dpo kernels: A semantically-aware, kernel-enhanced, and divergence-rich paradigm for direct preference optimization

URL: View paper

Brief Assessment

DPO Kernels[58] focuses on direct preference optimization for LLM alignment using kernel methods and divergence measures, not on adaptive optimizer design or covariance estimation in second-moment methods like Shampoo/SOAP.

---

3. On the normalized signal to noise ratio in covariance estimation

URL: View paper

Brief Assessment

Normalized Signal Noise[64] applies KL divergence to adaptive detection in signal processing (noise covariance estimation for target detection), not to second-moment estimation in neural network optimizers like Shampoo and SOAP.

---

4. Improving Mean Covariance Matrix Estimation by Minimizing Within-class Dissimilarities Using Asymmetry of Kullback-Leibler Divergence in MI-Based BCI

URL: View paper

Brief Assessment

Covariance KL Divergence[57] applies KL divergence to Brain-Computer Interface covariance matrix estimation for classification tasks, not to adaptive optimizer preconditioner design in deep learning.

---

5. A geometric unification of distributionally robust covariance estimators: Shrinking the spectrum by inflating the ambiguity set

URL: View paper

Brief Assessment

Geometric Covariance Estimators[59] focuses on distributionally robust covariance estimation for high-dimensional statistics using KL divergence, not on adaptive optimizer design or second-moment estimation in neural network training contexts like Shampoo/SOAP.

---

6. Estimation of clutter covariance matrix in stap based on knowledge-aided and geometric methods

URL: View paper

Brief Assessment

Clutter Covariance STAP[63] applies KL divergence to radar signal processing (STAP clutter covariance estimation), not to adaptive optimizer design for neural network training. The technical domains and applications are entirely distinct.

---

7. On the Minimum -Divergence Estimator

URL: View paper

Brief Assessment

Minimum Divergence Estimator[65] focuses on general statistical estimation via KL divergence minimization for covariance matrices, but does not address adaptive optimizers, second-moment estimation in neural network training, or the specific Shampoo/SOAP frameworks that are central to the original paper's contribution.

---

8. Covariance alignment: from maximum likelihood estimation to Gromovâ□□Wasserstein

URL: View paper

Brief Assessment

Covariance Alignment Gromov[61] focuses on covariance alignment using Gromov-Wasserstein methods, not on reinterpreting adaptive optimizer estimation schemes through KL divergence minimization. The candidate's mention of KL divergence relates to stationary processes and permutation estimation, which is a different context from optimizer preconditioner design.

---

9. Robust Gaussian Mixture Modeling: A -Divergence Based Approach

URL: View paper

Brief Assessment

Robust Gaussian Mixture[62] applies KL divergence to Gaussian mixture modeling problems, not to adaptive optimizer design or second-moment estimation in neural network training. The technical contexts are fundamentally different.

---

10. Differentially Private Distribution Release of Gaussian Mixture Models via KL-Divergence Minimization

URL: View paper

Brief Assessment

Private Gaussian Mixtures[66] applies KL divergence to privacy-preserving GMM parameter release, not to adaptive optimizer design. The candidate focuses on differential privacy mechanisms for statistical models, while the original work reinterprets second-moment estimation in neural network optimizers through KL minimization.

Contribution 2: KL-Shampoo and KL-SOAP optimization methods

Description: The authors develop two new optimization methods, KL-Shampoo and KL-SOAP, that implement improved estimation schemes based on their KL perspective. These methods achieve competitive or superior performance compared to existing Shampoo and SOAP optimizers while maintaining efficient per-iteration runtime.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Towards fast, specialized machine learning force fields: Distilling foundation models via energy hessians

URL: View paper

Brief Assessment

Energy Hessians Distillation[73] focuses on knowledge distillation for machine learning force fields in computational chemistry, not on second-order optimization methods for neural network training. The paper addresses a completely different domain and problem.

---

2. Recursion Newton-Like Algorithm for l2,0-ReLU Deep Neural Networks

URL: View paper

Brief Assessment

Recursion Newton-Like[74] focuses on training and compressing ReLU-DNNs with l2,0 regularization using Newton-like methods for sparse optimization, not on second-order optimization methods for general neural network pretraining like KL-Shampoo and KL-SOAP.

---

3. The Potential of Second-Order Optimization for LLMs: A Study with Full Gauss-Newton

URL: View paper

Brief Assessment

Gauss-Newton LLMs[70] focuses on full Gauss-Newton preconditioning and layerwise approximations for LLM training, not on KL-divergence-based improvements to Shampoo/SOAP's structured second-moment estimation schemes.

---

4. Sophia: A scalable stochastic second-order optimizer for language model pre-training

URL: View paper

Brief Assessment

Sophia[67] focuses on diagonal Hessian estimation with element-wise clipping for language model pre-training, while the original paper develops KL-divergence-based Kronecker-factored second-moment estimation methods. These are fundamentally different second-order optimization approaches with distinct mathematical frameworks and preconditioner structures.

---

5. Nonlinear discretizations and Newton's method: characterizing stationary points of regression objectives

URL: View paper

Brief Assessment

Newton Stationary Points[76] focuses on exact second-order methods and their failure modes in neural network training, not on structured second-moment estimation or Shampoo-based optimizers. The candidate examines why exact Hessian information fails, while the original develops improved Kronecker-factored preconditioners.

---

6. Practical Efficiency of Muon for Pretraining

URL: View paper

Brief Assessment

Muon Efficiency[71] focuses on demonstrating the practical efficiency of the Muon optimizer (a second-order method) compared to AdamW, without proposing new optimization methods or estimation schemes based on KL divergence minimization for Shampoo/SOAP.

---

7. 4-bit shampoo for memory-efficient network training

URL: View paper

Brief Assessment

4-bit Shampoo[72] focuses on memory-efficient training through 4-bit quantization of optimizer states, specifically quantizing eigenvector matrices of preconditioners. This is a different technical approach from developing new optimization methods based on KL divergence minimization for covariance estimation.

---

8. When Does Second-Order Optimization Speed Up Training?

URL: View paper

Brief Assessment

Second-Order Speed Up[69] focuses on empirically identifying conditions (batch size, dataset size) under which existing second-order methods like K-FAC and Shampoo outperform first-order methods. It does not propose new optimization methods or estimation schemes based on KL divergence minimization.

---

9. Unconstrained optimization in neural network training

URL: View paper

Brief Assessment

Unconstrained Neural Optimization[68] discusses Newton-type methods in general optimization contexts, but does not address structured second-order methods like Shampoo/SOAP or their KL-divergence-based improvements for neural network training.

---

10. Understanding data influence with differential approximation

URL: View paper

Brief Assessment

Data Influence Differential[75] focuses on data influence estimation for neural network training through second-order approximations of sample-wise influence across training iterations. This is fundamentally different from developing optimization methods like KL-Shampoo

and KL-SOAP, which address preconditioned gradient descent through Kullback-Leibler divergence minimization for second-moment estimation.

Contribution 3: Memory-efficient KL-Shampoo without Adam grafting

Description: The authors demonstrate that their KL-Shampoo method eliminates the need for step-size grafting with Adam, which is required by standard Shampoo for competitive performance. This design choice reduces memory overhead while maintaining or improving optimization performance.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Preconditioned Gradient Descent Algorithm for Inverse Filtering on Spatially Distributed Networks

URL: View paper

Brief Assessment

Preconditioned Gradient Networks[56] addresses inverse filtering on spatially distributed networks using preconditioned gradient descent, which is fundamentally different from the original paper's focus on memory-efficient second-order optimization methods for neural network training that eliminate Adam grafting overhead.

2. NysAct: A Scalable Preconditioned Gradient Descent using NystrÃ¶m Approximation

URL: View paper

Brief Assessment

NysAct[52] focuses on Nyström approximation for activation covariance preconditioning in neural networks, not on eliminating Adam grafting in Shampoo-type methods. The candidate addresses a different technical approach (eigenvalue-shifted Nyström) rather than the KL-divergence-based estimation that enables removing Adam grafting.

3. Extreme Tensoring for Low-Memory Preconditioning

URL: View paper

Brief Assessment

Extreme Tensoring[55] focuses on memory-efficient adaptive preconditioning through extreme tensor factorization for arbitrary models, not on eliminating Adam grafting specifically. The candidate does not discuss step-size grafting or Adam integration as a design consideration.

4. Pipelined Preconditioned s-step Conjugate Gradient Methods for Distributed Memory Systems

URL: View paper

Brief Assessment

Pipelined Conjugate Gradient[54] focuses on pipelined variants of the conjugate gradient method for solving linear systems in distributed memory systems, not on second-order optimization methods for neural network training or memory-efficient preconditioned gradient methods that eliminate Adam grafting.

5. Dual space preconditioning for gradient descent

URL: View paper

Brief Assessment

Dual Space Preconditioning[51] focuses on convex optimization using Bregman gradient methods with dual space preconditioning, not on neural network training or second-moment estimation schemes like KL-Shampoo.

6. A Scalable and Flexible Framework for Gaussian Processes via Matrix-Vector Multiplication

URL: View paper

Brief Assessment

Gaussian Processes Framework[53] focuses on matrix-vector multiplication methods for Gaussian processes, not on preconditioned gradient optimization methods or memory-efficient alternatives to Adam grafting in neural network training.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Understanding and improving Shampoo and SOAP via Kullback-Leibler Minimization View paper
- [1] Self-Supervised Learning for Covariance Estimation View paper
- [2] Scalable second order optimization for deep learning View paper
- [3] Constraining the Hubble constant with a simulated full covariance matrix using neural networks View paper
- [4] Adahessian: An adaptive second order optimizer for machine learning View paper
- [5] Tensor normal training for deep learning models View paper
- [6] A Second-Order Perspective on Model Compositionality and Incremental Learning View paper
- [7] Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods? View paper
- [8] Learning and inference with correlated neural variability View paper
- [9] Improving Adaptive Moment Optimization via Preconditioner Diagonalization View paper
- [10] A kronecker-factored approximate fisher matrix for convolution layers View paper
- [11] Gradient-based feature learning under structured data View paper
- [12] Multi-classification for EEG motor imagery signals using data evaluation-based auto-selected regularized FBCSP and convolutional neural network View paper
- [13] Understanding the covariance structure of convolutional filters View paper
- [14] Physics-Informed Deep Learning for Traffic State Estimation: A Hybrid Paradigm Informed By Second-Order Traffic Models View paper
- [15] Whitening and second order optimization both make information in the dataset unusable during training, and can reduce or prevent generalization View paper
- [16] ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks View paper
- [17] Review of second-order optimization techniques in artificial neural networks backpropagation View paper

- [18] Can We Remove the Square-Root in Adaptive Gradient Methods? A Second-Order Perspective View paper
- [19] Highâ□□Dimensional Covariance Estimation From a Small Number of Samples View paper
- [20] Structured Covariance Gaussian Networks for Orion Crew Module Aerodynamic Uncertainty Quantification View paper
- [21] Structured uncertainty prediction networks View paper
- [22] Mixed Precision Fermi-Operator Expansion on Tensor Cores from a Machine Learning Perspective. View paper
- [23] coVariance Neural Networks View paper
- [24] CCP-GNN: Competitive Covariance Pooling for Improving Graph Neural Networks View paper
- [25] Smooth Model Compression without Fine-Tuning View paper
- [26] Bayesian Deep Learning Detection of Anomalies and Failure: Application To Medical Images View paper
- [27] Learning Second Order Local Anomaly for General Face Forgery Detection View paper
- [28] Î»-FAdaMax: A novel fractional-order gradient descent method with decaying second moment for neural network training View paper
- [29] Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient View paper
- [30] Deep learning LSTM for predicting thermally induced geometric errors using rotary axes' powers as input parameters View paper
- [31] VAEs with structured image covariance applied to compressed sensing MRI View paper
- [32] Parameter Tuning Using Adaptive Moment Estimation in Deep Learning Neural Networks View paper
- [33] An empirical study of stochastic gradient descent with structured covariance noise View paper
- [34] L2M: Practical posterior Laplace approximation with optimization-driven second moment estimation View paper
- [35] Stock Prediction Based on Adaptive Gradient Descent Deep Learning View paper
- [36] Gradient descent on neurons and its link to approximate second-order optimization View paper
- [37] Explainable Multi-Module Semantic Guided Attention Network for Accurate Medical Image Segmentation View paper
- [38] Efficient federated graph aggregation for privacy-preserving GNN-based session recommendation View paper
- [39] PLUMAGE: Probabilistic Low rank Unbiased Min Variance Gradient Estimator for Efficient Large Model Training View paper
- [40] Deep learning-based moment closure for multi-phase computation of semiclassical limit of the Schr\" odinger equation View paper
- [41] An empirical study of large-batch stochastic gradient descent with structured covariance noise View paper
- [42] Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians View paper
- [43] Fair CoVariance Neural Networks View paper
- [44] Neural Networks for Parameter Estimation in Geometrically Anisotropic Geostatistical Models View paper
- [45] A survey of deep learning optimizers--first and second order methods View paper
- [46] Fast covariance parameter estimation of spatial Gaussian process models using neural networks View paper
- [47] Second Order Neural Network Optimization: Meta Analysis View paper
- [48] Doubly adaptive scaled algorithm for machine learning using second-order information View paper
- [49] Spatiotemporal Covariance Neural Networks View paper
- [50] Transferability of coVariance Neural Networks View paper
- [51] Dual space preconditioning for gradient descent View paper
- [52] NysAct: A Scalable Preconditioned Gradient Descent using NystrÃ¶m Approximation View paper
- [53] A Scalable and Flexible Framework for Gaussian Processes via Matrix-Vector Multiplication View paper
- [54] Pipelined Preconditioned s-step Conjugate Gradient Methods for Distributed Memory Systems View paper
- [55] Extreme Tensoring for Low-Memory Preconditioning View paper
- [56] Preconditioned Gradient Descent Algorithm for Inverse Filtering on Spatially Distributed Networks View paper
- [57] Improving Mean Covariance Matrix Estimation by Minimizing Within-class Dissimilarities Using Asymmetry of Kullback-Leibler Divergence in MI-Based BCI View paper
- [58] Dpo kernels: A semantically-aware, kernel-enhanced, and divergence-rich paradigm for direct preference optimization View paper
- [59] A geometric unification of distributionally robust covariance estimators: Shrinking the spectrum by inflating the ambiguity set View paper
- [60] Comparing KL Divergence and MSE for Covariance Estimation in Target Detection View paper
- [61] Covariance alignment: from maximum likelihood estimation to Gromovâ□□Wasserstein View paper
- [62] Robust Gaussian Mixture Modeling: A -Divergence Based Approach View paper
- [63] Estimation of clutter covariance matrix in stap based on knowledge-aided and geometric methods View paper
- [64] On the normalized signal to noise ratio in covariance estimation View paper
- [65] On the Minimum -Divergence Estimator View paper
- [66] Differentially Private Distribution Release of Gaussian Mixture Models via KL-Divergence Minimization View paper
- [67] Sophia: A scalable stochastic second-order optimizer for language model pre-training View paper
- [68] Unconstrained optimization in neural network training View paper
- [69] When Does Second-Order Optimization Speed Up Training? View paper
- [70] The Potential of Second-Order Optimization for LLMs: A Study with Full Gauss-Newton View paper
- [71] Practical Efficiency of Muon for Pretraining View paper
- [72] 4-bit shampoo for memory-efficient network training View paper
- [73] Towards fast, specialized machine learning force fields: Distilling foundation models via energy hessians View paper
- [74] Recursion Newton-Like Algorithm for l2,0-ReLU Deep Neural Networks View paper
- [75] Understanding data influence with differential approximation View paper
- [76] Nonlinear discretizations and Newton's method: characterizing stationary points of regression objectives View paper