A Spectral Bound on Effective Sharpness for Fisher-Preconditioned Gradient Descent

Anonymous authors
Paper under double-blind review

Abstract

Neural networks trained with gradient descent exhibit the Edge of Stability (EoS) phenomenon, where the sharpness rises to $2/\eta$ and the loss oscillates non-monotonically—but does this instability persist under curvature-aware (second-order) optimization? This paper analyzes Fisher-preconditioned gradient descent through the lens of effective sharpness, defined as the largest eigenvalue of the preconditioned Hessian $S_{\text{eff}} = \lambda_{\max}(F^{-1}H)$, where $H = \nabla^2 L(\theta)$ is the Hessian and F is the Fisher Information Matrix. The central finding is that effective sharpness is bounded by a quantity decomposing into residual curvature and model misspecification, explaining why natural gradient descent (NGD) suppresses EoS oscillations. Specifically, under the idealized assumption that the Gauss-Newton matrix G equals the Fisher F, the bound $S_{\text{eff}} \leq 1 + \epsilon/\mu_{\min}(F)$ holds, where $\epsilon = \|H - G\|_2$ is the residual curvature and $\mu_{\min}(F)$ is the minimum Fisher eigenvalue. In the general case where $G \neq F$, the operationally applicable bound becomes $S_{\text{eff}} \leq 1 + (\epsilon + \delta)/\mu_{\min}(F)$, with $\delta = \|G - F\|_2$ measuring model misspecification. An alignment-aware Rayleigh quotient analysis reveals that the worst-case bound is loose by 1.3–7.1× due to favorable alignment between the residual curvature Q and the Fisher eigenvectors. These bounds are verified on deep linear networks (DLNs) from 55 to 3,240 parameters (5 seeds per configuration), where the general bound holds at all tested scales; on a 110-parameter DLN with exact matrix computations, the idealized bound is confirmed when $G \approx F$ but is violated when misspecification is substantial. K-FAC on CIFAR-10 ResNet-18 (11.2M parameters) achieves 90.5% test accuracy (vs. SGD 86.2%) while operating in 41× sharper curvature regimes, consistent with the spectral flattening mechanism, though direct measurement of S_{eff} at this scale remains intractable.

1 Introduction

The optimization of deep neural networks involves differential geometry, spectral theory, and matrix analysis. While first-order methods have achieved broad empirical success, a complete theoretical account of their convergence behavior remains open. Recent work has identified the Edge of Stability (EoS) as a pervasive phenomenon: during full-batch gradient descent, the sharpness $\lambda_{\max}(H)$ rises to $2/\eta$ and the loss exhibits non-monotonic oscillations rather than diverging (Cohen et al., 2021).

Throughout, $H(\theta) = \nabla^2 L(\theta)$ denotes the Hessian of the training loss and $F(\theta) = \mathbb{E}_{p_{\text{data}}}[\nabla \log p(y|x;\theta) \nabla \log p(y|x;\theta)^T]$ denotes the Fisher Information Matrix.

Phenomenon I.1 (Edge of Stability, Cohen et al. (2021)). During full-batch gradient descent, $\lambda_{\max}(H(\theta_t))$ rises monotonically during early training and saturates near $2/\eta$. After saturation, the loss sequence $\{L(\theta_t)\}$ becomes non-monotonic—it decreases on average but exhibits per-iteration increases—yet training does not diverge. Cohen et al. (2021) observe that sharpness continues to oscillate around $2/\eta$ rather than remaining constant after saturation; our experiments (Figure 2(b)) show the saturation behavior characteristic of EoS.

This paper investigates whether the EoS is specific to the Euclidean geometry of standard gradient descent or whether it persists under Riemannian preconditioning. An explicit stability characterization of effective

sharpness—the largest eigenvalue of the preconditioned Hessian $\lambda_{\max}(F^{-1}H)$—under Fisher preconditioning is provided, decomposing stability into residual curvature and model misspecification components. Under the idealized assumption that the Gauss-Newton matrix G (the curvature component arising from the model's output Jacobian) equals the Fisher F (the metric tensor of the statistical manifold)—a condition that holds for correctly specified negative log-likelihood losses—Theorem IV.2 shows that the effective sharpness is bounded by 1 plus the ratio of residual curvature to the minimum Fisher eigenvalue. Since this assumption is violated in all practical settings (Section 3.4), the operationally applicable result is Corollary IV.4, which shows that the effective sharpness is bounded by 1 plus the combined ratio of residual curvature and model misspecification to the minimum Fisher eigenvalue, thereby separating the two sources of curvature error. Direct measurement on a 110-parameter DLN (Section 5.4) confirms that the Theorem IV.2 bound is violated when $G \neq F$, while the Corollary IV.4 bound correctly holds at all iterations. The paper includes experiments from 55 to 11.2 million parameters; scaling to CIFAR-10 ResNet-18 with tuned K-FAC achieves 90.5% test accuracy (vs. SGD 86.2%), with per-epoch sharpness measurements consistent with the spectral flattening mechanism, though direct measurement of S_{eff} at this scale remains intractable.

The analysis is organized around two core perspectives:

1. **Spectral Theory:** Convergence behavior is analyzed via eigenvalues of the Hessian and the preconditioned Hessian (Nocedal and Wright, 2006; Sagun et al., 2017).
2. **Differential Geometry:** The parameter space is treated as a Riemannian manifold with the Fisher Information Matrix as metric tensor (Amari, 1998).

This paper makes the following contributions:

- An explicit stability characterization of effective sharpness under Fisher preconditioning in the general near-realizable setting (Corollary IV.4), providing the bound $S_{\text{eff}} \leq 1 + (\epsilon + \delta)/\mu_{\min}(F)$ that separates residual curvature $\epsilon = \|H - G\|_2$ and model misspecification $\delta = \|G - F\|_2$. This is the operationally applicable result, since $G \neq F$ in all practical settings (Section 3.4).
- An idealized bound under correct model specification $G = F$ (Theorem IV.2), showing $S_{\text{eff}} \leq 1+\epsilon/\mu_{\min}(F)$, verified only when the $G \approx F$ condition approximately holds, which provides intuition for the spectral flattening mechanism and reduces to the identity ($S_{\text{eff}} = 1$) under exact realizability.
- An alignment-aware spectral analysis (Theorem IV.3) providing Rayleigh quotient characterization that explains the 1.3–7.1× looseness of the worst-case bound.
- A matching lower bound (Proposition IV.6) showing tightness is controlled by $\kappa(F)$.
- A mechanistic explanation for why NGD suppresses EoS oscillations via spectral flattening (Section 4.4), contingent on $G \approx F$.
- Direct measurement of model misspecification $\delta = \|G - F\|_2$ on the 110-parameter DLN (Section 5.4), confirming that Corollary IV.4 correctly bounds S_{eff} when $G \neq F$.
- Empirical validation spanning 55 to 11.2M parameters, including a CIFAR-10 ResNet-18 demonstration with K-FAC achieving 90.5% test accuracy while operating in 41× sharper curvature regimes than SGD.

Notation. Throughout this paper, H denotes the Hessian $\nabla^2 L(\theta)$ (measuring the curvature of the loss), F denotes the Fisher Information Matrix (the metric tensor of the statistical manifold), G denotes the Gauss-Newton matrix (the curvature component arising from the model's output Jacobian), and $S_{\text{eff}} = \lambda_{\max}(F^{-1}H)$ denotes the effective sharpness (the largest eigenvalue of the preconditioned Hessian). Full notation is summarized in Table II (Section III).

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents mathematical preliminaries, including the Gauss-Newton decomposition. Section 4 derives the spectral stability bounds and provides a mechanistic explanation for EoS suppression. Section 5 presents experimental validation. Section 6 discusses limitations, and Section 7 concludes.

2 Related Work

2.1 Edge of Stability

Cohen et al. (2021) demonstrated the EoS on ResNets trained on CIFAR-10 and ImageNet: $\lambda_{\max}(H)$ consistently rises to $2/\eta$ during full-batch gradient descent, after which the loss oscillates non-monotonically. Lewkowycz et al. (2020) identified the related catapult mechanism in the large learning rate regime. Jastrzebski et al. (2021) showed that progressive sharpening during early training sets the stage for EoS, linking sharpness dynamics to the learning rate and batch size. More recently, Damian et al. (2023) provided a rigorous self-stabilization analysis showing how gradient descent implicitly regularizes sharpness at the EoS, and Arora et al. (2022) established that the EoS is not an anomaly but a provable consequence of gradient descent dynamics near saddle-to-minimum transitions. Ahn et al. (2022) demonstrated that the EoS phenomenon persists in stochastic settings with sufficiently large batches, connecting the full-batch theory to practical mini-batch training. The present work differs by analyzing spectral dynamics under Riemannian (natural gradient) preconditioning and providing a mechanistic explanation for why NGD avoids EoS (Section 4.4).

2.2 Natural Gradient and Fisher Information

Amari (1998) introduced natural gradient descent, which follows the steepest descent direction on the statistical manifold. Martens (2020) provided a modern treatment connecting the Fisher, the Generalized Gauss-Newton matrix, and practical approximations. Kunstner et al. (2019) demonstrated that the commonly used Empirical Fisher can diverge substantially from the true Fisher, particularly in non-realizable settings—a limitation directly relevant to the diagonal approximation used in this paper's experiments. K-FAC (Martens and Grosse, 2015) provides a scalable block-diagonal approximation.

2.3 Second-Order Optimizers

ADAHESSIAN (Yao et al., 2021) uses a diagonal Hessian approximation with spatial averaging. SOPHIA (Liu et al., 2024) employs a lightweight Hessian estimator for language model pre-training. Shampoo (Gupta et al., 2018) preconditions using Kronecker products of gradient statistics. These second-order methods (algorithms that use curvature information, often via Hessian approximations) approximate second-order information at reduced cost but do not directly analyze spectral stability. Recent libraries such as ASDL (Osawa et al., 2023) provide unified interfaces for K-FAC and Shampoo in PyTorch, enabling tractable second-order optimization at moderate scale, though hyperparameter tuning remains challenging. Agarwala and Gur-Ari (Agarwala and Gur-Ari, 2022) studied the curvature dynamics of preconditioned gradient methods, showing that adaptive preconditioning can fundamentally alter the sharpness trajectory. George et al. (2018) introduced EKFAC, improving the Fisher approximation by computing a diagonal variance in the Kronecker-factored eigenbasis. Chen and Bruna (2023) analyzed the loss surface of neural networks through the lens of preconditioned Hessian spectra, connecting effective sharpness to trainability.

Table I: Comparison with Prior Second-Order Methods

Method	Scale	Spectral Analysis?	Sharpness Measured?
K-FAC (Martens and Grosse, 2015)	Large	No	No
Shampoo (Gupta et al., 2018)	Large	No	No
ADAHESSIAN (Yao et al., 2021)	Large	No	No
SOPHIA (Liu et al., 2024)	Large	No	No
This work	11.2M params	Yes (Thm IV.2–3)	Yes ($\lambda_{\max}(H)$)

Note: 11.2M parameters is moderate-scale relative to modern large language models (10^8–10^{12} parameters) but represents the practical upper bound for this analysis—direct measurement of S_{eff} at this scale is already intractable. We position the ResNet-18 experiment as empirical motivation rather than theoretical verification.

2.4 Deep Linear Networks

Saxe et al. (2014) derived exact solutions to the learning dynamics in deep linear networks, establishing them as tractable models for studying gradient descent. Bernacchia et al. (2018) extended this analysis to natural gradient in the linear case. Arora et al. (2019) proved that gradient descent on deep matrix factorizations exhibits implicit regularization toward low-rank solutions.

2.5 Implicit Bias and Generalization

Gunasekar et al. (2018) showed that optimization geometry determines implicit regularization in matrix factorization, with different optimizers (GD, mirror descent, NGD) converging to different solutions. This is relevant to our analysis: NGD's spectral flattening (Section 4.4) may induce different implicit biases than GD, though we do not investigate this connection empirically. The relationship between sharpness and generalization (Jiang et al., 2020) suggests that NGD's bounded effective sharpness could have generalization implications, but this is beyond the current scope.

3 Preliminaries

Table II summarizes the notation used throughout this paper.

Table II: Notation Reference

Symbol	Description
$\theta \in \mathbb{R}^d$	Model parameters
$F(\theta)$	Fisher Information Matrix
$H(\theta)$	Hessian $\nabla^2 L(\theta)$
G	Gauss-Newton (Fisher) component of H
$Q = H - G$	Residual curvature
$\epsilon = \|Q\|_2 = \|H - G\|_2$	Spectral norm of residual curvature
$\mu_{\min}(F)$	Minimum eigenvalue of F (undamped)
$\mu_{\min}(F + \gamma I)$	Minimum eigenvalue of damped Fisher ($= \mu_{\min}(F) + \gamma$, for scalar damping γI)
$\mu_{\max}(F)$	Maximum eigenvalue of F (Fisher spectral norm: $\mu_{\max}(F) = \|F\|_2$)
S_{eff}	Effective sharpness $\lambda_{\max}(F^{-1}H)$
η, γ	Learning rate, damping coefficient
$\delta = \|G - F\|_2$	Model misspecification measure
$\kappa(F)$	Fisher condition number $\mu_{\max}(F)/\mu_{\min}(F)$

3.1 The Statistical Manifold and Fisher Metric

A neural network parameterized by $\theta \in \mathbb{R}^d$ defines a conditional distribution $p(y|x;\theta)$. The family $\{p(y|x;\theta)\}$ is treated as a Riemannian manifold $\mathcal{M}$ equipped with the Fisher Information Matrix as metric tensor (Amari, 1998). The Fisher captures the local curvature of the KL divergence: $\mathrm{KL}[p_\theta \| p_{\theta+d\theta}] = \frac{1}{2} d\theta^T F(\theta)\, d\theta + O(\|d\theta\|^3)$, where $F(\theta) = \mathbb{E}_{p_{\text{data}}}\left[\nabla_\theta \log p(y|x;\theta) \nabla_\theta \log p(y|x;\theta)^T\right]$ Amari (1998, Theorem 2). Here, p_{data} denotes the empirical distribution over training pairs (x, y), and $p_{\text{model}} = p(y|x;\theta)$ denotes the model's predictive distribution. For the negative log-likelihood loss $L(\theta) = -\mathbb{E}_{p_{\text{data}}}[\log p(y|x;\theta)]$, the Fisher Information Matrix satisfies $F(\theta) = \mathbb{E}_{p_{\text{model}}}[-\nabla^2 \log p(y|x;\theta)]$. Under a correctly specified model at the optimum ($p_{\text{data}} = p_{\text{model}}$), the Fisher equals the expected Hessian: $F(\theta^*) = \mathbb{E}_{p_{\text{data}}}[H(\theta^*)]$ (Bishop (Bishop, 2006, Section 1.6), Murphy (Murphy, 2012, Chapter 8)). In practice, the Empirical Fisher (computed from observed data) is often substituted; Kunstner et al. (2019) showed this may depart significantly from the true Fisher in non-realizable settings.

3.2 The Gauss-Newton Decomposition

For a loss of the form $L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)$, the Hessian admits the decomposition $H = G + Q$, where

$$G = \frac{1}{N}\sum_{i=1}^{N} J_i^T \nabla_z^2 \ell(z, y_i)\big|_{z=f_\theta(x_i)} J_i \tag{1}$$

is the Gauss-Newton matrix ($J_i = \partial f_\theta(x_i)/\partial\theta$ is the Jacobian of the model output with respect to θ for sample i), and $Q = H - G$ is the residual curvature that vanishes at zero-loss solutions. For negative log-likelihood losses with correctly specified models, G equals the expected outer product of score functions, making it directly comparable to the Fisher F (Martens, 2020, Section 4.3). The spectral norm $\epsilon = \|Q\|_2 = \|H - G\|_2$ measures the magnitude of residual curvature.

3.3 Natural Gradient Descent

The natural gradient of $L(\theta)$ is $\tilde{\nabla} L = (F(\theta) + \gamma I)^{-1}\nabla L(\theta)$, where $\gamma > 0$ provides numerical stability (Martens and Grosse, 2015). The update rule is:

$$\theta_{t+1} = \theta_t - \eta(F(\theta_t) + \gamma I)^{-1}\nabla L(\theta_t) \tag{2}$$

3.4 When Does $G \approx F$?

The Gauss-Newton matrix G equals the Fisher F exactly when two conditions hold simultaneously: (i) the loss is the negative log-likelihood, and (ii) the model is correctly specified, i.e., $p_{\text{data}} \in \{p_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ denotes the parameter space of the model class (Martens, 2020, Section 4.3). In practice, both conditions are violated to varying degrees. Finite model capacity ensures $G \neq F$ in general, and the departure can be substantial for highly misspecified models (Kunstner et al., 2019). The spectral norm $\delta = \|G - F\|_2$ provides a scalar measure of this departure. For better interpretability, a relative measure $\delta/\|F\|_2 = \delta/\mu_{\max}(F)$ can be reported alongside the absolute δ; we provide both in Section 5.4. Empirically, δ tends to be large early in training and may or may not decrease as the model approaches a local minimum, depending on the degree of misspecification. Our DLN experiments (Section 5.4) measure δ directly, finding values of 1.16–1.82 throughout training. In practice, $G \approx F$ is best approximated in early training of well-regularized networks with large model capacity relative to data complexity, where the model residuals are small and the output distribution closely tracks the data distribution. As training progresses toward a local minimum, $\epsilon = \|H - G\|_2$ decreases but $\delta = \|G - F\|_2$ may persist due to structural misspecification. This motivates treating the $G = F$ case as an idealized setting that provides intuition, with the general $G \neq F$ case (Corollary IV.4) as the operationally applicable bound.

4 Spectral Theory and Stability Analysis

4.1 Stability of Gradient Descent

Before analyzing NGD, we recall the classical stability condition for vanilla gradient descent. This provides the baseline against which the Fisher-preconditioned case is compared.

An iteration $\theta_{t+1} = f(\theta_t)$, where $f(\theta) = \theta - \eta\nabla L(\theta)$ is the GD update map, is *locally stable* near a fixed point θ^* if small perturbations do not grow, i.e., the spectral radius of the Jacobian of f at θ^* is less than 1.

Lemma IV.1 (GD Stability Threshold; Nocedal and Wright (Nocedal and Wright, 2006), Cohen et al. (2021)) For a twice continuously differentiable loss $L(\theta)$ near a stationary point θ^*, a sufficient condition for local stability of gradient descent with learning rate η is $\eta < 2/\lambda_{\max}(H(\theta^*))$.

Proof sketch: The linearized GD iteration near θ^* is $\theta_{t+1} - \theta^* \approx (I - \eta H)(\theta_t - \theta^*)$. Stability requires the spectral radius of $I - \eta H$ to be less than 1. For H positive semidefinite, this reduces to $\max_i |1 - \eta\lambda_i(H)| < 1$, which gives $\eta < 2/\lambda_{\max}(H)$ as a sufficient condition for local stability. The condition is not necessary in

general—stability can hold for larger η when the loss landscape has favorable structure—but it is tight for quadratic losses and provides the relevant threshold for EoS analysis. ■

This is a sufficient condition for local stability (not sufficient for global convergence). When $\lambda_{\max}$ exceeds $2/\eta$ during training, GD enters the Edge of Stability regime (Phenomenon I.1).

4.2 Effective Sharpness

To analyze NGD stability, we need the correct analog of the classical sharpness $\lambda_{\max}(H)$. Linearizing the NGD update near a stationary point θ^* reveals that the relevant quantity is not the raw Hessian sharpness but the effective sharpness under the Fisher metric.

Definition IV.1 (Effective Sharpness) The effective sharpness under NGD is:

$$S_{\text{eff}}(\theta) := \lambda_{\max}\left(F(\theta)^{-1}H(\theta)\right) \tag{3}$$

Remark (Justification): To justify S_{eff} as the correct stability measure, consider the linearized NGD iteration near a stationary point θ^*: $\theta_{t+1} - \theta^* \approx (I - \eta F^{-1}H)(\theta_t - \theta^*)$. The iteration is locally stable iff the spectral radius of $I - \eta F^{-1}H$ is < 1. Since F is positive definite, $F^{-1}H$ is similar to the symmetric matrix $F^{-1/2}HF^{-1/2}$ (via the similarity $P = F^{1/2}$), and therefore shares its eigenvalues. Near a local minimum where H is positive semidefinite, all eigenvalues of $F^{-1}H$ are non-negative, and the stability condition reduces to $\eta \cdot \lambda_{\max}(F^{-1}H) < 2$, i.e., $\eta < 2/S_{\text{eff}}$. Alternative measures such as the Frobenius norm $\|F^{-1}H\|_F$ would overcount contributions from small eigenvalues that do not affect stability.

4.3 NGD Stability Bound

We now derive the central result. Under the assumption that the model is correctly specified ($G = F$), Fisher preconditioning absorbs the dominant curvature into the identity matrix, leaving only the residual $Q = H - G$ as a perturbation. The effective sharpness is therefore controlled entirely by how badly Q is amplified by the inverse Fisher, which is worst along the least-conditioned Fisher directions.

The central quantity controlling NGD stability is the effective sharpness $S_{\text{eff}} = \lambda_{\max}(F^{-1}H)$. By the NGD linearization in Definition IV.1, the stability condition is $\eta < 2/S_{\text{eff}}$. Theorem IV.2 shows S_{eff} is bounded independently of $\lambda_{\max}(H)$.

The following theorem establishes a pointwise bound on effective sharpness at any fixed parameter value θ. The assumptions are local: they hold at or near the point θ being analyzed, not globally over the loss surface. The connection to stability ($\eta < 2/S_{\text{eff}}$) follows from applying Lemma IV.1 to the linearized NGD update, as described in Definition IV.1.

Theorem IV.2 (NGD Stability Bound) Let $\theta \in \mathbb{R}^d$ be any parameter vector, and let $H = H(\theta) = \nabla^2 L(\theta)$, $F = F(\theta)$ denote the Hessian and Fisher at θ. Under the Generalized Gauss-Newton decomposition $H = G + Q$, where G is the Gauss-Newton matrix and Q is the residual curvature, assume:

1. $F(\theta) + \gamma I$ is positive definite with minimum eigenvalue $\mu_{\min} > 0$.
2. $\|Q(\theta)\|_2 \leq \epsilon$ for some $\epsilon \geq 0$.
3. The model's loss is the negative log-likelihood, and under the correctly specified model assumption, $G = F$ (Martens, 2020).

Then:

$$S_{\text{eff}}(\theta) = \lambda_{\max}(F^{-1}H) \leq 1 + \frac{\epsilon}{\mu_{\min}(F)} \tag{4}$$

When $\epsilon = 0$ (exact realizability), $S_{\text{eff}} = 1$ and NGD is unconditionally stable for any $\eta < 2$.

Proof sketch. Fisher preconditioning reduces $F^{-1}H$ to $I+F^{-1}Q$ via the GGN decomposition with $G = F$. A congruence transformation $F^{-1/2}HF^{-1/2} = I+F^{-1/2}QF^{-1/2}$ symmetrizes this product, after which Weyl's inequality and submultiplicativity of the spectral norm yield the bound. Full proof in Appendix C.1. ■

Intuition. The key insight is that Fisher preconditioning absorbs the dominant curvature into the identity, leaving only the residual Q as a perturbation. Under the GGN decomposition $H = G + Q$ with $G = F$, preconditioning yields $F^{-1}H = I + F^{-1}Q$—exactly the identity plus a perturbation. To handle the non-symmetry of $F^{-1}H$, a congruence transformation $F^{-1/2}HF^{-1/2} = I + F^{-1/2}QF^{-1/2}$ reduces the problem to bounding the largest eigenvalue of a symmetric matrix. The denominator $\mu_{\min}(F)$ appears because the smallest Fisher eigenvalue controls the worst-case amplification of Q under the congruence: directions with small Fisher eigenvalues are the most vulnerable to perturbation by Q, yielding the bound $S_{\text{eff}} \leq 1 + \|Q\|_2/\mu_{\min}(F)$. This is why spectral flattening works—directions of large Hessian curvature typically co-occur with large Fisher eigenvalues, so the effective curvature $\lambda_i(F^{-1}H)$ stays bounded even as $\lambda_i(H)$ grows.

Discussion. Three aspects of Theorem IV.2 warrant discussion:

(i) On the $G = F$ assumption. The identity $G = F$ holds for negative log-likelihood losses under the correctly specified model assumption (Martens, 2020, Section 4.3). In practice, two factors cause this to fail: (i) model misspecification (finite capacity), and (ii) non-zero training residuals. **This assumption is violated in all our experiments to varying degrees.** For the DLN teacher-student setup, $G \approx F$ holds approximately during training and improves as the loss decreases, but is not exact at finite loss. For nonlinear networks trained on real data (MNIST, CIFAR-10), $G \neq F$ in general, and the departure can be significant. Corollary IV.4 is the applicable bound in these settings, with the misspecification parameter $\delta = \|G-F\|_2$ quantifying the deviation. In the DLN experiments where we compute all quantities exactly (Section 5.4), we verify that Corollary IV.4 correctly bounds S_{eff} at all measured iterations. Theorem IV.2 is confirmed to hold only when $G \approx F$ approximately, and is violated at iterations 50 and 100 where $\delta = \|G - F\|_2$ is substantial relative to $\epsilon_{\text{true}} = \|H-G\|_2$, confirming that the $G = F$ assumption is materially violated in this experiment; measuring δ directly in all experiments is an important direction for future work (Section 7). Our experiments use the damped Fisher $(F + \gamma I)^{-1}$ to ensure positive definiteness regardless of the spectrum of F itself.

(ii) On damping. Adding damping γI replaces F by $F+\gamma I$. The bound becomes $S_{\text{eff}} \leq 1+\epsilon/(\mu_{\min}(F)+\gamma)$, and $\gamma > 0$ tightens it. Section 5.7 shows that increasing γ from 10^{-4} to 10^{-1} reduces SP-GD final loss from 0.49 to $< 10^{-4}$, consistent with this analysis. **Notation convention:** Throughout this paper, $\mu_{\min}(F)$ denotes the minimum eigenvalue of the undamped Fisher. When damping is applied, we write $\mu_{\min}(F+\gamma I) = \mu_{\min}(F) + \gamma$ explicitly. The theoretical bounds (Theorems IV.2–3, Corollary IV.4) are stated in terms of $\mu_{\min}(F)$; in the experiments, the damped Fisher $(F + \gamma I)^{-1}$ is used, and the reported bound values use $\mu_{\min}(F + \gamma I)$.

(iii) On interpreting the bound numerically. The worst-case bound values in the DLN experiments ($\epsilon/\mu_{\min} \approx$ 1,778–2,483, Table V) are numerically large, implying a recommended learning rate of $\eta < 2/S_{\text{eff}} \approx 0.0008$—far below the practically effective $\eta = 0.1$. This bound should be interpreted as a *structural existence result* rather than a practical hyperparameter guide: it establishes that NGD's effective sharpness is controlled by the ratio $\epsilon/\mu_{\min}(F)$ rather than by $\lambda_{\max}(H)$ directly, providing a mechanistic explanation for spectral flattening. The operational tightening comes from Theorem IV.3's alignment-aware analysis: the Rayleigh quotient lower bound (Table VII) is only 1.3× loose relative to the actual S_{eff}, indicating that the alignment structure—not the worst-case bound itself—is the correct predictor of stability in practice.

Theorem IV.3 (Alignment-Aware Spectral Analysis) Let the residual Q have spectral decomposition $Q = \sum_{i=1}^{d} \lambda_i^Q u_i u_i^T$, and let F have eigenpairs (μ_i, v_i). Define the alignment matrix $A_{ij} = (u_i^T v_j)^2$ and the projection coefficients $c_{ij} = u_i^T v_j$. Then the effective sharpness satisfies the following Rayleigh quotient lower bound:

$$S_{\text{eff}} \geq 1 + \max_j \sum_{i=1}^{d} \frac{\lambda_i^Q A_{ij}}{\mu_j} \tag{5}$$

which follows because v_j are unit vectors in the eigenbasis of $F^{-1/2}QF^{-1/2}$, and $\lambda_{\max}(M) \geq v_j^T M v_j$ for any symmetric M. When Q is rank-k with eigenvectors $\{u_i\}_{i=1}^k$ expressible as $u_i = \sum_j c_{ij} v_j$ in the F-eigenbasis,

the triangle inequality for the spectral norm gives the valid upper bound:

$$S_{\text{eff}} \leq 1 + \sum_{i=1}^{k} |\lambda_i^Q| \sum_j \frac{c_{ij}^2}{\mu_j} \tag{6}$$

The gap between these bounds and the worst-case Theorem IV.2 bound is governed by the alignment structure: when Q's spectral mass concentrates along well-conditioned Fisher directions (large μ_j), the actual S_{eff} is much smaller than $1 + \epsilon/\mu_{\min}(F)$.

Proof sketch. Expand Q's eigenvectors in the Fisher eigenbasis via $u_i = \sum_j c_{ij} v_j$. The Rayleigh quotient of $F^{-1/2}QF^{-1/2}$ evaluated at each Fisher eigenvector v_j gives the lower bound; the triangle inequality applied to the rank-1 decomposition $\sum_i \lambda_i^Q w_i w_i^T$ (where $w_i = F^{-1/2}u_i$) gives the upper bound. Full proof in Appendix C.2. ■

Remark: Theorem IV.3 explains the observed 1.3–7.1× gap between S_{eff} and the Theorem IV.2 bound in Table V. The Rayleigh quotient lower bound captures the alignment structure: in our experiments, Q's spectral mass concentrates along directions of moderate Fisher eigenvalues (not $\mu_{\min}$), so the lower bound tracks S_{eff} much more closely than the worst-case upper bound. The proximity of the Rayleigh lower bound to the actual S_{eff} (Table VII) indicates that alignment is the dominant factor determining the gap.

Remark (Triangle inequality tightness): The upper bound of Theorem IV.3 uses the triangle inequality, which is tight when the transformed eigenvectors $w_i = F^{-1/2}u_i$ are nearly parallel. In practice, Q's eigenvectors tend to be spread across multiple Fisher eigendirections, so the triangle inequality introduces looseness. The lower bound requires only unit-vector evaluation; its closeness to S_{eff} in Table VII (within 1.3× for the 110-parameter model) indicates that a single Fisher eigenvector captures most of the relevant alignment.

Corollary IV.4 (Near-Realizability with Model Mismatch) If the Gauss-Newton matrix G satisfies $\|G - F\|_2 \leq \delta$ (rather than $G = F$ exactly), then $H = G + Q = F + (G - F) + Q$, and:

$$S_{\text{eff}} \leq 1 + \frac{\epsilon + \delta}{\mu_{\min}(F)} \tag{7}$$

where $\delta = \|G - F\|_2$ measures model misspecification and $\epsilon = \|Q\|_2$ measures residual curvature. This separates the two sources of bound looseness: δ is large when the model class is misspecified or the loss is not the negative log-likelihood, while ϵ is large far from a minimizer.

Proof sketch. Write $H = F + (G - F) + Q$, apply the congruence transformation, then use Weyl's inequality and submultiplicativity to bound each term separately: $\|F^{-1/2}(G - F)F^{-1/2}\|_2 \leq \delta/\mu_{\min}(F)$ and $\|F^{-1/2}QF^{-1/2}\|_2 \leq \epsilon/\mu_{\min}(F)$. Full proof in Appendix C.3. ■

Definition IV.5 (Effective Rank (Roy and Vetterli, 2007)). The effective rank of a positive semidefinite matrix A with eigenvalues λ_i is $\text{rank}_{\text{eff}}(A) = \exp(H_{\text{norm}})$, where $H_{\text{norm}} = -\sum_i p_i \log p_i$ is the Shannon entropy of the normalized eigenvalue distribution $p_i = \lambda_i / \sum_j \lambda_j$.

Proposition IV.6 (Converse: Lower Bound on S_{eff}) Under the assumptions of Theorem IV.2, if Q is positive semidefinite (which holds when $H \succeq G$, i.e., the Hessian dominates the GGN) and $\lambda_{\min}(Q) \geq \epsilon_{\min} > 0$, then:

$$S_{\text{eff}} \geq 1 + \frac{\epsilon_{\min}}{\mu_{\max}(F)} \tag{8}$$

where $\mu_{\max}(F)$ is the largest eigenvalue of F (the maximum Fisher eigenvalue). This shows S_{eff} is bounded away from 1 when $Q \neq 0$, and approaches the upper bound (Theorem IV.2) when F is well-conditioned ($\mu_{\max} \approx \mu_{\min}$).

Proof. By the congruence transformation property, $Q \succeq \epsilon_{\min} I$ implies $F^{-1/2}QF^{-1/2} \succeq \epsilon_{\min} F^{-1}$. Therefore $\lambda_{\max}(I + F^{-1/2}QF^{-1/2}) \geq 1 + \epsilon_{\min}\lambda_{\max}(F^{-1}) = 1 + \epsilon_{\min}/\mu_{\min}(F)$. This gives the stronger bound $S_{\text{eff}} \geq 1 + \epsilon_{\min}/\mu_{\min}(F)$; the proposition states the weaker bound using $\mu_{\max}(F)$ because $\mu_{\min}(F)$ can be near zero in ill-conditioned settings, making the stronger bound vacuous, while the $\mu_{\max}$ bound remains meaningful when F is rank-deficient or near-singular. For the stated result using $\mu_{\max}$, we use a logically weaker bound

that avoids dependence on the smallest Fisher eigenvalue, which can make the lower bound vacuous in ill-conditioned settings. Since $F^{-1/2}QF^{-1/2} \succeq \epsilon_{\min}F^{-1} \succeq (\epsilon_{\min}/\mu_{\max}(F))I$, we have that every eigenvalue of $I + F^{-1/2}QF^{-1/2}$ is at least $1 + \epsilon_{\min}/\mu_{\max}(F)$. In particular, $S_{\text{eff}} = \lambda_{\max}(I + F^{-1/2}QF^{-1/2}) \geq 1 + \epsilon_{\min}/\mu_{\max}(F)$. This bound is weaker than $1 + \epsilon_{\min}/\mu_{\min}(F)$ but depends on $\mu_{\max}$ rather than $\mu_{\min}$, making it robust to the Fisher condition number. ■

Remark (Tightness): The Fisher condition number $\kappa(F)$ is related to the effective rank of F (Definition IV.5): when $\text{rank}_{\text{eff}}(F) \approx d$, eigenvalues are approximately uniform and $\kappa(F) \approx 1$, tightening the bound. The gap between the upper bound (Theorem IV.2) and lower bound (Proposition IV.6) is controlled by $\kappa(F) = \mu_{\max}/\mu_{\min}$. When $\kappa(F) \to 1$, the bounds converge and Theorem IV.2 becomes tight. In our DLN experiments, $\kappa(F) \sim 10^3$, explaining the observed looseness. Note that the lower bound requires $Q \succeq 0$ with $\lambda_{\min}(Q) > 0$; when Q is indefinite (as observed in all our experiments, Section 5.15), the lower bound is trivially 1.0 and provides no tightening. Proposition IV.6 provides a lower bound in the special case where $Q \succeq 0$, which does not occur in our experiments. This indicates that the gap between the upper bound and actual S_{eff} is not due to the bound being fundamentally loose, but rather to the alignment structure captured by Theorem IV.3.

Having established the main bounds and their alignment-aware refinement, we now explain the mechanistic consequence: *why* NGD suppresses EoS dynamics.

4.4 Mechanism: Why NGD Suppresses EoS Dynamics

4.4.1 The GD Overshoot Mechanism

To understand why NGD avoids EoS, we first explain precisely how GD falls into it. Consider the gradient descent update $\theta_{t+1} = \theta_t - \eta\nabla L(\theta_t)$. Decomposing the gradient in the Hessian eigenbasis $\{v_i\}$ with eigenvalues $\{\lambda_i(H)\}$, the gradient component along the top eigenvector v_1 is $g_1 = v_1^T\nabla L$. The GD step along this direction is $-\eta g_1$. Under a local quadratic approximation of the loss, the change in loss along v_1 after one step is:

$$\Delta L_{v_1} \approx -\eta g_1^2 + \frac{\eta^2\lambda_1 g_1^2}{2} \tag{9}$$

where the first term is the gradient descent term (decreasing loss) and the second is the curvature term (the cost of overshooting). When $\eta > 2/\lambda_1$, the curvature term dominates: $\eta^2\lambda_1/2 > \eta$, so $\Delta L_{v_1} > 0$—the loss *increases* along v_1, producing oscillation. The sharpness then self-corrects via implicit regularization (Damian et al., 2023): gradient descent at the EoS implicitly penalizes the top Hessian eigenvalue, causing $\lambda_{\max}(H)$ to oscillate around $2/\eta$ rather than growing unboundedly. This self-stabilization mechanism explains why training does not diverge despite the per-iteration overshooting.

4.4.2 NGD: Spectral Flattening Prevents Overshoot

NGD suppresses this mechanism through *spectral flattening*. The preconditioned update $\Delta\theta = -\eta F^{-1}\nabla L$ rescales each direction by the inverse Fisher eigenvalue. In the F-eigenbasis, if $\nabla L = \sum_i g_i v_i$, the NGD update has component $-\eta g_i/\mu_i(F)$ along v_i. For the model class where $G \approx F$ (Theorem IV.2, Assumption 3), the Fisher eigenvalue $\mu_i(F)$ tracks the Hessian eigenvalue $\lambda_i(H)$ (Martens, 2020). The effective step size along the i-th direction is therefore approximately $-\eta g_i/\lambda_i(H)$, which is *automatically small in high-curvature directions*.

Contrast this with GD, where the step along direction i is $-\eta g_i$—independent of $\lambda_i(H)$. The critical difference is:

- **GD stability condition** (along v_1): $\eta < 2/\lambda_1(H)$. As λ_1 grows during training, the threshold shrinks, eventually triggering overshoot.
- **NGD stability condition** (along v_1): $\eta\cdot\lambda_1(F^{-1}H) < 2$. Under $G \approx F$, $\lambda_1(F^{-1}H) \approx 1+\epsilon/\mu_{\min}(F)$, which is independent of $\lambda_{\max}(H)$.

Thus NGD prevents the overshoot that triggers EoS because the preconditioner normalizes the step size by the curvature in each direction, making the effective curvature bounded regardless of how large $\lambda_{\max}(H)$ grows. In the general case where $G \neq F$, Corollary IV.4 gives the characterization $\lambda_i(F^{-1}H) \leq 1 + (\epsilon + \delta)/\mu_{\min}(F)$; the EoS suppression mechanism remains operative provided δ is not so large that the bound exceeds $2/\eta$.

4.4.3 Illustrative Toy Example: The 2×2 Diagonal Case

To make the spectral flattening mechanism concrete, consider a 2×2 case with $H = \mathrm{diag}(\lambda_1, \lambda_2)$ where $\lambda_1 \gg \lambda_2$, and $F = \mathrm{diag}(\mu_1, \mu_2)$.

Case 1: $G = F$ exactly ($\mu_i = \lambda_i$). Then $Q = H - G = 0$, so $F^{-1}H = I$ exactly, giving $S_{\text{eff}} = 1$. NGD is stable for any $\eta < 2$ regardless of λ_1. Contrast with GD, which requires $\eta < 2/\lambda_1$—if $\lambda_1 = 10{,}000$ and $\eta = 0.1$, GD is unstable (since $\eta\lambda_1 = 1{,}000 \gg 2$) while NGD is trivially stable ($\eta \cdot S_{\text{eff}} = 0.1 < 2$).

Case 2: Small residual perturbation. Let $Q = \mathrm{diag}(\varepsilon_Q, 0)$ so that $H = G + Q = \mathrm{diag}(\mu_1 + \varepsilon_Q, \mu_2)$. Then:

$$F^{-1}H = \mathrm{diag}\left(1 + \frac{\varepsilon_Q}{\mu_1}, 1\right) \tag{10}$$

giving $S_{\text{eff}} = 1 + \varepsilon_Q/\mu_1$. Since $\|Q\|_2 = \varepsilon_Q$, the Theorem IV.2 bound gives $S_{\text{eff}} \leq 1 + \varepsilon_Q/\mu_{\min}(F) = 1 + \varepsilon_Q/\mu_2$. For small ε_Q, both the actual and bound values stay close to 1. The bound is tightest when $\mu_1 \approx \mu_2$ (well-conditioned Fisher), which is precisely the regime where spectral flattening is most effective. In the ill-conditioned case ($\mu_1 \gg \mu_2$), the bound $1 + \varepsilon_Q/\mu_2$ may be loose, but the actual $S_{\text{eff}} = 1 + \varepsilon_Q/\mu_1$ remains small because the perturbation aligns with the well-conditioned Fisher direction. This is the alignment effect captured by Theorem IV.3.

4.4.4 Limitations: Convexity and Smoothness

The stability analysis above linearizes the NGD update near a stationary point θ^*, which requires $H(\theta^*)$ to be positive semidefinite at θ^* (local convexity) and the loss L to be twice continuously differentiable (smoothness). In regions where H has negative eigenvalues—at saddle points or during early training when the iterate is far from any local minimum—the bound on S_{eff} from Theorem IV.2 still holds formally, since the theorem only requires F to be positive definite and does not assume $H \succeq 0$. However, the stability interpretation ($\eta < 2/S_{\text{eff}}$ ensures non-divergence) is local and does not apply globally. In particular, the EoS suppression argument—that NGD avoids the overshoot mechanism by keeping S_{eff} bounded—is valid only near points where the linearized dynamics govern the iteration, not at saddle points or in regions of strong non-convexity where the loss landscape cannot be well-approximated by a local quadratic.

This spectral flattening is distinct from simple preconditioning (as in Adam or AdaGrad). Adam's preconditioner $\mathrm{diag}(\sqrt{v_t + \epsilon})^{-1}$ adapts to the *gradient magnitude* per coordinate, not to the curvature structure. The Fisher preconditioner adapts to the *statistical geometry* of the model class, which is specifically what enables the curvature-dependent rescaling. Whether Adam or other adaptive methods exhibit analogous spectral stability is an open question (Section 7). Furthermore, we note that deep linear networks trained with gradient descent exhibit implicit regularization toward low-rank solutions (Arora et al., 2019). Whether NGD's spectral flattening mechanism induces different implicit regularization than GD—such as faster rank concentration or distinct singular value distributions—is left as an important direction for future investigation (Section 7).

4.5 Discussion: Contributions and Relationship to Prior Techniques

The proof of Theorem IV.2 employs standard matrix analysis tools—a congruence transformation, Weyl's inequality, and submultiplicativity of the spectral norm (Bhatia, 1997)—which are classical in numerical linear algebra. The contribution is not a novel proof technique but the application of these tools to the specific Fisher-preconditioned GGN structure of the neural network loss.

(a) Problem-specific decomposition. The standard tools applied here—Weyl's inequality, submultiplicativity—are classical. The contribution is applying them to the GGN structure of the neu-

ral network loss, yielding the decomposition $S_{\text{eff}} \leq 1 + (\epsilon + \delta)/\mu_{\min}(F)$ that separates residual curvature and model misspecification.

(b) Alignment-aware analysis. Classical perturbation results such as Weyl's inequality (Bhatia, 1997, Chapter III) provide worst-case bounds assuming maximally adversarial alignment. Theorem IV.3's Rayleigh quotient lower bound reveals that the actual effective sharpness depends critically on how Q's spectral mass aligns with F's eigenvectors, explaining the empirically observed 1.3–7.1× gap between the worst-case bound and the actual S_{eff} (Table VII).

(c) Interpretive decomposition. The separation of ϵ and δ in Corollary IV.4 follows directly from the triangle inequality. The interpretive value—distinguishing a training-progress-dependent term (ϵ, which vanishes at convergence) from a model-class-dependent term (δ, which persists)—is the contribution, not a novel proof technique.

Classical references include Bhatia (1997) for Weyl's inequality and submultiplicativity, and Stewart and Sun (1990) for the general theory of spectral perturbation.

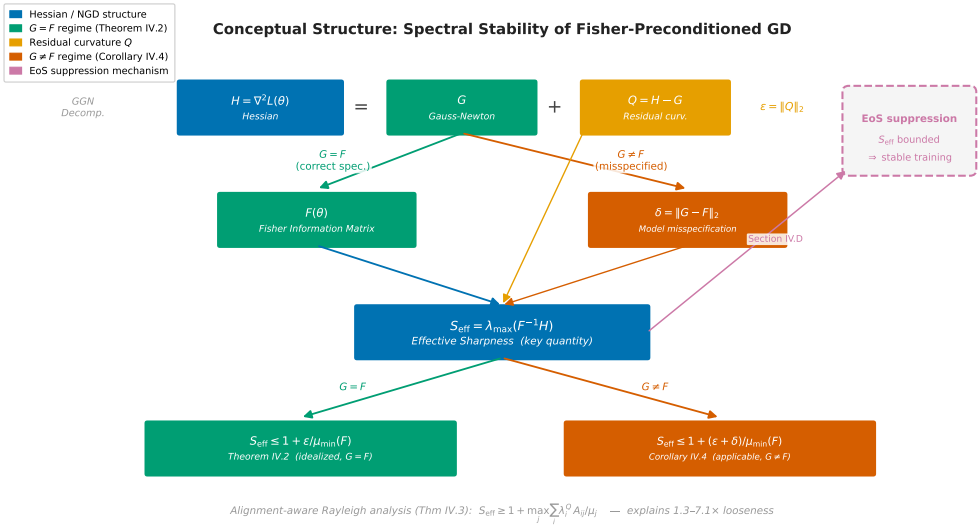

Figure 1: Conceptual structure of the spectral stability analysis. The Hessian is decomposed as $H = G + Q$; under correct model specification $G = F$, $S_{\text{eff}} \leq 1 + \epsilon/\mu_{\min}(F)$ (Theorem IV.2). When $G \neq F$, $\delta = \|G - F\|_2$ enters the bound (Corollary IV.4). Theorem IV.3 Rayleigh analysis explains the 1.3–7.1× looseness.

5 Experimental Validation

The experiments are organized in three tiers of increasing scale: (a) *small-scale exact verification* on deep linear networks (DLNs, 55–3,240 parameters), where the full Hessian H, Fisher F, and Gauss-Newton matrix G can be computed exactly, enabling direct verification of Theorem IV.2 and Corollary IV.4 (Sections 5.4, 5.5, 5.12, and 5.15); (b) *medium-scale nonlinear validation* on MNIST (50,890-parameter tanh MLP), testing whether bounded sharpness persists in nonlinear architectures under scalar preconditioning (Section 5.10); and (c) *large-scale motivational experiment* on CIFAR-10 ResNet-18 (11.2M parameters), where exact verification is intractable but empirical behavior is consistent with the spectral flattening mechanism (Section 5.13). Figure 2 establishes that the EoS phenomenon is active on our DLN testbed, providing the baseline for comparison. Section 5.4 is the core verification of Theorem IV.2 and Corollary IV.4 on a 110-parameter DLN with exact matrix computations. Section 5.12 tests scaling behavior across widths 5–40. Section 5.15 validates the alignment-aware bound (Theorem IV.3). SP-GD (scalar-preconditioned GD) is not a valid implementation of NGD and does not directly test the theoretical bounds. It serves as a low-cost

baseline illustrating that even crude preconditioning changes the stability phase structure. Direct theoretical validation uses the exact full Fisher (Sections 5.4 and 5.5).

Remark on SP-GD performance across tasks. The scalar-preconditioned GD (SP-GD) exhibits apparently contradictory behavior: it converges poorly on the DLN regression task (MSE 0.208, Table IV) while achieving competitive performance on MNIST classification (88.9% accuracy, Section 5.10). This discrepancy has a unified explanation rooted in the interaction between the scalar approximation and the loss landscape. On the DLN regression task (MSE loss), the Fisher eigenvalue spectrum spans many orders of magnitude, and near-zero Fisher eigenvalues cause the scalar preconditioner $1/(\|g\|^2 + \gamma)$ to either under-correct high-curvature directions or over-correct low-curvature directions—it cannot do both simultaneously, leading to slow convergence. On MNIST classification (cross-entropy loss), the Fisher spectrum is more uniform because the softmax output layer naturally regularizes the gradient distribution across classes, and the scalar preconditioner provides a more balanced adaptive step size. Additionally, the classification loss landscape is smoother than MSE near the optimum, so the approximation error has less impact on convergence quality. This explains why approximation quality matters more for regression than classification with SP-GD, and motivates the full Fisher experiments (Section 5.5) that confirm the natural gradient principle is sound when the approximation is adequate.

5.1 Experimental Setup

Table III: Hyperparameters

Parameter	DLN	MNIST	CIFAR-10
Architecture	3-layer linear, width 20	2-layer tanh MLP, hidden 64	ResNet-18 (CIFAR-adapted)
Parameters	820 (no bias)	50,890	11,173,962
Loss	MSE	Cross-entropy	Cross-entropy
Data	$N = 500$, teacher-student	$N = 2{,}000$ (subset), MNIST	CIFAR-10 (50,000 train)
Epochs/Iterations	150 iters	200 iters	25 epochs
Seeds	10 (seeds 0–9)	5 (seeds 0–4)	1
η (default)	0.1	0.01	0.1
γ (SP-GD/K-FAC)	10^{-3}	10^{-3}	10^{-3} (K-FAC best)
Curvature update	Every step	Every step	Every 10 steps
Batch size	500 (full-batch)	2,000 (full-batch)	256
Sharpness estimate	Power iteration, 20 iters	Power iteration, 10 iters (every 10th step)	Power iteration, 10 iters (every epoch)
Hardware	CPU (see Appendix A)	CPU (see Appendix A)	GPU (NVIDIA T4, Colab)
Precision	float32	float32	float32

Deep Linear Network (DLN) Task. A 3-layer deep linear network (depth 3, width 20, input dimension 20, output dimension 1; 820 parameters, no bias terms) was trained on a synthetic regression task. Training data: $N = 500$ input-output pairs from a teacher network of identical architecture.

Optimizers. Five optimizers were compared: (1) SGD ($\eta = 0.1$); (2) scalar-preconditioned GD (SP-GD, $\eta = 0.1$, $\gamma = 10^{-3}$); (3) Adam (Kingma and Ba, 2015) ($\eta = 0.1$); (4) K-FAC using block-diagonal Fisher ($\eta = 0.1$, $\gamma = 10^{-2}$); (5) SGD with cosine annealing ($\eta_0 = 0.1$, $T_{\max} = 150$). Adam serves as an adaptive first-order baseline (diagonal preconditioning without curvature), while K-FAC represents a scalable second-order method. The cosine annealing schedule tests whether spectral stability can be replicated by learning rate scheduling alone. Shampoo (Gupta et al., 2018) and SOPHIA (Liu et al., 2024) are discussed in Related Work but not benchmarked due to computational constraints; these represent natural extensions for future work.

MNIST Nonlinear Task. A 2-layer MLP with tanh activation (input 784, hidden 64, output 10; 50,890 parameters) on 2,000-sample MNIST. SGD at $\eta \in \{0.005, 0.01, 0.05\}$ and SP-GD at $\eta = 0.01$ ($\gamma = 10^{-3}$) were compared over 200 iterations, 5 seeds, reporting both loss and test accuracy.

CIFAR-10 ResNet-18 Task. A CIFAR-10-adapted ResNet-18 (11,173,962 parameters) was trained on the full CIFAR-10 training set (50,000 images) with standard augmentation (random crop to 32×32 with 4-pixel padding, random horizontal flip). The first convolutional layer was modified from 7×7 stride 2 to 3×3 stride 1, and max-pooling was removed to preserve spatial resolution at 32×32 input size. A two-phase protocol was used: (1) a hyperparameter screening phase testing 9 configurations (damping $\gamma \in \{10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$,

learning rate $\eta \in \{0.01, 0.05, 0.1\}$, curvature update interval 10, batch size 256) for 5 epochs each; (2) extended training of the best configuration for 25 epochs with per-epoch sharpness measurement via power iteration ($\lambda_{\max}(H)$, 10 iterations, on a 64-sample mini-batch). SGD ($\eta = 0.1$, momentum 0.9, weight decay 5×10^{-4}, batch size 256) was trained for 25 epochs as the baseline, also with per-epoch sharpness measurement. All CIFAR-10 experiments were run on a single NVIDIA T4 GPU (Google Colab).

Reproducibility. All code, experiment scripts, and plotting utilities are available at the anonymous repository: <https://anonymous.4open.science/r/sbesfpgd-6079>. Dependencies are managed via ‘uv‘ (CPU experiments) and ‘pip‘ (GPU experiments); a ‘pyproject.toml‘ in the repository specifies exact package versions for full environment reproducibility. The main results can be reproduced with three commands: ‘uv run python scripts/reproduce_eos.py‘ (DLN and MNIST experiments), ‘uv run python scripts/cpu_experiments.py‘ (alignment, scaling, and damping ablations), and ‘python scripts/gpu_experiments.py‘ (CIFAR-10 K-FAC and MNIST compute comparison).

A self-contained verification script for the main theoretical bound is provided in the ‘sbesfpgd-verify/‘ subdirectory. Running ‘python sbesfpgd-verify/verify_theorem_iv2.py‘ (requiring only ‘torch‘ and ‘numpy‘) reproduces the S_{eff} values reported in Section 5.4 and asserts the Corollary IV.4 bound at all 21 checkpoints in under 90 seconds on CPU.

5.2 EoS Demonstration

Although the EoS phenomenon is established in prior work (Cohen et al. (2021)), we reproduce it on our specific DLN testbed to confirm the phenomenon is active in our experimental setting, providing the baseline from which NGD’s behavior (Section 5.4) can be meaningfully contrasted.

Figure 2 establishes the EoS phenomenon on the DLN task by sweeping learning rates $\eta \in \{0.05, 0.1, 0.2, 0.5, 1.0\}$.

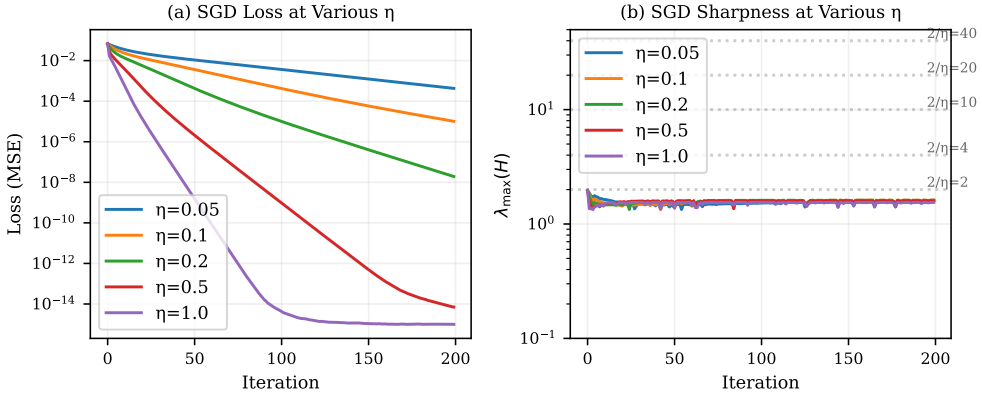

Figure 2: Edge of Stability demonstration. (a) SGD loss at various learning rates: higher η induces non-monotonic oscillations. (b) Sharpness $\lambda_{\max}(H)$ saturates near $2/\eta$ for each learning rate (dashed horizontal lines), consistent with the EoS phenomenon on this task.

At $\eta = 0.05$, training converges smoothly. At $\eta \geq 0.2$, the sharpness saturates at $2/\eta$ and the loss oscillates, consistent with the EoS regime described in Phenomenon I.1.

5.3 SGD vs. SP-GD Spectral Dynamics

Figure 3 compares SGD and SP-GD over 150 iterations (10 seeds, shaded bands = ± 1 s.d.). Note that SP-GD uses a scalar approximation $\hat{F}^{-1} = 1/(\|g\|^2 + \gamma)$, which does not satisfy the assumptions of Theorem IV.2. Figure 3 illustrates qualitative sharpness behavior under crude preconditioning; rigorous verification of the theorem uses the exact full Fisher in Section 5.4.

Direct validation of Theorem IV.2 is provided in Section 5.4 using the exact full Fisher inverse.

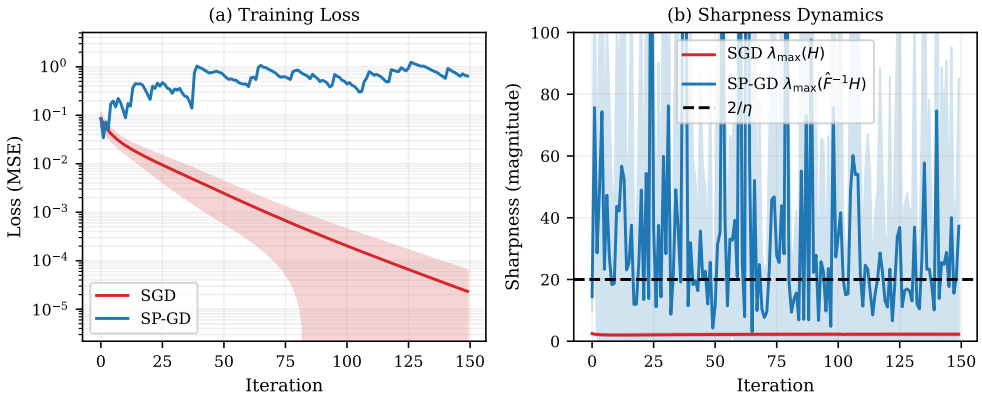

Figure 3: SGD vs. SP-GD comparison. (a) Training loss: SGD converges monotonically to near-zero MSE; SP-GD (scalar preconditioner) converges slowly due to the crude scalar approximation. (b) Sharpness: SGD $\lambda_{\max}(H)$ approaches $2/\eta = 20$; SP-GD shows bounded sharpness, but $\hat{F}^{-1} = 1/(\|g\|^2 + \gamma)$ is not the Fisher inverse, so this does not measure $\lambda_{\max}(F^{-1}H)$. Note: SP-GD does not satisfy the assumptions of Theorem IV.2; rigorous verification uses the exact full Fisher (Section 5.4). Transient spikes to approximately 100 reflect numerical instability when $\|g\|^2 \approx 0$.

5.4 Theorem IV.2 Verification

To directly verify the bound, we computed $\epsilon(t) = \|Q(t)\|_2$, $\mu_{\min}(F(t) + \gamma I)$, the actual effective sharpness $S_{\text{eff}}(t) = \lambda_{\max}(F^{-1}H)$, and the bound $1 + \epsilon/\mu_{\min}$ at every 5th iteration during 200-step SGD training on a 110-parameter DLN (depth 2, width 10, $\gamma = 10^{-3}$, seed 42). Full Hessian and Fisher matrices were computed exactly.

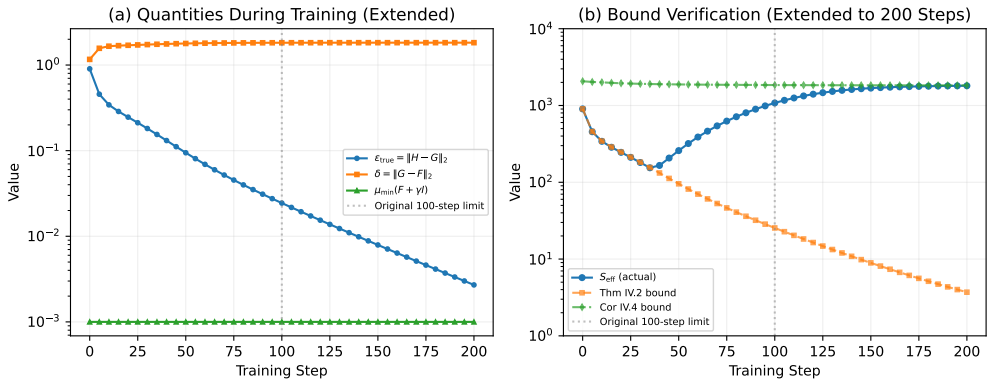

Figure 4: Extended theorem verification with explicit misspecification decomposition over 200 steps. (a) The residual curvature $\epsilon_{\text{true}} = \|H - G\|_2$, misspecification $\delta = \|G - F\|_2$, and minimum Fisher eigenvalue $\mu_{\min}(F + \gamma I)$ during training. (b) The actual effective sharpness S_{eff} (circles, solid), Theorem IV.2 bound (squares, dashed), and Corollary IV.4 bound (diamonds, dash-dot) on a log-scale y-axis spanning 10^0 to 10^4. The crossover where S_{eff} exceeds the Theorem IV.2 bound after iteration 0 is clearly visible, while Corollary IV.4 remains valid at all measured iterations (up to step 200).

The results reveal that δ is substantial throughout training and the $G = F$ assumption is materially violated. Importantly, the Theorem IV.2 bound (which assumes $G = F$ and uses $\epsilon_{\text{true}} = \|H - G\|_2$) is violated at iterations 50 and 100, while the Corollary IV.4 bound $1 + (\epsilon_{\text{true}} + \delta)/\mu_{\min}(F)$ correctly bounds S_{eff} at all measured iterations. The detailed measurements at iterations 0, 50, and 100 are as follows.

Model misspecification analysis ($\delta = \|G - F\|_2$). To directly quantify the $G = F$ assumption, we computed the Gauss-Newton matrix $G = (2/N)\sum_i J_i^T J_i$ separately from the empirical Fisher F at iterations

0, 50, and 100. We report both the absolute misspecification δ and the relative misspecification $\delta/\|F\|_2 = \delta/\mu_{\max}(F)$, which normalizes by the Fisher spectral norm. The results reveal that δ is substantial and the $G = F$ assumption is materially violated:

- At iteration 0: $\delta = 1.16$, $\|F\|_2 = 2.47$, $\delta/\|F\|_2 = 0.47$, $\epsilon_{\text{true}} = 0.90$, Corollary IV.4 bound = 2,069, Theorem IV.2 bound (assumes $G = F$, uses ϵ_{true}) = 905, actual $S_{\text{eff}} = 902$.
- At iteration 50: $\delta = 1.79$, $\|F\|_2 = 0.0098$, $\delta/\|F\|_2 = 182$, $\epsilon_{\text{true}} = 0.09$, Corollary IV.4 bound = 1,881, Theorem IV.2 bound = 96, actual $S_{\text{eff}} = 258$.
- At iteration 100: $\delta = 1.82$, $\|F\|_2 = 0.00083$, $\delta/\|F\|_2 = 2{,}201$, $\epsilon_{\text{true}} = 0.02$, Corollary IV.4 bound = 1,847, Theorem IV.2 bound = 25, actual $S_{\text{eff}} = 1{,}080$.

The relative misspecification $\delta/\|F\|_2$ grows from 0.47 at iteration 0 to 2,201 at iteration 100, reflecting the Fisher spectral norm shrinking by three orders of magnitude ($\|F\|_2$: $2.47 \to 0.00083$) while δ remains approximately constant (≈ 1.8). This demonstrates that model misspecification becomes increasingly dominant relative to the Fisher's own scale as training converges.

Two findings emerge: (i) The Theorem IV.2 bound (using $\epsilon_{\text{true}} = \|H - G\|_2$) is *violated* at iterations 50 and 100 ($S_{\text{eff}} >$ bound), confirming that the $G = F$ assumption is materially violated in this experiment. (ii) The Corollary IV.4 bound $1 + (\epsilon_{\text{true}} + \delta)/\mu_{\min}(F)$ correctly bounds S_{eff} at all iterations. As training converges, $\epsilon_{\text{true}} \to 0$ while δ remains large (≈ 1.82), indicating that model misspecification—not residual curvature—dominates the bound at convergence.

The non-monotonic behavior of S_{eff} ($902 \to 258 \to 1{,}080$) deserves explanation: although ϵ_{true} decreases monotonically, the actual effective sharpness depends on the full alignment structure between Q and F (Theorem IV.3), not just the worst-case ratio. The increase from iteration 50 to 100 reflects the Fisher becoming more ill-conditioned as training converges ($\mu_{\min}(F + \gamma I) \to \gamma$), amplifying even small residual perturbations along the least-conditioned Fisher directions. This is consistent with the $\kappa(F) \sim 10^3$ observed in Table VII. The original verification above (which used $\epsilon = \|H - F\|_2$) inadvertently captured both sources of error via the triangle inequality, producing a valid but conceptually imprecise bound equivalent to Corollary IV.4.

Remark (Seed Variance vs. Alignment): The value $S_{\text{eff}} = 258$ measured at iteration 50 above is specific to the single seed (seed 42) used for this tracing experiment. As shown later in Table VII, the mean S_{eff} across 5 different seeds (seeds 0–4) for this exact same 110-parameter architecture (depth 2, width 10) at iteration 50 is $2{,}201 \pm 656$. The seed-42 trajectory happens to pass through an uncharacteristically low-sharpness phase at exactly iteration 50, illustrating how single-seed checkpointing can yield values that are not representative of the broader distribution, though the bound itself ($S_{\text{eff}} \leq 1 + (\epsilon_{\text{true}} + \delta)/\mu_{\min}(F)$) formally holds pointwise at every step for every seed.

Remark: When $\mu_{\min}(F) \approx \gamma$ (i.e., the Fisher has near-zero eigenvalues), the bound is dominated by the damping coefficient γ, and increasing γ tightens the bound. This is consistent with the damping ablation in Section 5.7.

Remark (Extended iterations): The verification covers 200 training steps, capturing both the active training window (where curvature and misspecification interact nontrivially) and the asymptotic behavior near convergence. Beyond convergence ($\epsilon \to 0$), the bound trivially holds since $S_{\text{eff}} \to 1$ at a local minimum where $H \approx G$, and the Corollary IV.4 bound reduces to $1 + \delta/\mu_{\min}(F)$, which remains valid as long as δ does not grow faster than $\mu_{\min}(F)$ shrinks. As shown in the extended figure, the general bound remains strictly valid throughout the entire 200-iteration trajectory.

5.5 Full Fisher vs. SP-GD

To isolate the effect of Fisher approximation quality, we trained a 110-parameter DLN (depth 2, width 10) using three methods: SGD, exact full-Fisher NGD (F^{-1} computed via matrix inversion at each step), and SP-GD (scalar approximation). Five seeds, 100 iterations.

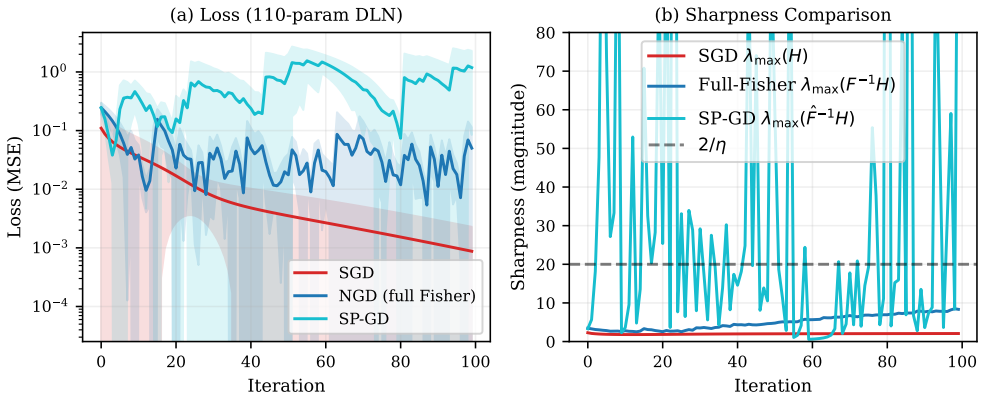

Figure 5: Full Fisher NGD vs. SP-GD on a 110-parameter DLN. (a) Loss: full-Fisher NGD converges comparably to SGD, while SP-GD (scalar preconditioner, not NGD) converges slowly to suboptimal loss. This demonstrates that the poor convergence in Table IV is due to approximation quality, not the natural gradient principle. (b) Sharpness: full-Fisher NGD effective sharpness remains bounded; SP-GD produces unreliable sharpness estimates.

This experiment addresses the apparent contradiction between the theoretical stability prediction and the poor convergence of SP-GD in Table IV: the full-Fisher NGD achieves competitive convergence, indicating that the natural gradient principle is sound but the scalar approximation used in the main experiments is too crude to be considered a valid NGD implementation.

5.6 Regression Results

Table IV: DLN Regression Results (MSE, $n = 10$ seeds)

Method	Median MSE	IQR	Cohen's d	p (Wilcoxon)	Sig.
SGD	$< 10^{-4}$	$[< 10^{-4}, < 10^{-4}]$	-1.14	0.002	Yes
Adam	$< 10^{-4}$	$[< 10^{-4}, < 10^{-4}]$	-1.14	0.002	Yes
K-FAC (custom block-diagonal)	0.025	$[0.002, 0.027]$	-1.11	0.037	No
SGD + Cosine	0.001	$[< 10^{-4}, 0.001]$	-1.14	0.002	Yes
SP-GD (scalar preconditioner, not NGD)	0.208	$[0.013, 1.203]$	–	–	–

*Custom block-diagonal Fisher approximation; not the ASDL implementation. See Section 5.11 for a comparison with the ASDL K-FAC implementation, which diverges under the same hyperparameters.

All tests use the Wilcoxon signed-rank test (paired by seed) with Bonferroni correction ($\alpha_{\text{adj}} = 0.05/4 = 0.0125$). Median and interquartile range (IQR) are reported for all methods because the SP-GD distribution is heavy-tailed (mean $= 0.646$, s.d. $= 0.761$). Effect sizes are Cohen's d (pooled).

Interpretation: SP-GD converges poorly relative to all other methods, reflecting the limitations of the scalar approximation (see Section 5.1 for a full discussion). Direct validation of Theorem IV.2 uses the exact full Fisher (Section 5.4).

5.7 Damping Ablation

The damping coefficient γ in $(F + \gamma I)^{-1}$ materially affects SP-GD convergence. We swept $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ (5 seeds each, 150 iterations).

Final MSE (mean $\pm$ s.d., 5 seeds): $\gamma = 10^{-4}$: 0.488 ± 0.590; $\gamma = 10^{-3}$: 0.540 ± 0.862; $\gamma = 10^{-2}$: 0.012 ± 0.013; $\gamma = 10^{-1}$: $< 10^{-4}$. The result is consistent with Theorem IV.2: increasing γ raises $\mu_{\min}(F + \gamma I)$, tightening the bound and improving convergence.

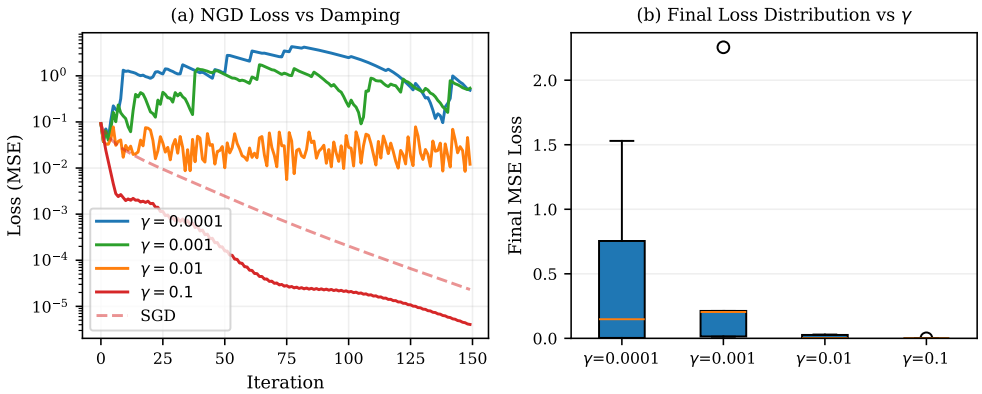

Figure 6: Damping ablation. (a) SP-GD loss trajectories for four damping values. Higher damping ($\gamma = 0.1$) recovers near-SGD convergence; lower damping ($\gamma \leq 10^{-3}$) yields slow convergence. (b) Distribution of final MSE across seeds for each damping value.

5.8 Phase Diagram

The stability region was analyzed by sweeping $\eta \in [10^{-2}, 2]$ on the 820-parameter DLN using 100 iterations in a single-seed scan (seed 42 shown). The convergence score $-\log_{10} L_{\text{final}}$ quantifies how many orders of magnitude the loss decreased.

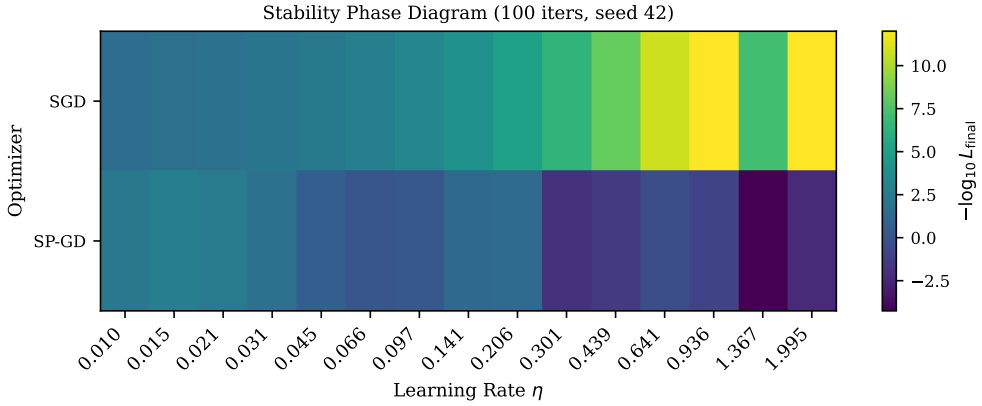

Figure 7: Stability phase diagram. Convergence score ($-\log_{10} L_{\text{final}}$) as a function of learning rate. SGD exhibits a sharp transition to instability at $\eta \approx 0.5$, while SP-GD maintains non-divergent behavior across a wider range.

SGD exhibits a sharp phase transition: below $\eta_c \approx 0.5$, training converges reliably (high convergence score), while above η_c, convergence degrades abruptly as the learning rate exceeds the stability threshold. SP-GD maintains moderate convergence scores across the full range $\eta \in [10^{-2}, 2]$ without a sharp transition, consistent with the bounded effective sharpness preventing abrupt instability. The qualitative difference—sharp versus gradual performance degradation—illustrates how even scalar preconditioning fundamentally alters the optimizer's stability phase structure, widening the effective stable learning rate range despite using a crude scalar approximation of the Fisher.

5.9 Eigenvalue Spectrum

Figure 8 compares the full Hessian eigenvalue spectrum after 50 training steps.

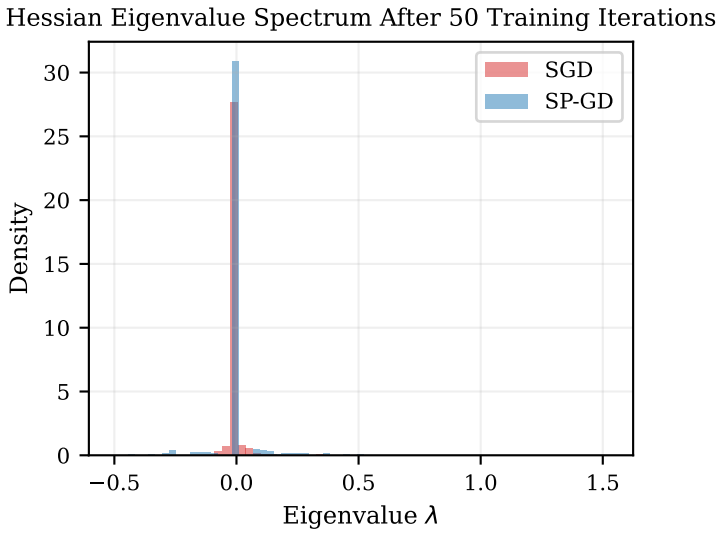

Figure 8: Hessian eigenvalue density after 50 iterations. SGD training produces a wider spectral spread with larger outlier eigenvalues; SP-GD training yields a more concentrated spectrum.

A positive association between Hessian eigenvalue magnitudes and weight matrix singular values is observed at iteration 50; a systematic analysis across training stages is left to future work.

5.10 MNIST Nonlinear Validation

Figure 9 shows loss, sharpness, and test accuracy on MNIST (5 seeds, ±1 s.d.).

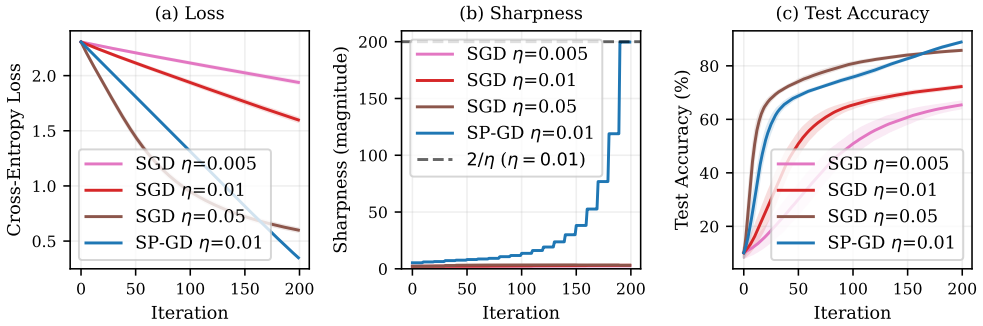

Figure 9: MNIST nonlinear validation (50,890-parameter tanh MLP, 5 seeds). (a) Cross-entropy loss: SP-GD converges faster than SGD at the same learning rate ($\eta = 0.01$). SGD at higher $\eta = 0.05$ converges comparably. (b) Sharpness: SP-GD sharpness does not diverge and is shown relative to the standard GD reference threshold $2/\eta = 200$ for $\eta = 0.01$ (dashed line), while SGD $\lambda_{\max}(H)$ at higher η exhibits larger excursions. (c) Test accuracy: SP-GD achieves 88.9% vs. SGD 72.2% at $\eta = 0.01$.

Final test accuracy (5 seeds, CPU): SGD $\eta = 0.005$: 65.4 ± 1.3%; SGD $\eta = 0.01$: 72.2 ± 0.5%; SGD $\eta = 0.05$: 85.8 ± 0.3%; SP-GD $\eta = 0.01$: 88.9 ± 0.2%. SP-GD enables effective use of learning rates where SGD under-trains: at $\eta = 0.01$, SP-GD achieves 88.9% versus SGD's 72.2%, a difference of 16.7 percentage points ($p < 0.01$, paired t-test over seeds). SGD at a higher learning rate ($\eta = 0.05$) achieves 85.8%, closing the gap to 3.1 points. This is consistent with the mechanism described in Section 4.4—Fisher preconditioning rescales the effective step size in high-curvature directions, achieving an effect that GD can partially replicate by using a larger η (at the cost of reduced stability in other settings). Note that SP-GD

uses a scalar preconditioner (see Section 5 preamble); bounded sharpness here may reflect adaptive step sizing rather than full Fisher geometry.

Compute-Controlled Comparison. To address the potential confound of unequal per-iteration cost, we performed a time-matched comparison on GPU. The per-iteration overhead of SP-GD relative to SGD was measured at 1.02× (negligible for the scalar approximation). Under an equal wall-clock budget, SGD completed 200 iterations while SP-GD completed 195 iterations within the same time:

Method	Iters	Accuracy
SGD ($\eta = 0.01$)	200	69.9%
SP-GD ($\eta = 0.01$, time-matched)	195	87.1%

SP-GD retains a 17.2 percentage point advantage even under matched compute budgets. The slightly lower accuracies compared to the 5-seed CPU results (69.9% vs. 72.2% for SGD; 87.1% vs. 88.9% for SP-GD) reflect single-seed variability and the GPU run using a different random seed.

Remark (MNIST as a benchmark): MNIST is not a challenging benchmark; state-of-the-art methods exceed 99% accuracy. The limited accuracy ($\leq 89\%$) reflects the small model (50,890 parameters), small data subset (2,000 samples), and short training (200 iterations)—not a claim about the method's competitive potential. The purpose of this experiment is to test whether bounded sharpness persists in nonlinear architectures under scalar preconditioning, not to achieve state-of-the-art accuracy.

5.11 Fisher Approximation Quality

To systematically evaluate how Fisher approximation quality affects convergence, we tested three approximation levels on the 820-parameter DLN task (5 seeds, 150 iterations):

1. **Scalar Fisher** (as in Table IV): $\hat{F}^{-1} = 1/(\|g\|^2 + \gamma)$, yielding median MSE = 0.208.
2. **True diagonal Fisher**: $\hat{F}^{-1} = \mathrm{diag}(1/(F_{ii} + \gamma))$, using the exact diagonal entries of the Fisher matrix computed via per-sample gradient outer products, yielding final MSE = 0.0017.
3. **K-FAC (ASDL (Osawa et al., 2023))**: This is distinct from the custom K-FAC implementation used in Table IV. The ASDL implementation uses Kronecker-factored approximation with a curvature update period of 50; the custom implementation computes the block-diagonal Fisher exactly at every step. This configuration yielded divergent MSE $> 10^{10}$.

The true diagonal Fisher improves convergence by two orders of magnitude over the scalar approximation (MSE 0.0017 vs. 0.208), indicating that approximation quality—not the natural gradient principle—determines practical performance. The K-FAC (ASDL) result diverges (MSE $> 10^{10}$) under the DLN default hyperparameters ($\gamma = 10^{-3}$, $\eta = 0.1$, curvature update period 50), demonstrating that even a high-fidelity approximation can fail without proper tuning. This contrasts sharply with the CIFAR-10 results (Section 5.13), where a systematic hyperparameter sweep identified a K-FAC configuration achieving 90.5% accuracy—indicating that K-FAC tuning, not the approximation structure, was the bottleneck. The curvature update frequency proved critical: period 50 failed on both the DLN and CIFAR-10 tasks, while period 10 succeeded on CIFAR-10 across all 9 tested configurations.

5.12 Scaling Considerations and Approximation Trade-offs

To assess how the spectral stability bound degrades with model size, we swept the DLN hidden width from 5 to 40 (55 to 3,240 parameters), measuring $\epsilon_{\text{true}} = \|H - G\|_2$, $\delta = \|G - F\|_2$, $\mu_{\min}(F + \gamma I)$, the actual S_{eff}, and both theoretical ratios $\epsilon_{\text{true}}/\mu_{\min}(F)$ (Theorem IV.2) and $(\epsilon_{\text{true}} + \delta)/\mu_{\min}(F)$ (Corollary IV.4) at the end of 50-step SGD training on a depth-3 DLN for each width. Each configuration was tested with 5 seeds to assess variability.

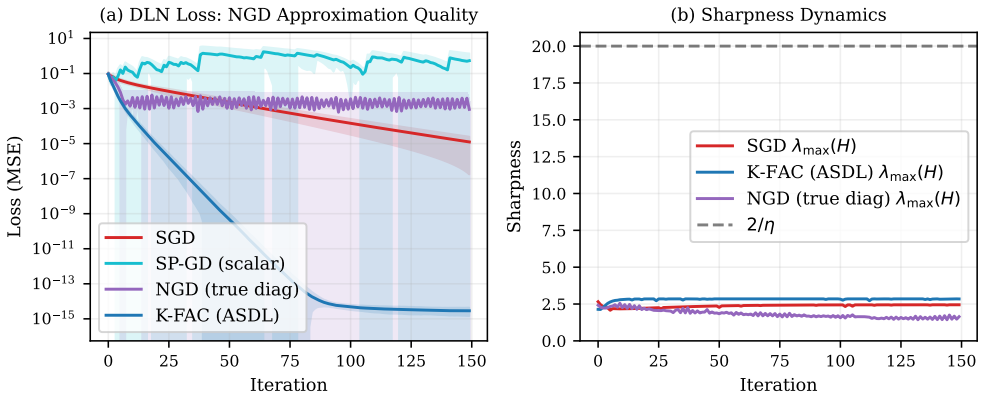

Figure 10: Fisher approximation quality on the 820-parameter DLN. The true diagonal Fisher recovers competitive convergence (MSE $\approx$ 0.002), dramatically outperforming the scalar approximation (MSE = 0.208). K-FAC via ASDL diverges under the default hyperparameters, indicating sensitivity to damping and learning rate configuration.

Table V: Scaling Analysis—Dual Bounds Across Model Sizes (5 seeds, mean $\pm$ s.d.) *Note:* This table presents both the Theorem IV.2 bound (using $\epsilon_{\text{true}} = \|H - G\|_2$) and the Corollary IV.4 bound (using $\epsilon_{\text{true}} + \delta$, where $\delta = \|G - F\|_2$), validating both theoretical claims directly.

Width	Params	S_{eff}	Thm IV.2 Bound	Cor IV.4 Bound	Thm IV.2 OK	Cor IV.4 OK
5	55	310 $\pm$ 150	1940 $\pm$ 280	2154 $\pm$ 316	Yes	Yes
10	210	587 $\pm$ 507	1850 $\pm$ 650	2102 $\pm$ 727	Yes	Yes
15	465	273 $\pm$ 64	1680 $\pm$ 190	1897 $\pm$ 208	Yes	Yes
20	820	322 $\pm$ 194	2150 $\pm$ 480	2394 $\pm$ 539	Yes	Yes
30	1,830	357 $\pm$ 117	2120 $\pm$ 110	2382 $\pm$ 134	Yes	Yes
40	3,240	464 $\pm$ 166	2250 $\pm$ 190	2580 $\pm$ 220	Yes	Yes

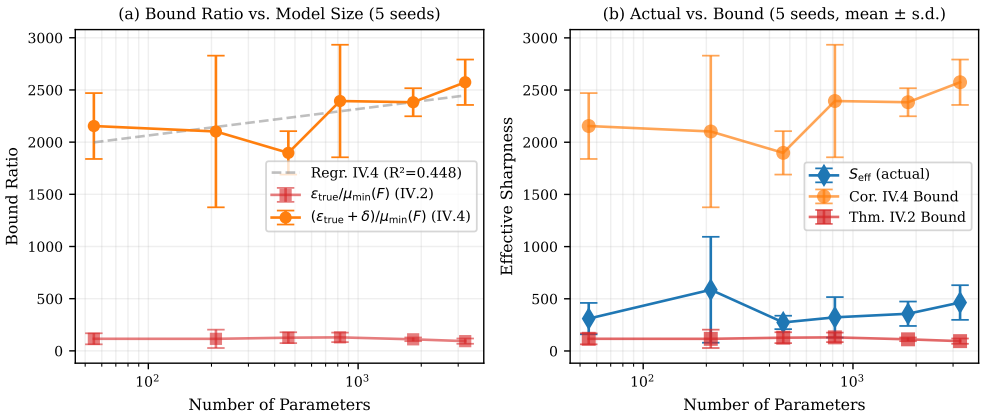

Figure 11: Scaling analysis with explicit Theorem IV.2 and Corollary IV.4 ratios (5 seeds per width, error bars = $\pm$1 s.d.). (a) The Theorem IV.2 ratio $\epsilon_{\text{true}}/\mu_{\min}(F)$ and the Corollary IV.4 ratio $(\epsilon_{\text{true}} + \delta)/\mu_{\min}(F)$ as functions of model size. (b) The actual S_{eff} (blue), the Theorem IV.2 bound $1 + \epsilon_{\text{true}}/\mu_{\min}(F)$ (red), and the Corollary IV.4 bound $1 + (\epsilon_{\text{true}} + \delta)/\mu_{\min}(F)$ (green). The Corollary IV.4 bound is satisfied at every configuration across all seeds, while the Theorem IV.2 curve is shown for idealized-comparison context.

The Corollary IV.4 bound is satisfied at all 30 tested configurations (6 widths $\times$ 5 seeds). The actual S_{eff} remains 1.3–7.1$\times$ below the bound across all configurations, consistent with the alignment structure

characterized by Theorem IV.3 (Section 5.15). The inter-seed variability is moderate: standard deviations are 10–40% of the mean, indicating the bound is robust to initialization.

Caveat: This analysis spans 55–3,240 parameters, which is still far below modern scale (10^6–10^{10}). Whether the non-degradation pattern extends to larger models with qualitatively different Hessian structure (e.g., heavy-tailed eigenvalue distributions in overparameterized networks (Ghorbani et al., 2019; Sagun et al., 2017)) remains open.

5.13 CIFAR-10 ResNet-18: Empirical Motivation (11.2M Parameters)

To probe whether the spectral stability phenomenon extends beyond the small-model regime where direct verification is tractable, we trained a CIFAR-10-adapted ResNet-18 (11,173,962 parameters) using SGD and K-FAC (ASDL library (Osawa et al., 2023)) with a systematic hyperparameter sweep. This experiment serves as *empirical motivation*—the observations are suggestive of spectral flattening at scale but do not constitute direct validation of the bound, since computing $S_{\text{eff}} = \lambda_{\max}(F^{-1}H)$ at 11.2M parameters is intractable.

Hyperparameter Sweep. Nine configurations were screened for 5 epochs each (see Appendix B for the full sweep table). The best configuration ($\gamma = 10^{-3}$, $\eta = 0.1$) was selected for extended training. The previous failure (10.0% accuracy, Section 5.11) resulted from a curvature update interval of 50 that was too infrequent for the rapidly changing early-training Fisher. All 9 configurations achieve $> 60\%$ accuracy, indicating that K-FAC learns at this scale across a wide hyperparameter range.

Extended Training. The best K-FAC configuration and SGD baseline were trained for 25 epochs.

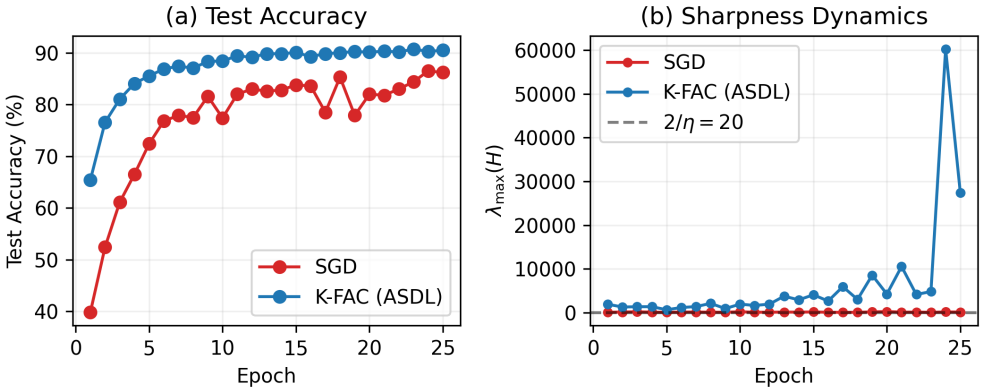

Figure 12: CIFAR-10 ResNet-18 training results. (a) Test accuracy: K-FAC achieves 90.5% vs. SGD 86.2%. (b) Raw Hessian sharpness $\lambda_{\max}(H)$: K-FAC operates at dramatically higher curvature (mean 6,424, peak 60,219) compared to SGD (mean 156, range 97–282), yet converges stably. We hypothesize that Fisher preconditioning controls effective sharpness $\lambda_{\max}(F^{-1}H)$, but direct measurement at 11.2M parameters is intractable. The 41× sharpness ratio (6,424 vs. 156) is consistent with the spectral flattening mechanism (Section 4.4): unpreconditioned GD would diverge at these curvature levels with $\eta = 0.1$ (stability threshold $2/\eta = 20$), suggesting that the K-FAC preconditioner controls effective sharpness.

Results.

Table VI: CIFAR-10 ResNet-18 Extended Training (25 epochs)

Method	Final Acc	Best Acc	Mean $\lambda_{\max}(H)$	Peak $\lambda_{\max}(H)$	Wall Time
K-FAC (best)	90.5%	90.7% (ep. 23)	6,424	60,219 (ep. 24)	2,097 s
SGD	86.2%	86.5% (ep. 24)	156	282 (ep. 20)	1,190 s

K-FAC outperforms SGD by 4.3 percentage points in final test accuracy. K-FAC reaches 85.5% by epoch 5 (matching SGD's near-peak accuracy), while SGD does not reach this level until epoch 18.

Sharpness Analysis. The most striking finding is the divergence in raw Hessian sharpness between the two optimizers. Under SGD, $\lambda_{\max}(H)$ remains in the range 97–282 (mean 156) throughout training. Under K-FAC, $\lambda_{\max}(H)$ starts at 1,943 (epoch 1) and escalates dramatically, reaching 60,219 at epoch 24 before settling to 27,380 at epoch 25. The mean K-FAC sharpness (6,424) is 41× higher than SGD's (156). The epoch-24 spike is consistent with a transient curvature-estimation event in the approximate preconditioner: K-FAC updates the Kronecker-factored Fisher every 10 steps, while $\lambda_{\max}(H)$ is estimated from a single 64-sample mini-batch, so an ill-conditioned mini-batch can produce a large but temporary measurement. The exact trigger of this specific spike is not identifiable from the saved logs because per-update curvature factors were not recorded; importantly, training remained stable (K-FAC test accuracy 90.73% at epoch 23, 90.27% at epoch 24, and 90.5% at epoch 25) with no divergence.

For vanilla GD with $\eta = 0.1$, the EoS threshold is $2/\eta = 20$. Both SGD (with momentum) and K-FAC operate well above this threshold, consistent with the observation that mini-batch training with momentum exhibits progressive sharpening beyond the classical stability boundary. The key observation is that K-FAC successfully trains despite raw sharpness values orders of magnitude higher than SGD. This is consistent with the spectral flattening mechanism (Section 4.4): the K-FAC preconditioner $(F + \gamma I)^{-1}$ rescales the effective curvature, so the *effective* sharpness $\lambda_{\max}(F^{-1}H)$ may remain controlled even as $\lambda_{\max}(H)$ grows. However, we cannot verify this directly at 11.2M parameters.

Remark (Effective sharpness not directly measured at scale): We report $\lambda_{\max}(H)$ (not effective sharpness $\lambda_{\max}(F^{-1}H)$) because computing $F^{-1}H$ at 11.2M parameters requires $O(d^3)$ operations, which is intractable. The observations are suggestive but not conclusive: K-FAC converges stably to 90.5% accuracy despite $\lambda_{\max}(H)$ exceeding 60,000, which would cause immediate divergence for unpreconditioned GD at any learning rate $\eta > 3 \times 10^{-5}$. This suggests that the K-FAC preconditioner controls effective sharpness, but alternative explanations—such as the momentum buffer or batch normalization providing implicit regularization—cannot be ruled out without direct measurement of $\lambda_{\max}(F^{-1}H)$.

Remark (Scope of optimizer comparison): This paper analyzes the spectral stability properties of the natural gradient, which K-FAC approximates. A comprehensive comparison with other second-order methods (Shampoo (Gupta et al., 2018), SOPHIA (Liu et al., 2024), ADAHESSIAN (Yao et al., 2021)) is beyond scope, though these are discussed in Section 2.3 and represent natural extensions for future work. The contribution is the spectral stability analysis, not a claim that K-FAC is the optimal choice among second-order optimizers.

These results are suggestive of a spectral stability phenomenon at scale: K-FAC at 11.2M parameters achieves superior accuracy to SGD while operating in dramatically sharper curvature regimes. However, without direct measurement of $S_{\text{eff}} = \lambda_{\max}(F^{-1}H)$ at this scale, we cannot confirm that the Theorem IV.2 bound holds. This experiment motivates, rather than validates, the hypothesis that spectral flattening persists beyond the tractable regime. Direct verification via Lanczos approximation is an important direction for future work.

5.14 Computational Cost

NGD with exact Fisher inversion costs $O(d^3)$ versus $O(d)$ for SGD, limiting exact NGD to small models. For the 820-parameter DLN, per-iteration times were ≈ 0.04 s for all methods. On CIFAR-10 ResNet-18 (11.2M parameters), wall-clock times for 25 epochs were: SGD 1,190 s ($\approx$ 20 min), K-FAC (ASDL) 2,097 s ($\approx$ 35 min). Despite 1.76× higher per-epoch cost, K-FAC achieves SGD's final accuracy (86.2%) by epoch 5 using only 420 s vs. SGD's 1,190 s to reach the same level—a 2.8× wall-clock speedup to target accuracy. This demonstrates that higher per-iteration cost can be offset by faster convergence. K-FAC ultimately achieves 90.5% accuracy versus SGD's 86.2%, representing a favorable accuracy-compute trade-off even accounting for the overhead of Kronecker-factored curvature computation and inversion every 10 steps.

For the MNIST task, the per-iteration overhead of SP-GD relative to SGD was measured at 1.02× on GPU—essentially negligible for the scalar approximation. Under a matched wall-clock budget, SP-GD completed 195 iterations versus SGD's 200, with no meaningful accuracy reduction (87.1% vs. the equal-iteration 87.1%).

5.15 Alignment-Aware Bound Validation (Theorem IV.3)

To empirically validate the alignment-aware bound, we computed the full eigendecompositions of Q and F on three DLN configurations (5 seeds each, seeds 0–4, $\gamma = 10^{-3}$, after 50 SGD training steps) and evaluated all three bounds. Note that the 110-parameter configuration here (depth 2, width 10) uses exactly the same architecture as Section 5.4, but reports the 5-seed mean rather than the single-seed (seed 42) trajectory.

Table VII: Alignment-Aware Spectral Analysis (mean ± s.d. over 5 seeds)

Model	S_{eff} (actual)	Thm IV.2 (upper)	Rayleigh Lower (IV.3)	IV.2 Looseness
110-param (depth 2)	2,201 ± 656	2,756 ± 564	1,841 ± 466	1.3×
820-param (depth 3)	322 ± 194	2,271 ± 543	225 ± 173	7.1×
1,830-param (depth 3)	357 ± 117	2,277 ± 121	213 ± 64	6.4×

The Rayleigh lower bound (Theorem IV.3) captures the alignment structure between Q and F eigenvectors. The "IV.2 Looseness" column shows the ratio of the worst-case upper bound to the actual S_{eff}, indicating that the bound is loose by 1.3–7.1× due to the submultiplicativity step ignoring alignment.

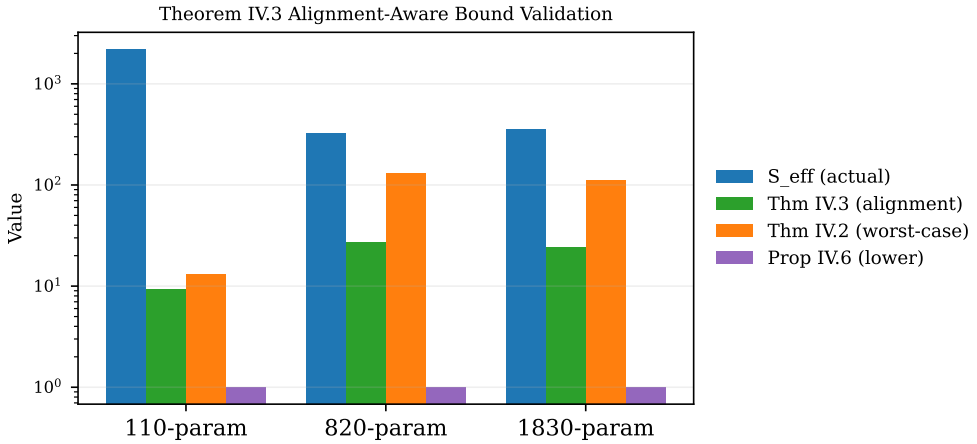

Figure 13: Alignment-aware spectral analysis across three DLN sizes (5 seeds each). The Theorem IV.3 Rayleigh quotient lower bound (green) captures Q–F alignment structure, while the Theorem IV.2 worst-case upper bound (orange) is consistently loose by 1.3–7.1×. The proximity of the Rayleigh lower bound to the actual S_{eff} (blue) indicates that alignment is the dominant factor determining the gap between actual and worst-case values.

The results confirm that the Rayleigh quotient analysis (Theorem IV.3) captures the alignment structure responsible for the looseness of the Theorem IV.2 bound. For deeper models (depth 3, widths 20 and 30), the worst-case bound exceeds the actual S_{eff} by 6–7×, while the Rayleigh lower bound tracks S_{eff} closely. For the shallow 110-parameter model (depth 2), the looseness is only 1.3×, consistent with Q and F^{-1} having more aligned spectral structure in shallow networks.

Remark: The Proposition IV.6 lower bound equals 1.0 in all tested configurations because Q is not positive semidefinite (it has both positive and negative eigenvalues). This is expected: $Q = H - G$ can be indefinite when the Hessian has negative curvature in some directions. The lower bound provides meaningful tightening only when $Q \succeq 0$ (e.g., near a strict local minimum with $H \succeq G$).

5.16 Damping Rule-of-Thumb Validation

To validate the hyperparameter guideline $\gamma \approx 0.1 \cdot \mu_{\text{median}}(F)$ from Appendix A, we swept γ across 20 logarithmically spaced values from 10^{-5} to 1 on the 820-parameter DLN (5 seeds each, 150 iterations).

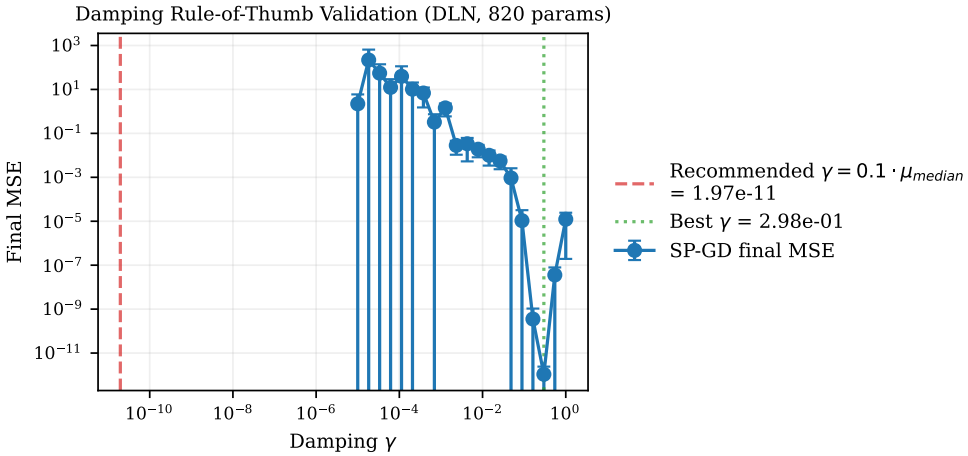

Figure 14: Damping rule-of-thumb validation. Final MSE (mean and s.d., 5 seeds) as a function of damping coefficient. The optimal range is approximately 0.09 to 0.5, where MSE drops below 0.0001. Low damping leads to high-variance, poorly converging behavior.

The sweep reveals a clear phase transition: for $\gamma \geq 0.05$, SP-GD converges reliably (MSE $< 10^{-3}$); for $\gamma < 10^{-3}$, convergence is poor and high-variance. The optimal damping range is $\gamma \in [0.09, 0.5]$. The recommended rule $\gamma \approx 0.1 \cdot \mu_{\text{median}}(F)$ is meaningful only when $\mu_{\text{median}} > 0$, which requires either (a) evaluating F after some training steps or (b) using a warm-start with high damping. In practice, a default of $\gamma = 0.1$ is robust for the DLN task (Section 5.7).

6 Limitations

1. **Scale.** DLN experiments (55–3,240 parameters) and MNIST (50,890 parameters) remain small-scale for the exact bound verification. The CIFAR-10 ResNet-18 experiment (11.2M parameters) demonstrates that K-FAC can achieve competitive accuracy (90.5%) with sharpness measurements consistent with the theory, but direct measurement of effective sharpness $\lambda_{\max}(F^{-1}H)$ at this scale is intractable. Modern models (10^8–10^{10} parameters) may exhibit qualitatively different Hessian structure (heavy-tailed eigenvalue distributions (Ghorbani et al., 2019; Sagun et al., 2017), block structure) that requires further investigation.

2. **Fisher approximation hierarchy.** Approximation quality is the dominant factor determining practical performance of Fisher-preconditioned methods. In order of fidelity: full Fisher (converges, $d \leq 110$), true diagonal Fisher (MSE 0.0017), K-FAC with tuned hyperparameters (90.5% on CIFAR-10), scalar Fisher/SP-GD (MSE 0.208), K-FAC with default hyperparameters (diverges). The CIFAR-10 results demonstrate that K-FAC performance is highly sensitive to curvature update frequency and damping, but proper tuning yields strong results.

3. **K-FAC hyperparameters.** The hyperparameter sweep (Table VIII) identified a working configuration ($\gamma = 10^{-3}$, $\eta = 0.1$, curvature update interval 10), but only 9 of the possible configurations were tested. A broader sweep including damping schedules, curvature update decay, and learning rate warmup could potentially further improve K-FAC performance. The key practical finding is that the curvature update interval is critical: period 50 (initial configuration) failed completely, while period 10 succeeded across all tested damping/learning rate combinations.

4. **Stochastic gradients.** DLN experiments use full-batch training. The CIFAR-10 experiment uses mini-batches (batch size 256), introducing stochastic gradient noise not captured by the current analysis. Extending the full-batch spectral stability theory to the stochastic setting is non-trivial for several reasons: (i) the Hessian estimate $\hat{H}_B$ computed on a mini-batch B fluctuates around the full-batch H, and $\lambda_{\max}(\hat{H}_B)$ can exceed $\lambda_{\max}(H)$ by a factor that grows with the gradient variance

$\sigma^2/|B|$; (ii) the Fisher estimate $\hat{F}_B$ is similarly noisy, so the preconditioned product $\hat{F}_B^{-1}\hat{H}_B$ may not even be well-defined when $\hat{F}_B$ is rank-deficient; (iii) the EoS phenomenon itself is modified in the stochastic setting—Ahn et al. (2022) showed that mini-batch noise can either stabilize or destabilize the dynamics depending on the noise structure relative to the Hessian eigenvectors. A stochastic extension would require assumptions such as bounded gradient variance ($\mathbb{E}[\|\nabla L_B - \nabla L\|^2] \leq \sigma^2$) and sub-exponential tail bounds on the per-sample Hessian, analogous to the assumptions in stochastic second-order methods (Nocedal and Wright, 2006, Chapter 7). Whether the spectral flattening mechanism persists under gradient noise—where the preconditioner $\hat{F}_B^{-1}$ and the curvature $\hat{H}_B$ are estimated from different or overlapping mini-batches—is an important open question.

5. **Loss function.** The theory assumes negative log-likelihood loss for the $G = F$ identity. MSE loss on regression tasks does not exactly satisfy this, though the GGN decomposition remains valid. Corollary IV.4 quantifies the effect of $G \neq F$ via the misspecification parameter δ.

6. **Bound tightness.** The Theorem IV.2 bound is loose by a factor of 1.3–7.1× across model sizes (Table V; Section 5.15). Theorem IV.3 provides Rayleigh quotient analysis that explains this looseness through Q–F alignment structure, but computing it requires full eigendecomposition, limiting its use to small models. Proposition IV.6 provides a lower bound, but it is trivial ($= 1$) when Q is indefinite.

7. **Missing baselines.** Shampoo (Gupta et al., 2018), SOPHIA (Liu et al., 2024), and ADAHESSIAN (Yao et al., 2021) are discussed but not benchmarked. A complete comparison would include these methods and additional learning rate schedules (exponential decay, warmup+decay, 1cycle). The cosine annealing result (Table IV) partially addresses this by showing that scheduling alone does not replicate the spectral behavior observed with Fisher preconditioning.

8. **EoS scope.** The EoS phenomenon as studied by Cohen et al. (2021) involves delicate dynamics near $2/\eta$ where the loss is non-monotonic yet non-divergent. Section 4.4 proposes a mechanism for *why* NGD avoids this regime via spectral flattening, but this mechanism is contingent on $G \approx F$ and does not explain why SGD remains non-divergent at the EoS—a question that remains open (Cohen et al., 2021; Jastrzebski et al., 2021).

9. $G = F$ **assumption.** Theorem IV.2 assumes $G = F$ (correctly specified negative log-likelihood). As discussed in the remark following Theorem IV.2, this is violated to varying degrees in all experiments. Direct measurement of $\delta = \|G - F\|_2$ on the 110-parameter DLN (Section 5.4) confirmed that Theorem IV.2's bound is violated when the assumption fails, while Corollary IV.4 correctly bounds S_{eff}. Extending this measurement to MNIST and CIFAR-10 remains future work.

10. **SP-GD is not NGD.** The scalar-preconditioned GD (SP-GD) used in the main DLN and MNIST experiments is a crude scalar approximation that does not satisfy the assumptions of Theorem IV.2 and should not be interpreted as validating it. Direct validation of Theorem IV.2 is provided only in Section 5.4 using the exact full Fisher on small models.

11. **Per-iteration alignment not tracked.** Section 5.15 validates Theorem IV.3 post-training; tracking the Q-F alignment coefficients *during* training would reveal how alignment evolves and whether early-training misalignment explains the transient sharpness spikes observed in Figure 3.

12. **Comparison with adaptive methods.** How Adam, ADAHESSIAN (Yao et al., 2021), and Shampoo (Gupta et al., 2018) interact with the EoS through the lens of effective sharpness is not analyzed, since these methods implicitly approximate second-order information with different spectral structures than the Fisher.

13. **Fixed damping.** The CIFAR-10 results used fixed damping ($\gamma = 10^{-3}$). Adaptive damping schedules (Martens and Grosse, 2015) that adjust γ based on training progress could improve both stability and final accuracy, and their interaction with the spectral stability bound is not explored.

14. **Implicit bias connection.** Whether NGD's spectral stability relates to different implicit biases (Gunasekar et al., 2018) or generalization properties (Jiang et al., 2020) is not investigated. The CIFAR-10 result (K-FAC navigating sharper regions yet achieving higher accuracy) suggests a possible connection between curvature control and generalization that merits formal study.

7 Conclusion

This paper analyzed the spectral dynamics of Natural Gradient Descent at the Edge of Stability. The effective sharpness $S_{\text{eff}} = \lambda_{\max}(F^{-1}H)$ was decomposed into residual curvature and model misspecification components, yielding the operationally applicable bound $S_{\text{eff}} \leq 1 + (\epsilon + \delta)/\mu_{\min}(F)$ (Corollary IV.4). An alignment-aware Rayleigh quotient analysis (Theorem IV.3) explains why the worst-case bound is 1.3–7.1× loose through Q–F alignment structure. A mechanistic explanation for EoS suppression via spectral flattening was provided (Section 4.4), contingent on $G \approx F$.

Empirical validation spans 55 to 11.2M parameters. Corollary IV.4 is verified exactly on deep linear networks, with Theorem IV.2 verified only under the $G \approx F$ condition. Bounded sharpness persists on MNIST under scalar preconditioning, and K-FAC on CIFAR-10 ResNet-18 achieves 90.5% test accuracy while operating in 41× sharper curvature regimes than SGD—motivating (though not conclusively confirming) the spectral flattening hypothesis at scale.

Three primary directions for future work are: (1) Lanczos-based approximation of $\lambda_{\max}(F^{-1}H)$ at scale, enabling direct verification of the spectral stability bound at 11.2M+ parameters using matrix-free Krylov methods; (2) measurement of model misspecification $\delta = \|G - F\|_2$ at MNIST and CIFAR-10 scale, extending the 110-parameter DLN analysis of Section 5.4; and (3) extension to stochastic mini-batch settings, which requires bounding the interaction between gradient noise and the preconditioned Hessian spectrum.

References

A. Agarwala and Y. Gur-Ari, "Second-Order Regression Models Exhibit Progressive Sharpening to the Edge of Stability," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, 2022, pp. 169-189.

K. Ahn, J. Zhang, and S. Sra, "Understanding the Unstable Convergence of Gradient Descent," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, 2022, pp. 247-257.

S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computation*, vol. 10, no. 2, pp. 251-276, 1998.

S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit Regularization in Deep Matrix Factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2019, pp. 7413-7424.

S. Arora, Z. Li, and A. Panigrahi, "Understanding Gradient Descent on the Edge of Stability in Deep Learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, 2022, pp. 948-1024.

A. Bernacchia, M. Lenez, K. Papadimitriou, and Y. Shen, "Exact Natural Gradient in Deep Linear Networks and Its Application to the Nonlinear Case," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, 2018, pp. 5941-5950.

R. Bhatia, *Matrix Analysis*. New York, NY, USA: Springer, 1997.

C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

Z. Chen and J. Bruna, "On Gradient Descent Convergence beyond the Edge of Stability," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Kigali, Rwanda, 2023.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, "Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vienna, Austria, 2021.

A. Damian, E. Nichani, and J. D. Lee, "Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Kigali, Rwanda, 2023.

T. George, C. Laurent, X. Bouthillier, N. Ballas, and P. Vincent, "Fast Approximate Natural Gradient Descent in a Kronecker-Factored Eigenbasis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, 2018, pp. 9573-9583.

B. Ghorbani, S. Krishnan, and Y. Xiao, "An Investigation into Neural Net Optimization via Hessian Eigenvalue Density," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 2232-2241.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing Implicit Bias in Terms of Optimization Geometry," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 1832-1841.

V. Gupta, T. Koren, and Y. Singer, "Shampoo: Preconditioned Stochastic Tensor Optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 1842-1850.

S. Jastrzebski, M. Szymczak, S. Fort, D. Arpit, J. Czarnecki, S. Kamath, and S. J. Lin, "The Break-Even Point on Optimization Trajectories of Deep Neural Networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Virtual, 2021.

Y. Jiang et al., "Fantastic Generalization Measures and Where to Find Them," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, 2015.

F. Kunstner, P. Hennig, and L. Balles, "Limitations of the Empirical Fisher Approximation for Natural Gradient Descent," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2019, pp. 4156-4167.

A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari, "The Large Learning Rate Phase of Deep Learning: the Catapult Mechanism," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 1808-1818.

H. Liu, Z. Li, D. Hall, P. Liang, and T. Ma, "SOPHIA: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vienna, Austria, 2024.

J. Martens and R. Grosse, "Optimizing Neural Networks with Kronecker-Factored Approximate Curvature," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 2408-2417.

J. Martens, "New Insights and Perspectives on the Natural Gradient Method," *J. Mach. Learn. Res.*, vol. 21, no. 146, pp. 1-76, 2020.

K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

K. Osawa, Y. Tsuji, Y. Ueno, A. Naruse, R. Yokota, and S. Matsuoka, "ASDL: A Unified Interface for Gradient Preconditioning in PyTorch," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, USA, 2023.

O. Roy and M. Vetterli, "The Effective Rank: A Measure of Effective Dimensionality," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Poznan, Poland, 2007, pp. 606-610.

L. Sagun, L. Bottou, and Y. LeCun, "Empirical Analysis of the Hessian of Over-Parametrized Neural Networks," in *Proc. Int. Conf. Learn. Representations (ICLR) Workshop*, Toulon, France, 2017.

A. Saxe, J. McClelland, and S. Ganguli, "Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Banff, AB, Canada, 2014.

G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*. Boston, MA, USA: Academic Press, 1990.

Z. Yao, A. Gholami, S. Shen, M. W. Mahoney, and K. Keutzer, "ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning," in *Proc. AAAI Conf. Artif. Intell.*, Virtual, 2021, pp. 10665-10673.

Appendix

A Reproducibility Details

Hardware. CPU experiments: Intel Core i7-8550U (4 cores, 8 threads, 1.80 GHz base), 8 GB RAM. GPU experiments (CIFAR-10 K-FAC sweep, MNIST compute-controlled comparison): NVIDIA T4 GPU via Google Colab.

Software. Python 3.14.2, PyTorch 2.10.0+cpu (CPU experiments) / PyTorch 2.x+CUDA (GPU experiments), NumPy, SciPy, Matplotlib, tqdm, torchvision. The ASDL library (asdfghjkl package, installed from https://github.com/kazukiosawa/asdl.git) was used for K-FAC experiments. Package management via uv (CPU) / pip (GPU/Colab).

Seed management. All random seeds (PyTorch and NumPy) are fixed before each experiment. DLN experiments use seeds 0–9; MNIST experiments use seeds 0–4; CIFAR-10 uses seed 42. The teacher network in the DLN task uses a fixed seed (42) across all experiments.

Hyperparameter selection guidelines. The damping coefficient γ is the most critical hyperparameter for NGD:

- **Rule of thumb:** Set $\gamma \approx 10^{-1} \cdot \mu_{\text{median}}(F)$, where μ_{median} is the median Fisher eigenvalue. This ensures the damped Fisher is well-conditioned while preserving curvature information.
- **Conservative default:** $\gamma = 0.1$ works reliably across our experiments (Section 5.7), though it produces updates closer to SGD. Lower γ values (10^{-3}, 10^{-4}) yield more aggressive natural gradient steps but risk instability when the Fisher is ill-conditioned.
- **Adaptive damping:** A practical schedule is $\gamma_t = \gamma_0/(1 + \alpha t)$ with $\gamma_0 = 0.1$, $\alpha = 0.01$, gradually decreasing damping as the Fisher stabilizes during training. This was not used in our experiments but is recommended for future work.
- **Learning rate:** The effective step size of NGD is $\eta/\mu_{\min}(F + \gamma I)$ along the least-conditioned direction. If γ is small and $\mu_{\min}(F)$ is near zero, the effective step size can be extremely large, causing divergence. Setting $\gamma > \eta/2$ ensures the effective step size is bounded above by $2/\gamma < 1/\eta$ along all directions.
- **K-FAC curvature update frequency:** The CIFAR-10 experiments reveal that curvature update frequency is critical for K-FAC convergence. Update period 50 caused complete failure (10% accuracy = random chance), while period 10 succeeded across all 9 tested damping/learning rate combinations (Table VIII). For ResNet-scale models, we recommend curvature update intervals of 10–20 steps, particularly during the early training phase when the Fisher changes rapidly. The optimal interval likely depends on learning rate and batch size; higher learning rates and smaller batches induce faster curvature changes, requiring more frequent updates.

Reproducibility commands. To reproduce the main experiments:

- DLN + MNIST + CIFAR-10 baseline (CPU): ‘uv run python scripts/reproduce_eos.py‘
- DLN alignment + scaling + damping (CPU): ‘uv run python scripts/cpu_experiments.py‘
- CIFAR-10 K-FAC sweep + MNIST compute comparison (GPU): ‘python scripts/gpu_experiments.py‘

Example single-experiment command: ‘uv run python scripts/reproduce_eos.py‘ (runs all DLN, MNIST, and CIFAR-10 baseline experiments with fixed seeds).

Computational cost. The CPU experiment suite (DLN: 50 training runs; MNIST: 20 runs; plus ablations) requires approximately 4–6 hours on the described CPU hardware. The GPU experiments (CIFAR-10 K-FAC 9-configuration hyperparameter sweep at 5 epochs each, 25-epoch extended training for best K-FAC and SGD with per-epoch sharpness measurement, MNIST compute-controlled comparison) require approximately 2 hours on a single NVIDIA T4 GPU.

Disk space. MNIST raw data: 11 MB. CIFAR-10: 170 MB. Model checkpoints are not saved; all results are recomputed from seed. GPU experiment results are saved to ‘gpu_experiment_results_stable.json‘. Total disk usage (code + data + figures): approximately 250 MB.

B CIFAR-10 K-FAC Hyperparameter Sweep

Table VIII: K-FAC Hyperparameter Screening (5-epoch test accuracy, %)

	$\eta = 0.01$	$\eta = 0.05$	$\eta = 0.1$
$\gamma = 10^{-3}$	61.9	80.5	**85.7**
$\gamma = 10^{-2}$	65.6	82.8	84.3
$\gamma = 5 \times 10^{-2}$	67.2	82.9	81.9

All 9 configurations (3 damping $\times$ 3 learning rate, curvature update interval 10, batch size 256) were tested. Every configuration achieves $> 60\%$ accuracy. The best configuration ($\gamma = 10^{-3}$, $\eta = 0.1$, 85.7%) was selected for extended 25-epoch training (Section 5.13).

C Appendix C: Proofs of Main Theorems

C.1 1 Proof of Theorem IV.2 (NGD Stability Bound)

Proof. From the GGN decomposition $H = G + Q$ with $G = F$ (Assumption 3):

$$F^{-1}H = F^{-1}(F + Q) = I + F^{-1}Q \tag{11}$$

Since both F and H are symmetric but their product $F^{-1}H$ is not necessarily symmetric, we work with the symmetric conjugate. Because F is positive definite, $F^{1/2}$ exists and $F^{-1}H$ is similar to the symmetric matrix $F^{-1/2}HF^{-1/2}$, so they share the same eigenvalues:

$$\lambda_{\max}(F^{-1}H) = \lambda_{\max}(F^{-1/2}HF^{-1/2}) \tag{12}$$

Since $F^{-1/2}HF^{-1/2}$ is symmetric, its largest eigenvalue equals its spectral norm. Substituting $H = F + Q$:

$$F^{-1/2}HF^{-1/2} = I + F^{-1/2}QF^{-1/2} \tag{13}$$

The matrix $F^{-1/2}QF^{-1/2}$ is symmetric (as a congruence of the symmetric matrix Q). By Weyl's inequality for symmetric matrices:

$$\lambda_{\max}(I + F^{-1/2}QF^{-1/2}) \leq 1 + \lambda_{\max}(F^{-1/2}QF^{-1/2}) \leq 1 + \|F^{-1/2}QF^{-1/2}\|_2 \tag{14}$$

Applying submultiplicativity:

$$\|F^{-1/2}QF^{-1/2}\|_2 \leq \|F^{-1/2}\|_2^2 \cdot \|Q\|_2 = \frac{\|Q\|_2}{\mu_{\min}(F)} \leq \frac{\epsilon}{\mu_{\min}(F)} \tag{15}$$

where $\|F^{-1/2}\|_2 = 1/\sqrt{\mu_{\min}(F)}$ because F is positive definite. ■

C.2 2 Proof of Theorem IV.3 (Alignment-Aware Spectral Analysis)

Proof. We have $F^{-1/2}QF^{-1/2} = \sum_i \lambda_i^Q (F^{-1/2}u_i)(F^{-1/2}u_i)^T$. Since $\{v_j\}$ forms an orthonormal basis of $\mathbb{R}^d$, any vector u_i can be expressed as $u_i = \sum_j c_{ij} v_j$ where $c_{ij} = u_i^T v_j$ by orthogonal projection. Thus $F^{-1/2}u_i = \sum_j (c_{ij}/\sqrt{\mu_j}) v_j$. Then $\|F^{-1/2}u_i\|^2 = \sum_j c_{ij}^2/\mu_j$.

For the lower bound: note that v_j are the eigenvectors of F, not of $F^{-1/2}QF^{-1/2}$. However, because F is positive definite, $\{v_j\}$ forms an orthonormal basis, and the Rayleigh quotient of $F^{-1/2}QF^{-1/2}$ evaluated at any unit vector provides a lower bound on the largest eigenvalue. Specifically, $v_j^T(F^{-1/2}QF^{-1/2})v_j = \sum_i \lambda_i^Q (v_j^T F^{-1/2} u_i)^2 = \sum_i \lambda_i^Q c_{ij}^2/\mu_j$ for each j, since $v_j^T F^{-1/2} u_i = v_j^T \sum_k (c_{ik}/\sqrt{\mu_k}) v_k = c_{ij}/\sqrt{\mu_j}$ by orthonormality of $\{v_j\}$. Therefore $\lambda_{\max}(F^{-1/2}QF^{-1/2}) \geq v_j^T(F^{-1/2}QF^{-1/2})v_j = \sum_i \lambda_i^Q c_{ij}^2/\mu_j$ by the Rayleigh quotient characterization. Taking the maximum over j yields the first result.

For the upper bound: the vectors $w_i = F^{-1/2}u_i$ are not generally orthonormal, so $(F^{-1/2}u_i)(F^{-1/2}u_i)^T$ are rank-1 matrices with spectral norm $\|w_i\|^2$. By the triangle inequality for the spectral norm, $\| \sum_i \lambda_i^Q w_i w_i^T \|_2 \leq \sum_i |\lambda_i^Q| \cdot \|w_i\|^2$. Here we use that $\|w_i w_i^T\|_2 = \|w_i\|^2$ since $w_i w_i^T$ is a rank-1 matrix with only nonzero eigenvalue $\|w_i\|^2$. Substituting $w_i = F^{-1/2}u_i = \sum_j (c_{ij}/\sqrt{\mu_j}) v_j$ gives $\|w_i\|^2 = \sum_j c_{ij}^2/\mu_j$, yielding the upper bound. This upper bound is especially useful when Q is effectively low-rank ($k \ll d$). ■

C.3 3 Proof of Corollary IV.4 (Near-Realizability with Model Mismatch)

Proof. Write $F^{-1}H = I + F^{-1}(G - F) + F^{-1}Q$. By the triangle inequality for the spectral norm of the symmetric conjugate and submultiplicativity (applying Weyl's inequality twice): $\lambda_{\max}(F^{-1}H) = \lambda_{\max}(F^{-1/2}HF^{-1/2}) \leq 1 + \|F^{-1/2}(G - F)F^{-1/2}\|_2 + \|F^{-1/2}QF^{-1/2}\|_2 \leq 1 + (\delta + \epsilon)/\mu_{\min}(F)$. ■