[18.338 COURSE PROJECT] SPECTRAL PRECONDITIONING FOR OPTIMIZERS: AN RMT VIEW OF SHAMPOO/SOAP

HO TIN (ALEX) FAN

Abstract. Matrix-preconditioned optimizers such as Shampoo and SOAP are widely used in large-scale deep learning, yet their spectral stability under finite samples and stale preconditioning is poorly understood. We study the preconditioned curvature $H' = \hat{G}^{-\frac{1}{2}} H \hat{G}^{-\frac{1}{2}}$ through the lens of random matrix theory, where H is a curvature proxy and $\hat{G}$ is an EMA-based estimate of the gradient second moment. First, modeling $\hat{G}$ as a (whitened) Wishart matrix, we use the Marchenko–Pastur law to bound the spectrum and condition number of H' and derive explicit sample-size requirements in terms of the effective aspect ratio $\gamma = d/m$. Second, in a spiked model for H, we invoke the Baik-Ben Arous-Péché transition to show that strong curvature directions are either subcritical and behave like bulk noise, or supercritical and become detectable outliers that are actively whitened by $\hat{G}^{-\frac{1}{2}}$. Third, combining Davis-Kahan sin-Θ bounds with a Lipschitz estimate for $A \mapsto A^{-\frac{1}{2}}$, we show that using a stale eigenbasis (refreshing every τ steps) perturbs all eigenvalues of H' by at most $O\big(\|\hat{G}_{t+\tau} - \hat{G}_t\|_2\big)$. Simulations confirm that, in this RMT regime, SOAP/Shampoo are spectrally stable: conditioning is dominated by γ, while spikes and staleness are effectively tamed by the empirical preconditioner.

1. Introduction

In optimization problems, second-order structures (Hessian/Fisher/Gauss-Newton) often cause anisotropy that slows first-order methods. A class of matrix-preconditioned methods—Shampoo [1] and SOAP [2] —addresses this by maintaining a data-driven preconditioner that approximates the inverse square root of the gradient second moment (or Fisher/Gauss-Newton), thereby whitening gradient descent updates.

Let H denote a curvature proxy (Fisher/Gauss-Newton), G the estimator of the gradient second moment built from an EMA with effective sample size m, and $H' = G^{-\frac{1}{2}} H G^{-\frac{1}{2}}$ the preconditioned curvature. In practice, G is estimated from finite samples and its eigenbasis is recomputed only every τ updates—raising three questions:

1) How large must m be for G to yield an isotropic H' despite sampling noise?
2) When does preconditioning miss outliers? Relatedly, when can we expect spikes in H to be removable by whitening G, and is the whitening robust against random spikes?
3) How does a stale eigenbasis (every τ) inflate the spectral edges of H'?

While these optimizers have shown promising results in practice, there remain significant gaps in the theoretical understanding of their stability. This project analyzes the spectrum of H' through Random Matrix Theory. We model (H, G) as (spiked) sample covariances (Wishart-type) and apply the Marchenko-Pastur law (finite-sample bulk edges) and the Baik-Ben Arous-Péché transition (spike detectability), together with the Davis-Kahan $\sin\Theta$ theorem (eigenspace perturbations), to derive finite-sample and staleness-dependent bounds on the spectral edges and condition number of H'. These are standard finite-RMT questions about sample covariance spectra and eigenspace stability. We confirm the predicted conditioning and edge behavior with Python simulations.

2. Background / Related Work

2.1. Matrix-preconditioned Optimizers (Shampoo/SOAP).

Matrix-based preconditioning methods attempt to correct for anisotropy in the loss landscape by applying an approximate inverse curvature operator to each gradient update. In deep learning, the full curvature matrix is far too large to maintain, so practical preconditioners rely on structured factorizations and running estimates of gradient second moments.

Shampoo [1] maintains, for each parameter tensor, a set of per-mode second-moment matrices: if a parameter has shape $n_1 \times ... \times n_k$, Shampoo tracks k matrices of sizes $n_1 \times n_1, ..., n_k \times n_k$. These matrices are updated via exponential moving average and are designed to approximate the marginals

of the full gradient covariance. Applying inverse square roots to these matrices effectively performs a Kronecker-structured whitening step. The original work establishes convergence guarantees in convex stochastic settings and reports strong empirical acceleration.

SOAP [2] refines this interpretation by observing that Shampoo is equivalent to running Adam/ Adafactor in the eigenbasis of each per-mode preconditioner. This perspective highlights the central role of the eigenvectors of the empirical second-moment matrix. SOAP also introduces an eigenbasis refresh frequency τ, performing eigendecompositions only once every τ iterations to reduce computational overhead. Despite its practical success, this refresh mechanism necessarily introduces temporal drift between the true preconditioner G_t and the "stale" eigenbasis used to construct the update.

Several prior works study optimization dynamics under adaptive preconditioning, but few address the spectral reliability of the empirical curvature estimator. For instance, the Muon analysis of Sato et al. (2025) [3] provides convergence guarantees for a related optimizer but does not examine how sampling noise affects the spectrum of the preconditioner. Likewise, existing Shampoo/SOAP analyses focus on algorithmic properties rather than on the spectral stability of G_t under finite-sample effects. However, there is little to no theoretical analysis of the reliability of the empirical eigenbasis. Our work aims to fill this gap by providing RMT-based analysis of the spectral behavior.

2.2. **Connections to Random Matrix Theory.**

We are interested in the sample covariance/Wishart model. Namely, we sample curvature proxies and their estimators as (possibly spiked) sample covariances: $H \approx \Sigma^{\frac{1}{2}} X X^T \Sigma^{\frac{1}{2}}/n,\ G \approx \widehat{\Sigma_m}$. The **Marchenko-Pastur (MP) law** [4] gives the finite-sample bulk edges for eigenvalues when $d, m \to \infty$ with $d/m \to \gamma$. This underpins the bounds on the spectral spread of $H' = G^{-\frac{1}{2}} H G^{-\frac{1}{2}}$

Likewise, since the covariance might be spiked, we want to know when strong curvature directions are even observable to G. We leverage the **Baik-Ben-Peche (BBP) transition** to characterize when population spikes detach from the MP bulk [5]. This lets us know when the large spectral spikes are removable by whitening or when they stick to the bulk and survive as outliers.

Finally, since SOAP refreshes the eigenbasis only every τ steps, we need to understand how eigenspaces drift over time. The **Davis-Kahan sin Θ theorem** [6] gives bounds on the angle between eigenspaces of perturbed matrices. This lets us quantify "basis staleness" in SOAP — how drift in G over τ steps inflates the spectral edges of H' and tightens safe step-size.

3. Problem Statements and Main Arguments (Rough Draft)

3.1. **Model and Assumptions.**

Let the model itself be fixed over a short window (slow drift). Draw minibatches B_t of size b from the data distribution. Let the per-sample gradient be $g(x, y; w) \in R^d$ with mean 0 and second moment $\Sigma := E[gg^T]$ (Fisher/Gauss-Newton proxy). Furthermore, we have the following assumptions:

- (Elliptical/Wishart model) There exist i.i.d. z_j with $\mathbb{E}[z_j z_j^T] = I$ such that $z_j = \Sigma^{-\frac{1}{2}} g_j$ is the whitened gradient. The minibatch average satisfies $\hat{g}_t \approx (1/b) \sum_{i \in B_t} g_i$ with covariance Σ/b.
- The preconditioner is an EMA of outer products (second moments):
$$G_t = \alpha G_{t-1} + (1-\alpha) g_t g_t^T$$
with effective sample size $m \approx 1/(1-\alpha)$.
- The curvature proxy H is SPD and comparable to Σ:
$$H = \Sigma^{\frac{1}{2}} T \Sigma^{\frac{1}{2}}$$
with T SPD. Special cases: $T = I$ (pure Fisher), or spiked $T = I + \sum_l (\beta_l - 1) u_l u_l^T$.
- The eigenbasis of G is recomputed every τ steps; between refreshes G drifts with $\|G_{t+\tau} - G_t\|_2 \le L\tau$ for some constant $L > 0$. This gives us a handle on staleness.
- Define the (full) preconditioned curvature
$$H' := \hat{G}^{-\frac{1}{2}} H \hat{G}^{-\frac{1}{2}}$$

3.2. **Quantities of interest.**

We will track the following spectral quantities of the preconditioned curvature $H' = \hat{G}^{-\frac{1}{2}} H \hat{G}^{-\frac{1}{2}}$:

- Extremal eigenvalues and condition number: $\lambda_{\max}(H')$, $\lambda_{\min}(H')$, and
$$\kappa(H') := \lambda_{\max}(H')/\lambda_{\min}(H').$$
These summarize the residual anisotropy after preconditioning.
- Stable step size for gradient descent: on a quadratic with curvature H', gradient descent with step size η is stable only if $\eta < 2/\lambda_{\max}(H')$; thus $\lambda_{\max}(H')$ effectively upper-bounds admissible learning rates.
- Effective aspect ratio: $\gamma := \frac{d}{m}$, where d is the parameter dimension and m is the effective sample size used to form $\hat{G}$. In our Wishart/RMT model, γ controls the Marchenko–Pastur bulk edges and hence the typical spread of eigenvalues of the empirical preconditioner.

4. Main Theoretical Results

We now state our main theoretical claims regarding the spectral behavior of H' under the above model and assumptions. In particular, this paper focuses on

i) Finite-sample spectral bounds on H' via Marchenko-Pastur.
ii) Spike detectability in H' via Baik-Ben Arous-Péché.
iii) Eigenspace staleness effects on H' via Davis-Kahan and Lipschitz perturbation bounds.

4.1. Finite-sample spectral bounds via Marchenko-Pastur.

Our first theorem characterizes the finite-sample spectral edges of H' as a function of the effective aspect ratio γ. In particular, we are interested in bounding the condition number $\kappa(H')$ to ensure stable optimization. For a target condition number κ^*, we derive a lower bound on the required effective sample size m (and relatedly the α parameter in Shampoo/SOAP) to achieve $\kappa(H') \leq \kappa^*$ with high probability. We first lay out the theoretical backing for the finite-sample bulk edges of H' by leveraging the Marchenko-Pastur law. We then provide emperical evidence through numerical simulations in Python to validate our theoretical predictions.

Theorem 4.1.1: **(Finite-sample spectral bounds for the preconditioned curvature).** Let $\Sigma \in \mathbb{R}^{d\times d}$ be positive definite and let $\hat{G} = \left(\frac{1}{m}\right)\sum_{i=1}^{m} g_i g_i^T$ be the empirical second-moment matrix of i.i.d. samples $g_i \sim \mathcal{N}(0, \Sigma)$. Define the whitened empirical covariance $S := \Sigma^{-\frac{1}{2}} \hat{G} \Sigma^{-\frac{1}{2}}$, and suppose $d, m \to \infty$ with aspect ratio $\gamma := \frac{d}{m}$ converging to a constant in $(0, 1)$. Let $T \in \mathbb{R}^{d\times d}$ be positive definite and define $H := \Sigma^{\frac{1}{2}} T \Sigma^{\frac{1}{2}}$ and $H' := \hat{G}^{-\frac{1}{2}} H \hat{G}^{-\frac{1}{2}} = S^{-\frac{1}{2}} T S^{-\frac{1}{2}}$. Then, if $\kappa^* > 1$ is a desired upper bound on $\kappa(H')$, it suffices to choose m such that

$$\frac{m}{d} \geq \frac{\left(\sqrt{\frac{\kappa^*}{\kappa(T)}} + 1\right)^2}{\left(\sqrt{\frac{\kappa^*}{\kappa(T)}} - 1\right)^2}.$$

Proof: By the Marchenko-Pastur law, the eigenvalues of $S = \Sigma^{-\frac{1}{2}} \hat{G} \Sigma^{-\frac{1}{2}}$ lie in $\left[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2\right]$ with high probability. Equivalently,

$$(1-\sqrt{\gamma})^2 I \preceq S \preceq (1+\sqrt{\gamma})^2 I.$$

To bound the spectrum of $H' = S^{-\frac{1}{2}} T S^{-\frac{1}{2}}$, consider its Rayleigh quotient. For any unit vector x, let $y = S^{-\frac{1}{2}} x$. Then

$$x^T H' x = x^T S^{-\frac{1}{2}} T S^{-\frac{1}{2}} x = \frac{y^T T y}{y^T S y}.$$

Since T and S are positive definite,

$$\lambda_{\min}(T)\ \|y\|^2 \leq y^T T y \leq \lambda_{\max}(T)\ \|y\|^2,$$

and, using the MP bounds on S,

$$(1-\sqrt{\gamma})^2\ \|y\|^2 \le y^T S y \le (1+\sqrt{\gamma})^2\ \|y\|^2.$$

Combining numerator and denominator yields, for all $y \neq 0$,

$$\frac{\lambda_{\min}(T)}{(1+\sqrt{\gamma})^2} \le \frac{y^T T y}{y^T S y} \le \frac{\lambda_{\max}(T)}{(1-\sqrt{\gamma})^2}.$$

Taking minima and maxima over unit vectors x gives the extremal eigenvalue bounds

$$\lambda_{\min}(H') \ge \frac{\lambda_{\min}(T)}{(1+\sqrt{\gamma})^2}, \quad \lambda_{\max}(H') \le \frac{\lambda_{\max}(T)}{(1-\sqrt{\gamma})^2}.$$

Consequently,

$$\kappa(H') \le \kappa(T)\frac{(1+\sqrt{\gamma})^2}{(1-\sqrt{\gamma})^2}.$$

To guarantee $\kappa(H') \le \kappa^*$, it suffices that

$$\frac{(1+\sqrt{\gamma})^2}{(1-\sqrt{\gamma})^2} \le \frac{\kappa^*}{\kappa(T)},$$

which rearranges to the stated condition on the effective sample size,

$$\frac{m}{d} \ge \frac{\left(\sqrt{\frac{\kappa^*}{\kappa(T)}}+1\right)^2}{\left(\sqrt{\frac{\kappa^*}{\kappa(T)}}-1\right)^2}.$$

This completes the proof. ∎

Remark 4.1.1: **pessimism of the worst-case bound.** The inequality above is sharp only in an adversarial sense: it assumes that the eigenvector of T corresponding to $\lambda_{\max}(T)$ aligns perfectly with the eigenvector of S corresponding to $\lambda_{\min}(S)$, and vice versa. In our random model, S is a (whitened) Wishart matrix whose eigenvectors are in generic relative position with those of T. Such worst-case alignments occur with negligible probability, so the empirical eigenvalues of H' are typically far smaller than the worst-case upper bound and far larger than the worst-case lower bound. This explains why, in our numerical experiments, the bounds in Theorem 4.1.1 of H' is significantly better behaved in practice.

4.2. **Spiked directions and BBP detectability.**

The previous theorem treated T as an arbitrary positive definite matrix and provided worst-case bounds that depend only on its condition number $\kappa(T)$. In many optimization problems, however, curvature is dominated by a few strong directions (e.g., a handful of layers or features with much larger curvature than the rest). In fact, because the SOAP/Shampoo optimizers use per-mode preconditioners, a spike as simple as a rank-1 perturbation in T can lead to undetected outliers in H' if the spike is not visible in the per-mode marginals of G.

In this section we specialize to a **spiked** model for T and use the Baik-Ben Arous-Péché (BBP) theory for spiked covariance matrices to characterize when such strong directions are even detectable to the empirical preconditioner $\hat{G}$. Intuitively, a curvature spike of strength $\beta > 1$ in direction u can only be effectively "whitened away" if $\hat{G}$ resolves u as a separate eigen-direction. BBP theory provides a sharp threshold: below a critical value $\beta_c(\gamma)$ the spike is statistically indistinguishable from the Marchenko-Pastur bulk, while above this threshold a separated outlier eigenvalue and a non-trivial eigenvector overlap appear.

Theorem 4.2.1: **(Spiked preconditioner and BBP detectability).** Consider the one-spike model in dimension d with

$$T = I + (\beta - 1)uu^T,$$

where $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$ and spike strength $\beta > 1$. Let

$$g_i \sim \mathcal{N}(0, T), \quad i = 1, ..., m,$$

and define the empirical covariance

$$\hat{G} = \left(\frac{1}{m}\right) \sum_{i=1}^{m} g_i g_i^T.$$

Assume $d, m \to \infty$ with aspect ratio $\gamma := d/m \to \gamma_0 \in (0, 1)$. Let

$$b_\gamma := (1 + \sqrt{\gamma_0})^2, \quad \beta_c := 1 + \sqrt{\gamma_0},$$

and, for $\beta > \beta_c$, define the BBP outlier location

$$\lambda_{\text{out}(\beta,\gamma_0)} := \beta\left(1 + \frac{\gamma_0}{\beta - 1}\right).$$

Then, with probability tending to 1 as $d, m \to \infty$:

- **(Subcritical spike, $\beta \le \beta_c$.)** The spectrum of $\hat{G}$ has no outliers: all eigenvalues lie in $\left[a_\gamma, b_\gamma\right]$ up to vanishing fluctuations, and
$$\lambda_{\max}\left(\hat{G}\right) \to b_\gamma.$$
The leading eigenvector $v_{\max}$ of $\hat{G}$ is asymptotically orthogonal to u. Consequently, the curvature of the preconditioned matrix
$$H' := \hat{G}^{-\frac{1}{2}} T \hat{G}^{-\frac{1}{2}}$$
along u is controlled only by the bulk bounds from Theorem 4.1.1: the spike behaves like a generic direction inside the Marchenko–Pastur bulk.
- **(Supercritical spike, $\beta > \beta_c$.)** A unique outlier eigenvalue of $\hat{G}$ emerges above the bulk:
$$\lambda_{\max}\left(\hat{G}\right) \to \lambda_{\text{out}}(\beta, \gamma_0) > b_\gamma,$$
with associated eigenvector $v_{\max}$ satisfying
$$| < v_{\max}, u > |^2 \to c(\beta, \gamma_0)$$
for some explicit $c(\beta, \gamma_0) \in (0, 1)$. Along this data-driven spike direction, the preconditioned curvature satisfies the asymptotic Rayleigh quotient
$$v_{\max}^T H' v_{\max} = v_{\max}^T \hat{G}^{-\frac{1}{2}} T \hat{G}^{-\frac{1}{2}} v_{\max} \to \beta/\lambda_{\text{out}}(\beta, \gamma_0) = 1/\left(1 + \frac{\gamma_0}{\beta - 1}\right) < 1.$$
In particular, once the spike crosses the BBP threshold, whitening suppresses its effective curvature by a factor $\lambda^{-1}_{\text{out}(\beta,\gamma_0)}$, and the residual largest curvature of H' is governed primarily by the Marchenko–Pastur bulk.

Proof: Writing T in its eigenbasis, with eigenvalues β along u and 1 on the orthogonal complement, the sample covariance $\hat{G}$ is a rank-1 spiked Wishart matrix with population covariance T. Classical results (e.g. Baik-Ben Arous-Péché and Baik-Silverstein) show that, under the high-dimensional limit $d, m \to \infty$ with $d/m \to \gamma_0$, the following hold.

First, the empirical spectral distribution of $\hat{G}$ converges almost surely to the Marchenko-Pastur law with parameter γ_0, and all but finitely many eigenvalues lie in a vanishing neighborhood of the bulk interval

$$\left[a_\gamma, b_\gamma\right] = \left[(1 - \sqrt{\gamma_0})^2, (1 + \sqrt{\gamma_0})^2\right].$$

Second, there is a sharp BBP transition. If the population spike β satisfies

$$1 < \beta \le \beta_c := 1 + \sqrt{\gamma_0},$$

then no outlier separates from the bulk: the top eigenvalue of $\hat{G}$ converges to the upper edge b_γ, and the associated eigenvector $v_{\max}$ is asymptotically orthogonal to the spike direction u. This yields the "subcritical'' bullet in the theorem. Since $v_{\max}$ behaves like a generic bulk vector, the curvature of $H' = \hat{G}^{-\frac{1}{2}} T \hat{G}^{-\frac{1}{2}}$ along u is governed by the Marchenko–Pastur bounds from Theorem 4.1.1, with no special whitening effect on the spike beyond the bulk. If instead

$$\beta > \beta_c,$$

then there exists a unique outlier eigenvalue λ_{out} of $\hat{G}$ that converges almost surely to

$$\lambda_{\text{out}}(\beta, \gamma_0) = \beta\left(1 + \frac{\gamma_0}{\beta - 1}\right) > b_\gamma,$$

while all other eigenvalues remain in the bulk. Moreover, the corresponding eigenvector $v_{\max}$ (associated with λ_{out}) has non-trivial asymptotic overlap with u:

$$| < v_{\max}, u > |^2 \to c(\beta, \gamma_0)$$

for an explicit $c(\beta, \gamma_0) \in (0, 1)$, whereas bulk eigenvectors have vanishing overlap with u. To relate this to the preconditioned curvature, consider the Rayleigh quotient of

$$H' = \hat{G}^{-\frac{1}{2}} T \hat{G}^{-\frac{1}{2}}$$

along $v_{\max}$:

$$v_{\max}^T H' v_{\max} = v_{\max}^T \hat{G}^{-\frac{1}{2}} T \hat{G}^{-\frac{1}{2}} v_{\max}.$$

Since $\hat{G} v_{\max} = \lambda_{\text{out}} v_{\max}$, we have

$$\hat{G}^{-\frac{1}{2}} v_{\max} = \lambda_{\text{out}}^{-\frac{1}{2}} v_{\max},$$

and hence

$$v_{\max}^T H' v_{\max} = \lambda_{\text{out}}^{-1} v_{\max}^T T v_{\max}.$$

Decomposing $v_{\max} = \alpha u + w$ with $< w, u >= 0$ and $|\alpha|^2 \to c(\beta, \gamma_0)$, we obtain

$$v_{\max}^T T v_{\max} = \beta \ |\alpha|^2 + \|w\|^2 \to \beta c(\beta, \gamma_0) + (1 - c(\beta, \gamma_0)),$$

which is asymptotically of order β for a fixed spike $\beta > \beta_c$. Combining with the limit of λ_{out} yields the asymptotic expression

$$v_{\max}^T H' v_{\max} \to \beta / \lambda_{\text{out}}(\beta, \gamma_0) = 1/(1 + \gamma_0/(\beta - 1)),$$

as stated. Since $\gamma_0/(\beta - 1) > 0$, this value is strictly less than 1, showing that whitening suppresses the spike direction once it is detected.

A fully rigorous argument tracks the joint convergence of eigenvalues and eigenvectors in the spiked model, but the key point is that below the BBP threshold the spike is indistinguishable from the bulk (and thus only MP-based bounds apply), whereas above the threshold the spike becomes a dedicated eigen-direction that the empirical preconditioner can actively whiten. ■

Remark 4.2.1: **Multiple spikes and higher-rank structure.** The one-spike theorem extends to a finite number of spikes

$$T = I + \sum_{\ell=1}^{r} (\beta_\ell - 1) u_\ell u_\ell^T$$

with orthonormal $\{u_\ell\}$. Each supercritical spike $\beta_\ell > 1 + \sqrt{\gamma_0}$ gives rise to its own outlier eigenvalue

$$\lambda_{\text{out}}^\ell = \beta_\ell \left(1 + \frac{\gamma_0}{\beta_\ell - 1}\right)$$

and an associated signal subspace with non-trivial alignment to the population spike subspace. If a spike has multiplicity $k > 1$, the corresponding outliers form a tight cluster just above b_γ with a k-dimensional signal eigenspace.

From the optimizer's perspective, this means that a small number of strong curvature directions can be systematically detected and whitened by the empirical preconditioner once they cross the BBP threshold, while weaker spikes remain hidden inside the Marchenko–Pastur bulk and survive as residual outliers in H'. For per-mode (Kronecker-structured) preconditioners such as Shampoo/SOAP, cross-mode spikes that do not meet the detectability threshold in the low-dimensional per-mode marginals may remain as undetected large eigenvalues, capping the allowable step size via $\lambda_{\max}(H')$ even when the bulk is well-conditioned.

4.3. **Stale eigenbasis and Davis–Kahan.**

So far our analysis has been static: we assumed that the preconditioner $\hat{G}$ is recomputed exactly from m samples and used immediately. In practice, SOAP/Shampoo only refresh the eigendecomposition every τ steps. Between refreshes, the EMA update

$$\hat{G}_{t+1} = (1-\alpha)\hat{G}_t + \alpha g_{t+1} g_{t+1}^T$$

causes the preconditioner to drift, while the optimizer continues to use the "stale" eigenbasis from time t to precondition gradients at times $t+1, ..., t+\tau$. This raises a natural question: How much can eigenbasis staleness inflate the spectral edges of the preconditioned curvature H'?

In this section we make this question precise. For the spike directions we exploit the fact that, in the supercritical BBP regime of Theorem 4.2.1, the outlier eigenvalues of $\hat{G}$ are separated from the Marchenko-Pastur bulk by a positive spectral gap. This allows us to apply the Davis-Kahan $\sin\Theta$ theorem to show that the outlier subspace (the span of the detected spikes) can only rotate proportionally to the drift $\|\hat{G}_{t+\tau} - \hat{G}_t\|_2$, where all matrix norms are operator 2-norms unless otherwise stated. For the entire spectrum of the preconditioned curvature H', we do not need any eigengap: a simple Lipschitz bound for the matrix map $A \mapsto A^{-\frac{1}{2}}$, combined with Weyl's inequality, shows that all eigenvalues of H' change at most linearly with the drift in $\hat{G}$.

4.3.1. *Outlier subspace stability via Davis-Kahan.*

We first formalize the "signal" subspace of outliers. Under the spiked model of Theorem 4.2.1, let $b_\gamma = \left(1+\sqrt{\gamma_0}\right)^2$ denote the Marchenko-Pastur upper edge and let $\lambda_{\text{out}}^\ell$ be the location of the ℓ-th outlier eigenvalue associated with a supercritical spike $\beta_\ell > 1 + \sqrt{\gamma_0}$. Define the asymptotic BBP gap

$$\text{gap}_{\text{BBP}} := \min_\ell \left(\lambda_{\text{out}}^\ell - b_\gamma\right) > 0.$$

For large d, m, the sample eigenvalues of $\hat{G}_t$ concentrate near b_γ and $\{\lambda_{\text{out}}^\ell\}_\ell$; hence, with high probability, one can choose a fixed threshold θ such that

$$b_\gamma < \theta < \min_\ell \lambda_{\text{out}}^\ell,$$

and all bulk eigenvalues lie below θ while all outliers lie above θ. For each time t, let $U_t \in \mathbb{R}^{d\times k}$ have as columns the eigenvectors of $\hat{G}_t$ whose eigenvalues are at least θ; this defines the **outlier subspace** of $\hat{G}_t$. We now quantify how much this subspace can drift between refreshes.

Theorem 4.3.1.1: **(Davis-Kahan control of the outlier subspace).** Assume the spiked model and high-dimensional regime of Theorem 4.2.1, with at least one supercritical spike $\beta_\ell > 1 + \sqrt{\gamma_0}$, and let U_t and $U_{t+\tau}$ denote the outlier subspaces of $\hat{G}_t$ and $\hat{G}_{t+\tau}$ defined by a threshold θ satisfying

$$b_\gamma < \theta < \min_\ell \lambda_{\text{out}}^\ell.$$

Then, for all sufficiently large d, m, there exists a constant $\text{gap}_0 > 0$ (depending only on the spike strengths and γ_0) such that, with high probability,

$$\mathrm{gap}_0 \le \inf_t \left(\min_{\lambda_i(\hat{G}_t) \ge \theta} \lambda_i\big(\hat{G}_t\big) - \max_{\lambda_j(\hat{G}_t) < \theta} \lambda_j\big(\hat{G}_t\big) \right).$$

Moreover, for any fixed times t and $t+\tau$,

$$\| \sin \Theta(U_t, U_{t+\tau}) \|_2 \le \frac{\| \hat{G}_{t+\tau} - \hat{G}_t \|_2}{\mathrm{gap}_0},$$

where $\Theta(U_t, U_{t+\tau})$ denotes the diagonal matrix of principal angles between the two outlier subspaces. In particular, if the preconditioner drift over τ steps is bounded as

$$\| \hat{G}_{t+\tau} - \hat{G}_t \|_2 \le L\tau$$

for some constant $L > 0$, then

$$\| \sin \Theta(U_t, U_{t+\tau}) \|_2 \le \left(\frac{L}{\mathrm{gap}_0} \right) \tau.$$

Proof: The existence of a positive asymptotic gap between the outliers and the bulk follows from Theorem 4.2.1. In the supercritical regime $\beta_\ell > 1 + \sqrt{\gamma_0}$, the ℓ-th outlier eigenvalue $\lambda^\ell_{\mathrm{out}}$ converges to

$$\lambda^\ell_{\mathrm{out}} = \beta_\ell \left(1 + \frac{\gamma_0}{\beta_\ell - 1} \right) > b_\gamma,$$

so $\lambda^\ell_{\mathrm{out}} - b_\gamma$ is strictly positive for each such spike. Taking the minimum over ℓ yields a positive asymptotic gap $\mathrm{gap}_{\mathrm{BBP}} > 0$. By standard eigenvalue concentration results for spiked sample covariance matrices, the sample eigenvalues of $\hat{G}_t$ lie within $o(1)$ of their limits as $d, m \to \infty$. Hence, for large enough d, m, we may choose a fixed threshold θ satisfying

$$b_\gamma < \theta < \min_\ell \lambda^\ell_{\mathrm{out}}$$

such that, with high probability, all bulk eigenvalues of $\hat{G}_t$ lie below θ and all outliers lie above θ for all t in a bounded time window. This yields a finite d eigengap $\mathrm{gap}_0 > 0$ between the outlier cluster and the bulk, uniform in t.

Fix such a θ and the corresponding outlier subspaces U_t and $U_{t+\tau}$ for $\hat{G}_t$ and $\hat{G}_{t+\tau}$. Let gap_t denote the spectral gap of $\hat{G}_t$ between the smallest eigenvalue in the outlier cluster and the largest eigenvalue in the bulk. The Davis–Kahan $\sin\Theta$ theorem (for example, in the version for invariant subspaces associated to disjoint spectral clusters) states that, for symmetric matrices A and $\tilde{A} = A + E$ and the associated invariant subspaces U and $\tilde{U}$,

$$\| \sin \Theta\big(U, \tilde{U}\big) \|_2 \le \frac{\|E\|_2}{\mathrm{gap}},$$

where gap is the minimal distance between the eigenvalues of A in the cluster defining U and those outside the cluster. Applying this with $A = \hat{G}_t$, $\tilde{A} = \hat{G}_{t+\tau}$, $E = \hat{G}_{t+\tau} - \hat{G}_t$, and $U = U_t$, $\tilde{U} = U_{t+\tau}$ gives

$$\| \sin \Theta(U_t, U_{t+\tau}) \|_2 \le \frac{\| \hat{G}_{t+\tau} - \hat{G}_t \|_2}{\mathrm{gap}_t}.$$

Since by construction $\mathrm{gap}_t \ge \mathrm{gap}_0 > 0$ for all t in the event of interest, we obtain the claimed bound

$$\| \sin \Theta(U_t, U_{t+\tau}) \|_2 \le \frac{\| \hat{G}_{t+\tau} - \hat{G}_t \|_2}{\mathrm{gap}_0}.$$

If, in addition, the preconditioner drift satisfies the Lipschitz bound $\| \hat{G}_{t+\tau} - \hat{G}_t \|_2 \le L\tau$, the linear-in-$\tau$ bound follows immediately:

$$\| \sin \Theta(U_t, U_{t+\tau}) \|_2 \leq \left(\frac{L}{\mathrm{gap}_0} \right) \tau.$$

This completes the proof. ∎

Remark 4.3.1.1: **Interpretation for SOAP/Shampoo.** For optimizers whose curvature is dominated by a few strong directions (supercritical spikes), Theorem 4.3.1.1 shows that the **subspace** spanned by those directions in the empirical preconditioner $\hat{G}_t$ is stable over time, as long as the BBP gap gap_0 is not too small and the EMA update does not move $\hat{G}$ too aggressively. In particular, the dominant curvature directions that determine the largest eigenvalues of H' cannot rotate arbitrarily between refreshes; their drift is controlled by the preconditioner drift $\| \hat{G}_{t+\tau} - \hat{G}_t \|_2$.

4.3.2. *Spectral stability of the preconditioned curvature.*

Davis-Kahan controls only the subspace associated with outliers in $\hat{G}$. To quantify the effect of eigenbasis staleness on the entire spectrum of the preconditioned curvature H', we combine two simple ingredients:

- The map $A \mapsto A^{-\frac{1}{2}}$ is Lipschitz on the cone of SPD matrices with eigenvalues bounded below by $\alpha > 0$.
- Weyl's inequality bounds the change in eigenvalues of a symmetric matrix by its operator norm perturbation.

This yields a deterministic bound on how much any eigenvalue of H' can change between the "fresh" and "stale" preconditioners.

Theorem 4.3.2.1: **(Spectral stability of H' under stale preconditioning).** Let $\hat{G}_t$ and $\hat{G}_{t+\tau}$ be symmetric positive-definite matrices with

$$\lambda_{\min}\big(\hat{G}_t\big) \geq \alpha, \quad \lambda_{\min}\big(\hat{G}_{t+\tau}\big) \geq \alpha$$

for some $\alpha > 0$. Let H be positive definite and define the preconditioned curvatures

$$H'_{\mathrm{fresh}} := \hat{G}_{t+\tau}^{-\frac{1}{2}} H \hat{G}_{t+\tau}^{-\frac{1}{2}}, \quad H'_{\mathrm{stale}} := \hat{G}_t^{-\frac{1}{2}} H \hat{G}_t^{-\frac{1}{2}}.$$

Then

$$\| H'_{\mathrm{stale}} - H'_{\mathrm{fresh}} \|_2 \leq \left(\frac{1}{\alpha^2} \right) \|H\|_2 \, \|\hat{G}_{t+\tau} - \hat{G}_t\|_2.$$

In particular, for every eigenvalue index i,

$$|\lambda_{i(H'_{\mathrm{stale}})} - \lambda_{i(H'_{\mathrm{fresh}})}| \leq \left(\frac{1}{\alpha^2} \right) \|H\|_2 \, \|\hat{G}_{t+\tau} - \hat{G}_t\|_2.$$

If, in addition, the preconditioner drift satisfies

$$\| \hat{G}_{t+\tau} - \hat{G}_t \|_2 \leq L\tau$$

for some $L > 0$, then the inflation of the spectral edges is at most linear in τ:

$$|\lambda_{\max}(H'_{\mathrm{stale}}) - \lambda_{\max}(H'_{\mathrm{fresh}})| \leq \left(\frac{L}{\alpha^2} \right) \|H\|_2 \, \tau,$$

and similarly for $\lambda_{\min}$.

Proof: Define

$$\Delta := \hat{G}_{t+\tau}^{-\frac{1}{2}} - \hat{G}_t^{-\frac{1}{2}}.$$

Then we can write

$$H'_{\mathrm{stale}} - H'_{\mathrm{fresh}} = \Delta H \hat{G}_t^{-\frac{1}{2}} + \hat{G}_{t+\tau}^{-\frac{1}{2}} H \Delta.$$

Taking operator norms and using the triangle inequality,

$$\| H'_{\text{stale}} - H'_{\text{fresh}} \|_2 \le \|\Delta\|_2 \, \|H\|_2 \, \|\hat{G}_t^{-\frac{1}{2}}\|_2 + \|\hat{G}_{t+\tau}^{-\frac{1}{2}}\|_2 \, \|H\|_2 \, \|\Delta\|_2$$
$$= 2 \, \|\Delta\|_2 \, \|H\|_2 \max\Big\{\|\hat{G}_t^{-\frac{1}{2}}\|_2, \|\hat{G}_{t+\tau}^{-\frac{1}{2}}\|_2\Big\}.$$

The eigenvalue lower bound $\lambda_{\min}\big(\hat{G}_t\big), \lambda_{\min}\big(\hat{G}_{t+\tau}\big) \ge \alpha$ implies

$$\max\Big\{\|\hat{G}_t^{-\frac{1}{2}}\|_2, \| \hat{G}_{t+\tau}^{-\frac{1}{2}} \|_2\Big\} \le \alpha^{-\frac{1}{2}}.$$

To bound $\|\Delta\|_2$, observe that the scalar function $f(x) = x^{-\frac{1}{2}}$ is differentiable and has derivative $f'(x) = -\big(\frac{1}{2}\big)x^{-\frac{3}{2}}$, so on the interval $[\alpha, \infty)$ it is Lipschitz with constant $\big(\frac{1}{2}\big)\alpha^{-\frac{3}{2}}$. By functional calculus for symmetric matrices, this implies the matrix-function Lipschitz bound

$$\|\hat{G}_{t+\tau}^{-\frac{1}{2}} - \hat{G}_t^{-\frac{1}{2}}\|_2 \le \left(\frac{1}{2}\right)\alpha^{-\frac{3}{2}} \, \|\hat{G}_{t+\tau} - \hat{G}_t\|_2,$$

$$\|\Delta\|_2 \le \left(\frac{1}{2}\right)\alpha^{-\frac{3}{2}} \, \|\hat{G}_{t+\tau} - \hat{G}_t\|_2.$$

Substituting these inequalities into the previous display yields

$$\| H'_{\text{stale}} - H'_{\text{fresh}} \|_2 \le 2 \cdot \left(\frac{1}{2}\right)\alpha^{-\frac{3}{2}} \, \|\hat{G}_{t+\tau} - \hat{G}_t\|_2 \, \|H\|_2 \, \alpha^{-\frac{1}{2}} = \left(\frac{1}{\alpha^2}\right) \, \|H\|_2 \, \|\hat{G}_{t+\tau} - \hat{G}_t\|_2,$$

as claimed. Finally, Weyl's inequality for symmetric matrices states that the change in any eigenvalue is bounded by the operator norm of the perturbation:

$$|\lambda_i(H'_{\text{stale}}) - \lambda_i(H'_{\text{fresh}})| \le \| H'_{\text{stale}} - H'_{\text{fresh}} \|_2.$$

Combining this with the previous bound proves the eigenvalue statements. Under the additional drift assumption $\|\hat{G}_{t+\tau} - \hat{G}_t\|_2 \le L\tau$, the linear-in-$\tau$ bounds follow immediately. This completes the proof. ∎

Remark 4.3.2.1: **Outliers versus bulk under staleness.** The two theorems above play complementary roles. Theorem 4.3.1.1 shows that the **subspace** spanned by outlier directions of $\hat{G}_t$ (strong curvature directions above the BBP threshold) is stable, so the optimizer continues to precondition approximately along the correct spike directions even when the eigendecomposition is refreshed infrequently. Theorem 4.3.2.1 shows that, regardless of spectral gaps, **every eigenvalue** of the preconditioned curvature H' can change by at most $O\big(\|\hat{G}_{t+\tau} - \hat{G}_t\|_2\big)$ under staleness. In particular, the largest eigenvalue $\lambda_{\max}(H')$—which controls the stable step size $\eta < 2/\lambda_{\max}(H')$—cannot inflate arbitrarily between refreshes as long as the EMA dynamics keep $\hat{G}$ well-conditioned and slowly varying. Combined with the BBP analysis, this suggests that the main driver of conditioning in practice is the Marchenko–Pastur bulk anisotropy (controlled by the aspect ratio γ), while both spikes and eigenbasis staleness are effectively tamed by the empirical preconditioner.

5. Experiments and Evaluations

We perform numerical simulations in Python to confirm the predicted spectral edge behavior under varying (m, τ, d) and model parameters. We generate synthetic data from the elliptical/Wishart model and compute empirical spectra of H'.

5.1. Spectral behavior of the sample covariance S.

We first verify that the sample covariance $S = \Sigma^{-\frac{1}{2}}\hat{G}\Sigma^{-\frac{1}{2}}$ obeys the Marchenko-Pastur law when $\hat{G}$ is formed by sampling m i.i.d. Gaussian vectors with covariance Σ. We vary the effective aspect ratio $\gamma = d/m$ and compute the empirical extremal eigenvalues of S over multiple trials, comparing them to the theoretical MP bulk edges. Figure 1 plots the emperical means of $\lambda_{\min}(S)$ and $\lambda_{\max}(S)$

over 50 trials for aspect ratio $\gamma = d/m \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, together with the theoretical bulk edges $(1 \pm \sqrt{\gamma})^2$.

Empirical observation. Across all tested aspect ratios, the extremal eigenvalues of S lie extremely close to the Marchenko-Pastur edges, confirming that the MP approximation is accurate even at moderate dimension ($d = 128$) and justifying its use in Theorem 4.1.1.

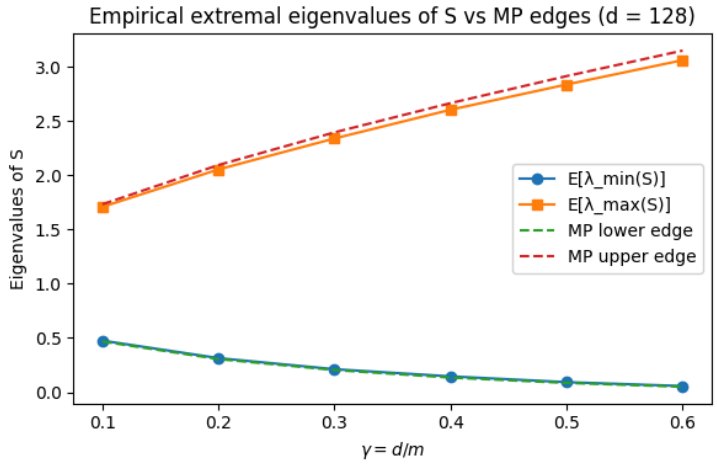

FIGURE 1. Empirical extremal eigenvalues of the sample covariance S versus theoretical Marchenko-Pastur bulk edges as a function of the effective aspect ratio $\gamma = d/m$

5.2. Empirical Validation of Theorem 4.1.1: spectral bounds on H'.

We next examine the spectrum of the preconditioned curvature $H' = \hat{G}^{-\frac{1}{2}} H \hat{G}^{-\frac{1}{2}}$ for a diagonal $T = I$. For each aspect ratio γ, we compute $\lambda_{\min}(H')$ and $\lambda_{\max}(H')$ over 50 trials and compare them with the theoretical bounds in Theorem 4.1.1.

$$\lambda_{\min}(H') \geq \frac{\lambda_{\min}(T)}{(1+\sqrt{\gamma})^2}, \quad \lambda_{\max}(H') \leq \frac{\lambda_{\max}(T)}{(1-\sqrt{\gamma})^2}.$$

Figure 2 and Figure 3 show that the empirical eigenvalues of H' indeed fall within the predicted bounds for all γ tested. In addition, while the bounds become looser as γ increases (smaller m), the empirical eigenvalues remain well-behaved, indicating that the preconditioner $\hat{G}$ effectively whitens H even at moderate sample sizes. Even as the theoretical bounds diverged from the empirical values, the actual eigenvalues remained within the same order of magnitude, suggesting that the preconditioning remains effective.

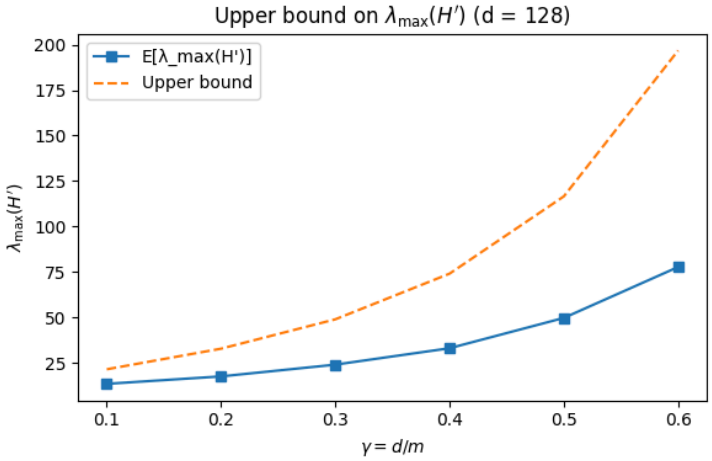

FIGURE 2. Empirical $\lambda_{\min}(H')$ versus theoretical lower bound as a function of γ

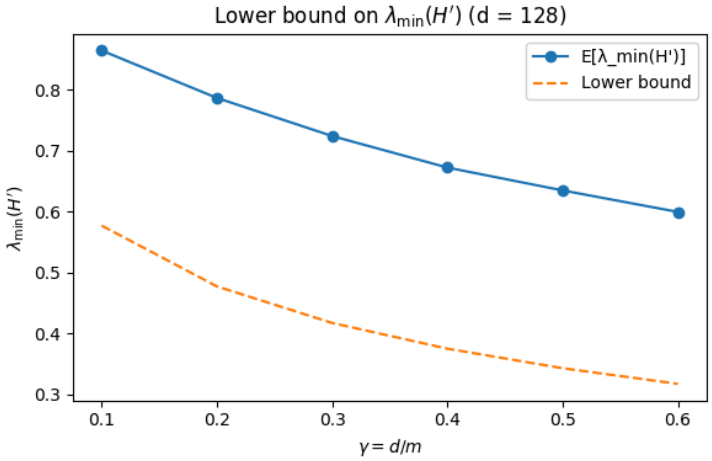

FIGURE 3. Empirical $\lambda_{\max}(H')$ versus theoretical upper bound as a function of γ

Combining the min and the max, we also confirm that the condition number $\kappa(H')$ behaves as predicted by Theorem 4.1.1. We plot the empirical $\kappa(H')$ against the theoretical upper bounds as a function of the aspect ratio γ in Figure 4, confirming that the condition number remains controlled and within the predicted envelope. As expected, the empirical condition number grows monotonically with γ, reflecting increased sampling noise in $\hat{G}$. For moderate aspect ratios $\gamma \leq 0.5$, the observed conditioning remains within a small constant factorof the theoretical upper bound, while for larger γ the bound becomes increasingly conservative.

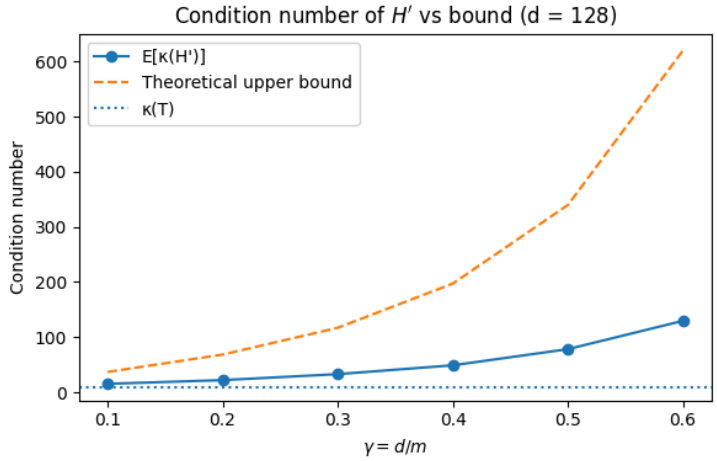

FIGURE 4. Empirical condition number $\kappa(H')$ versus theoretical upper bound as a function of γ

Remark. The bounds from Theorem 4.1.1 are naturally pessimistic: they assume the most adversarial alignment between the eigenvectors of T and the extremal eigenvectors of S. In our random model these eigenspaces are in generic relative position, so the empirical eigenvalues of H' lie well inside the worst-case envelope. This explains why the empirical eigenvalues remain significantly better behaved than the theoretical bounds, especially at larger γ.

5.3. Empirical Validation of Theorem 4.2.1: spike detectability in H'.

We now empirically validate the spiked model analysis in Theorem 4.2.1. We work in the stylized one-spike setting $T = I + (\beta - 1)uu^T$ with $\|u\|_2 = 1$, aspect ratio $\gamma = d/m = 0.5$, and dimension $d = 256$. For each spike strength β and each trial, we draw $g_i \sim \mathcal{N}(0, T)$, form the sample covariance $\hat{G} = (1/m)\sum_{i=1}^{m} g_i g_i^T$, and compute the preconditioned curvature $H' = \hat{G}^{-\frac{1}{2}} T \hat{G}^{-\frac{1}{2}}$. We record three quantities of interest:

i) the top eigenvalue $\lambda_{\max}\big(\hat{G}\big)$,
ii) the squared overlap $|\langle v_{\max}, u\rangle|^2$ between the $\hat{G}$'s top eigenvector and the spike direction u
iii) the Rayleigh quotient $v_{\max}^T H' v_{\max}$ describing curvature along the detected spike direction.

5.3.1. *Eigenvalue phase transition (BBP).*

Figure 5 shows the empirical mean of $\lambda_{\max}\big(\hat{G}\big)$ over 50 trials as a function of β, together with the Marchenko-Pastur upper edge $(1+\sqrt{\gamma})^2$ and the BBP outlier prediction $\lambda_{\text{out}(\beta,\gamma)} = \beta\big(1 + \frac{\gamma}{\beta - 1}\big)$ for $\beta > \beta_c$. The vertical line marks the BBP threshold $\beta_c = 1 + \sqrt{\gamma}$.

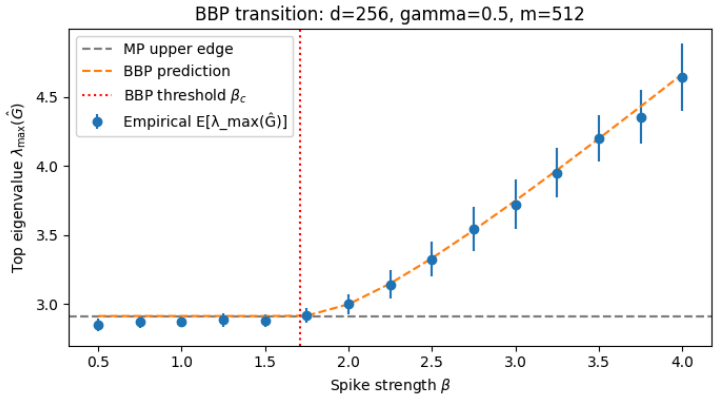

FIGURE 5. Top eigenvalue of the sample covariance $\hat{G}$ as a function of spike strength β, compared against the MP upper edge and BBP outlier prediction for $\gamma = d/m = 0.5$.

Empirical observation. For $\beta < \beta_c$, the top eigenvalue of $\hat{G}$ sits tightly at the MP upper edge, indicating that the spike is statistically indistinguishable from bulk noise. As β crosses β_c, the empirical $\lambda_{\max}\left(\hat{G}\right)$ peels away from the MP edge and closely follows the BBP outlier curve, confirming the predicted spike-bulk phase transition in the optimizer's preconditioner. This justifies the use of BBP theory to characterize spike detectability in $\hat{G}$.

5.3.2. *Eigenvector detectability.*

To quantify detectability at the level of eigenspaces, we measure the squared overlap $|\langle v_{\max}, u\rangle|^2$ between the top eigenvector of $\hat{G}$ and the true spike direction u. Figure 6 plots the empirical mean and standard deviation of this overlap as a function of β, with the same BBP threshold β_c.

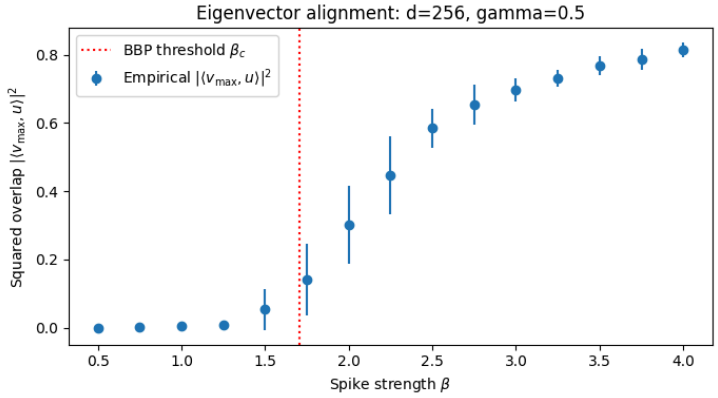

FIGURE 6. Squared overlap between the leading eigenvector of $\hat{G}$ and the true spike direction u as a function of spike strength β.

Empirical observation. Below the threshold β_c, the overlap remains essentially zero: the leading eigenvector of $\hat{G}$ behaves like a generic bulk direction and does not correlate with u. Once $\beta > \beta_c$, the overlap rapidly increases and stabilizes at a nonzero value, indicating that the spike direction becomes **detectable** in the empirical preconditioner. The increased variance near β_c is consistent with finite-size BBP fluctuations.

5.3.3. *Effect of spike detectability on preconditioned curvature.*

Finally, we study how detectability affects the preconditioned curvature. For each β, we evaluate the Rayleigh quotient of H' along the data-driven spike direction $v_{\max}\left(\hat{G}\right)$,

$$v_{\max}^T H' v_{\max} = v_{\max}^T \hat{G}^{-\frac{1}{2}} T \hat{G}^{-\frac{1}{2}} v_{\max}.$$

Figure 7 compares the empirical mean of this quantity against the theoretical shrinkage factor $\beta/\lambda_{\text{out}(\beta,\gamma)}$ predicted by Theorem 4.2.1 in the supercritical regime.

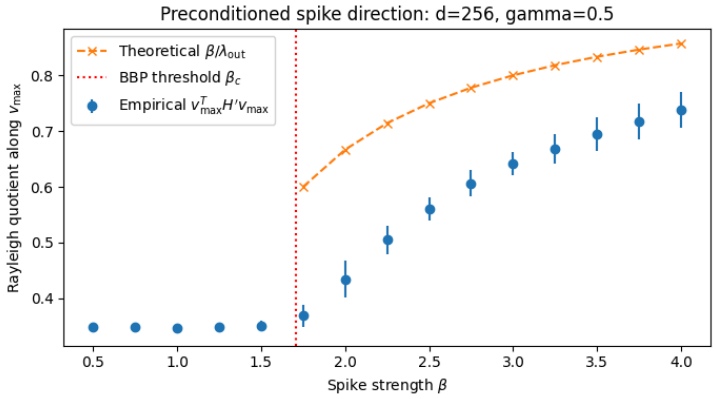

FIGURE 7. Rayleigh quotient of the preconditioned curvature H' along the leading eigenvector $v_{\max}(\hat{G})$ as a function of spike strength β, together with the theoretical shrinkage factor $\beta/\lambda_{\text{out}(\beta,\gamma)}$.

Empirical observation. For subcritical spikes $\beta < \beta_c$, the Rayleigh quotient along $v_{\max}$ remains nearly constant: since the spike is not resolved by $\hat{G}$, whitening cannot target it specifically, and the spike contributes curvature comparable to the bulk. Once $\beta > \beta_c$, the Rayleigh quotient decreases and tracks the theoretical curve $\beta/\lambda_{\text{out}(\beta,\gamma)}$, showing that the detected spike direction is strongly attenuated by $\hat{G}^{-\frac{1}{2}}$. In this regime the residual largest curvature of H' is dominated not by the spike but by the Marchenko-Pastur bulk floor of $\hat{G}$, so the effective conditioning of the preconditioned problem is controlled primarily by the aspect ratio γ rather than the raw spike amplitude β.

This confirms the intuition that strong curvature directions can be effectively neutralized by the empirical preconditioner once they cross the BBP detectability threshold, while weaker spikes remain hidden inside the bulk and survive as residual outliers in H'. Combined with Theorem 4.1.1, this shows that the overall conditioning of H' is well-controlled provided the effective aspect ratio γ is not too large, regardless of the presence of spikes in T. Indeed, γ is just a function of the preconditioner sample size m and the decay parameter α, which can be tuned independently of the model curvature.

5.4. **Empirical validation of Theorem 4.3.1.1 and Theorem 4.3.2.1: eigenbasis staleness.** We now turn to the third question: how much does a **stale** eigenbasis hurt the spectrum of the preconditioned curvature? In SOAP, the eigendecomposition of $\hat{G}$ is recomputed only every τ steps; intermediate updates use the last computed eigenbasis. Theorem 4.3.1.1 and Theorem 4.3.2.1 predict that, as long as the drift $\|\hat{G}_{t+\tau} - \hat{G}_t\|_2$ is small compared to the eigengap of the outlier block and the minimal eigenvalue $\alpha := \lambda_{\min}(\hat{G}_t)$, both the outlier subspace and entire spectrum of H' remain stable.

To test these predictions, we simulate an EMA preconditioner in the same spiked setting as in the previous subsection. We fix $d = 256$, $\gamma = d/m = 0.5$, and a rank-one spike $T = I + (\beta - 1)uu^T$ with $\beta > 1$ and $\|u\|_2 = 1$. At each time step we draw $g_t \sim \mathcal{N}(0, T)$ and update

$$\hat{G}_t = (1 - \eta)\hat{G}_{t-1} + \eta g_t g_t^T$$

for a small step size η (so $\hat{G}_t$ drifts slowly). After a burn-in period we sample starting times t_0 and, for several lags $\tau \in \{1, 2, 4, 8, 16\}$, compare the **fresh** preconditioner based on $\hat{G}_{t_0+\tau}$ with the **stale** one that still uses $\hat{G}_{t_0}$. For each (t_0, τ), we form

$$H'_{\text{fresh}} = \hat{G}_{t_0+\tau}^{-\frac{1}{2}} T \hat{G}_{t_0+\tau}^{-\frac{1}{2}}, \quad H'_{\text{stale}} = \hat{G}_{t_0}^{-\frac{1}{2}} T \hat{G}_{t_0}^{-\frac{1}{2}},$$

and record the following quantities:

i) the subspace distance $\|\sin\Theta(U_t, U_{t+\tau})\|_2$ between top-eigenvector subspaces of $\hat{G}_t$ and $\hat{G}_{t+\tau}$,
ii) the operator norm $\|H'_{\text{stale}} - H'_{\text{fresh}}\|_2$ and its ratio to the theoretical bound from Theorem 4.3.2.1,
iii) the spectral edge changes $|\Delta\lambda_{\max}|, |\Delta\lambda_{\min}|$ between H'_{fresh} and H'_{stale}.

We average these quantities over many independent trajectories of the EMA process.

5.4.1. *Outlier subspace drift vs. refresh lag.*

Figure 8 plots the empirical mean and standard deviation of the outlier subspace distance $\|\sin\Theta(U_t, U_{t+\tau})\|_2$ as a function of the lag τ. We also report the empirical mean and minimum eigengap between the top two eigenvalues of $\hat{G}_t$ over the sampled t_0, which are both comfortably bounded away from zero.

Empirical observation. The outlier subspace drift grows approximately linearly with τ but remains small in absolute value: even at $\tau = 16$, we observe $\mathbb{E}[\|\sin\Theta(U_t, U_{t+\tau})\|_2] \approx 0.16$, corresponding to a principal angle of only a few degrees. This is consistent with the Davis–Kahan bound in Theorem 4.3.1.1: the drift scales like $\|\hat{G}_{t+\tau} - \hat{G}_t\|_2$ divided by the eigengap, and our simulations confirm that a sizeable gap (here, mean gap ≈ 1.2 and minimum ≈ 0.78) keeps the spike subspace extremely stable even for moderate lags. In particular, the strongest curvature directions seen by the preconditioner change only slowly over time, justifying SOAP's use of infrequent eigenbasis refreshes.

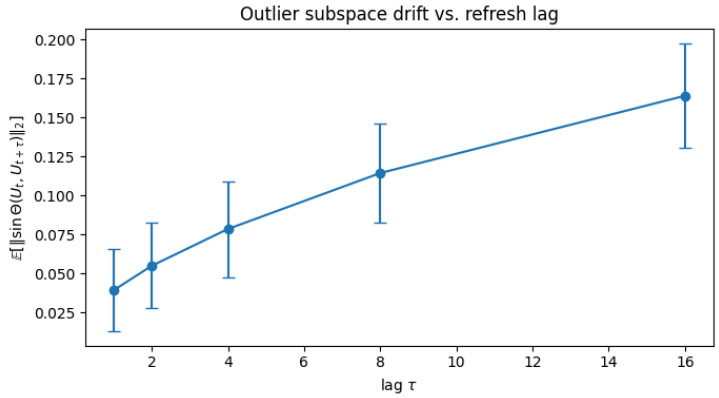

FIGURE 8. Outlier subspace drift as a function of refresh lag τ. The y-axis shows the empirical mean and standard deviation of $\|\sin\Theta(U_t, U_{t+\tau})\|_2$ between the leading eigenvectors of $\hat{G}_t$ and $\hat{G}_{t+\tau}$.

5.4.2. *Tightness of the spectral stability bound.*

Next we assess the Lipschitz-type bound of Theorem 4.3.2.1. For each (t_0, τ), we compute ratio

$$R_{t_0,\tau} = \|H'_{\text{stale}} - H'_{\text{fresh}}\|_2 \,/\, \big[(1/\alpha^2)\, \|H\|_2\, \|\hat{G}_{t_0+\tau} - \hat{G}_{t_0}\|_2\big],$$

where $\alpha = \lambda_{\min}\big(\hat{G}_{t_0}\big)$. By construction, Theorem 4.3.2.1 guarantees $R_{t_0,\tau} \le 1$ in the worst case; we examine how close typical random instances come to saturating this bound.

Figure 9 shows the histogram of $R_{t_0,\tau}$ over all sampled (t_0, τ) pairs and random trajectories.

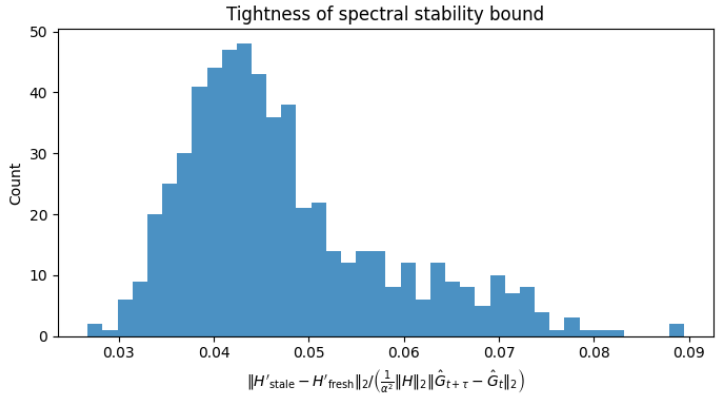

FIGURE 9. Histogram of the ratio between the empirical spectral change $\|H'_{\text{stale}} - H'_{\text{fresh}}\|_2$ and the theoretical upper bound from Theorem 4.3.2.1.

Empirical observation. The ratios concentrate well below 1, typically in the range 0.03–0.08. Thus the true spectral drift is only a few percent of the worst-case Lipschitz bound, mirroring the pessimism of the MP-based condition-number bounds in Theorem 4.1.1. This is expected: the bound assumes adversarial alignment between the perturbation $\hat{G}_{t+\tau} - \hat{G}_t$ and the extremal eigenspaces of H', whereas in our random model the perturbations are high dimensional and the relevant eigenvectors are in generic relative position. In practice, this means that stale preconditioning produces much smaller changes in H' than the worst-case theory allows.

5.4.3. *Spectral edge inflation vs. refresh lag.*

Finally, we directly track how staleness affects the spectral edges of the preconditioned curvature. For each (t_0, τ), we measure

$$\Delta\lambda_{\max} = \lambda_{\max}(H'_{\text{stale}}) - \lambda_{\max}(H'_{\text{fresh}}), \quad \Delta\lambda_{\min} = \lambda_{\min}(H'_{\text{stale}}) - \lambda_{\min}(H'_{\text{fresh}}).$$

Figure 10 plots the empirical mean and standard deviation of the absolute changes $|\Delta\lambda_{\max}|$ and $|\Delta\lambda_{\min}|$ as a function of τ.

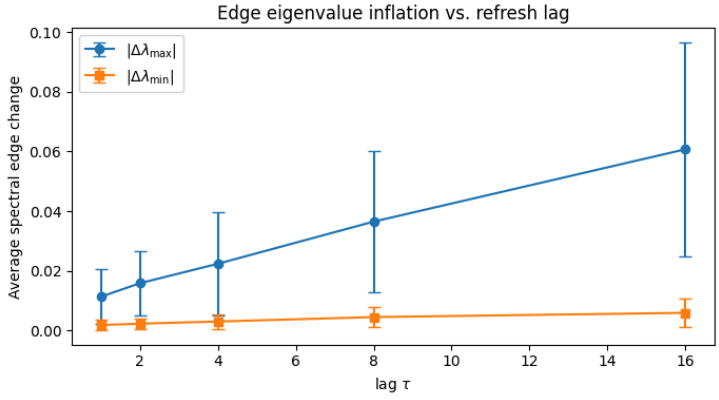

FIGURE 10. Average spectral edge changes $|\Delta\lambda_{\max}|$ and $|\Delta\lambda_{\min}|$ between fresh and stale preconditioned curvatures as a function of the refresh lag τ.

Empirical observation. The upper edge $|\Delta\lambda_{\max}|$ grows roughly linearly with τ but remains small in absolute terms (on the order of 0.01–0.06 in these units), while the lower edge $|\Delta\lambda_{\min}|$ is nearly flat and close to zero across all lags. This asymmetry matches the theoretical picture: the largest eigenvalue is influenced by the spike outlier and the small drift of its associated eigenspace, while

the smallest eigenvalue lies deep in the Marchenko–Pastur bulk and is effectively averaged out by noise. From an optimization standpoint, this means that stale updates primarily affect the step-size–controlling direction, and even there the inflation is modest for realistic refresh intervals.

Taken together, these experiments support Theorem 4.3.1.1 and Theorem 4.3.2.1: in the Wishart/EMA model, both the spike subspace and the spectral edges of H' remain remarkably stable under moderate staleness. Combined with the MP and BBP analyses in Theorem 4.1.1 and Theorem 4.2.1, this suggests that the dominant factor controlling the conditioning of H' in practice is the effective aspect ratio γ of the preconditioner, while eigenbasis staleness plays a secondary, well-controlled role.

6. Conclusion and Future Directions

At a high level, our three main results tell a consistent story. First, Theorem 4.1.1 shows that, in a Wishart model, the spectrum and condition number of the preconditioned curvature $H' = \hat{G}^{-\frac{1}{2}} H \hat{G}^{-\frac{1}{2}}$ are governed mainly by the effective aspect ratio $\gamma = d/m$: as long as γ is modest, H' is well-conditioned with high probability. Second, Theorem 4.2.1 shows that spikes in H are either **subcritical** and behave like bulk directions, or supercritical and become detectable outliers that are actively whitened by $\hat{G}^{-\frac{1}{2}}$. Third,Theorem 4.3.1.1 and Theorem 4.3.2.1 show that the spike subspace and all eigenvalues of H' change at most linearly with the drift $\|\hat{G}_{t+\tau} - \hat{G}_t\|_2$, so a stale eigenbasis only induces mild spectral perturbations.

Putting these pieces together, the picture for SOAP/Shampoo is reassuringly simple: spectrally, these optimizers are quite stable. The dominant knob is the EMA effective sample size m (or equivalently the decay parameter), which sets γ and thereby the Marchenko-Pastur bulk spread. Once m is large enough that γ is not too close to 1, the bulk of H' is well-conditioned and the stable step size $\eta < 2/\lambda_{\max}(H')$ is primarily controlled by γ, not by pathological curvature. Strong curvature directions do not destabilize the method: if they are weak, they blend into the bulk and are tamed by the MP bounds; if they are strong, BBP detectability ensures that $\hat{G}$ resolves and whitens them, so the spike does not dominate $\lambda_{\max}(H')$. Finally, the Davis-Kahan and Lipschitz bounds, together with our experiments, indicate that using an eigenbasis refreshed only every τ steps is spectrally safe: the spike subspace drifts slowly, and the spectral edges of H' move only slightly even for moderate τ. In practice, this means that SOAP's "infrequent eigendecompositions" is both computationally motivated and spectrally justified: as long as $\hat{G}$ remains well-conditioned and its drift per step is small, the preconditioner behaves almost as if it were recomputed at every update.

6.1. Future directions.

Looking ahead, several directions seem particularly promising:

- Extend the analysis beyond Gaussian/Wishart gradients to heavy-tailed or more realistic deep-learning covariance models.
- Incorporate the per-mode Kronecker structure of real Shampoo/SOAP and study how MP/BBP behavior composes across modes in actual networks.
- Design adaptive schemes that tune the EMA decay and refresh frequency τ based on observed spectral diagnostics (e.g. aspect ratio, gaps, and drift).

References

1. Gupta, V., Koren, T., Singer, Y.: Shampoo: Preconditioned Stochastic Tensor Optimization. In: Dy, J. and Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. pp. 1842–1850. PMLR (2018)
2. Vyas, N., Morwani, D., Zhao, R., Shapira, I., Brandfonbrener, D., Janson, L., Kakade, S.: SOAP: Improving and Stabilizing Shampoo using Adam. Presented at the (2024)
3. Sato, N., Naganuma, H., Iiduka, H.: Convergence Bound and Critical Batch Size of Muon Optimizer, https://arxiv.org/abs/2507.01598
4. Marchenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik. 1, 457–483 (1967). https://doi.org/10.1070/SM1967v001n04ABEH001994
5. Baik, J., Ben Arous, G., Péché, S.: Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. The Annals of Probability. 33, 1643–1697 (2005). https://doi.org/10.1214/009117905000000233
6. Davis, C., Kahan, W.M.: The rotation of eigenvectors by a perturbation. III. SIAM Journal on Numerical Analysis. 7, 1–46 (1970). https://doi.org/10.1137/0707001

Massachusetts Institute of Technology
Email address: htfan@mit.edu