Adam or Gauss-Newton?
A Comparative Study In Terms of Basis Alignment and SGD Noise

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Diagonal preconditioners are computationally tractable approximations to second-order optimizers and have been central to the recent push for faster training of deep models. Two predominant families are based on Adam and on the Gauss-Newton (GN) matrix: Adam tracks running statistics of gradients, while GN-based methods (e.g. Sophia) use the diagonal of the Gauss-Newton matrix. We compare these two families through the lens of two factors: the *basis* in which the diagonal preconditioner acts and the impact of *gradient noise* from mini-batching. Using linear and logistic regression as analytic testbeds, we show that (i) regardless of basis, there exist instances where Adam outperforms both GN^{-1} and $\text{GN}^{-1/2}$ in the full-batch regime—even in the GN eigenbasis for logistic regression—and (ii) in the stochastic regime, Adam behaves equivalently to $\text{GN}^{-1/2}$ for linear regression under Gaussian inputs. These predictions are corroborated by experiments on convex and non-convex problems including MLPs, CIFAR-10, and Transformers.

1. Introduction

Modern deep learning has shifted from vanilla (stochastic) gradient descent toward adaptive first-order optimizers with preconditioned updates of the form $\theta^{(t+1)} = \theta^{(t)} - \eta P g^{(t)}$, where $P \in \mathbb{R}^{d\times d}$ is often diagonal. Methods such as Adam [15], RMSProp [27], Adafactor [26], SignSGD [7], and Lion [10] build their preconditioners from empirical *gradient* statistics rather than true curvature. Recently, optimizers such as Sophia [19] have revisited the curvature connection by using a diagonal of the Gauss-Newton (GN) matrix. A complementary line of work [29] interprets non-diagonal preconditioners such as Shampoo [14] as a *diagonal preconditioner in a transformed basis*.

This naturally raises a question central to this work: *can we disentangle the role of the basis used for preconditioning from the diagonal scaling applied within that basis?* We compare two canonical diagonal scalings—one based on running-average squared gradients (Adam, the diagonal of the empirical Fisher [17]), and one based on the diagonal of the Gauss-Newton matrix—across two bases (the identity basis used by Adam and the GN eigenbasis used by curvature-aware methods). Since empirical-gradient methods effectively use a square root, we additionally consider both GN^{-1} and $\text{GN}^{-1/2}$. Under this lens we ask: *does the empirical Fisher (used by Adam, SOAP) offer real advantages over GN-derived curvature, or merely serve as a tractable proxy?*

Contributions. We show the comparison hinges on two axes: **(1) basis choice**—identity vs. the GN eigenbasis; **(2) gradient noise**—full-batch (population) vs. stochastic (batch size 1). Our results, summarized in Table 1, are:

- **Sensitivity to basis (§3,4).** For linear regression, GN preconditioning in the eigenbasis is well known to be optimal; under a misaligned (identity) basis, however, Adam can outperform both GN^{-1} and $\text{GN}^{-1/2}$. For logistic regression, Adam can outperform GN^{-1} *even under the GN eigenbasis* with full batches.
- **Equivalence under noise (§3.2).** For linear regression in the stochastic regime with Gaussian inputs, Adam behaves equivalently to $\text{GN}^{-1/2}$ regardless of basis—a surprising alignment between Adam's empirical design and curvature-based preconditioning.

We complement these findings with experiments on toy and real problems, including CIFAR-10 and Transformers (§C).

Table 1. Comparison of Adam vs. GN diagonal preconditioners across two axes: (i) basis and (ii) gradient noise. Theoretical results are based on quadratics (§3) and logistic regression (§4).

	Batch-size regime	
Basis	**Full batch**	**Small batch**
Eigenbasis	$\exists$ logistic example where Adam $>$ GN^{-1}	$\text{GN}^{-1} \geq$ Adam $\approx \text{GN}^{-1/2}$ (quadratics)
Identity	$\exists$ quadratic example where Adam $>$ GN^{-1}	Adam $\approx \text{GN}^{-1/2}$ (quadratics)

2. Preliminaries

Consider optimizing $f : \mathbb{R}^d \to \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$ under loss ℓ. We study preconditioned updates of the form

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \, (U \, D^p \, U^\top) \, g^{(t)},$$

which factorizes any preconditioner into two ingredients: an orthonormal *basis* U in which the gradient is rotated, and a *diagonal* preconditioner D^p applied within that basis. The exponent $p \in \{-\frac{1}{2}, -1\}$ controls how aggressively curvature is exploited; $g^{(t)}$ is the gradient at $\theta^{(t)}$ (population or stochastic).

The Hessian decomposes as $H = H^{(\text{GN})} + \nabla_f \ell \, \nabla_\theta^2 f$, where $H^{(\text{GN})} := \nabla_\theta f \, \nabla_f^2 \ell \, \nabla_\theta f^\top$ is PSD for convex losses and only needs first-order information of the network [25]; for squared loss, $H^{(\text{GN})} = \mathbb{E}[\nabla_\theta f \, \nabla_\theta f^\top]$. We use $H^{(\text{GN})}$ to define both a basis (its eigenbasis) and a diagonal preconditioner (its diagonal entries). The two basis choices we compare are the identity $U = \mathbf{I}$ (Adam's default) and the GN eigenbasis (the curvature-aligned target of Sophia/Shampoo). For experiments we additionally use the Kronecker-factored approximation [20, 29] as a tractable surrogate for the full eigenbasis (Section C).

Diagonal preconditioners. Given basis U, rotate the gradient $\tilde{g} := U^\top g$ and apply a diagonal D. We study

$$D_{ii}^{(A)} := \left(\mathbb{E}[(u_i^\top g(x))^2]\right)^{-1/2}, \qquad D_{ii}^{(\mathrm{GN})} := (u_i^\top H^{(\mathrm{GN})} u_i)^p. \tag{1}$$

The first matches Adam's running average of squared gradients (modulo $\beta_1 = \beta_2$); the second uses GN's diagonal curvature estimate. In the full-batch limit $D_{ii}^{(A)}$ approaches the rotated mean-gradient norm (giving rise to the *auto-tuning* effect of §3.1); in the stochastic limit it is dominated by per-sample gradient norms. At batch size 1 for cross-entropy / MSE losses, $D^{(A)}$ and $D^{(\mathrm{GN})}$ correspond, respectively, to the empirical Fisher and the Fisher matrices [17].

3. Linear regression: basis choice and gradient noise

We study the workhorse setting $f_\theta(x) = \theta^\top x$ with $\ell(\theta) = \frac{1}{2}\mathbb{E}[((\theta - \theta^*)^\top x)^2]$ and $x \sim \mathcal{N}(0, \Sigma_x)$. Here $H^{(\mathrm{GN})} = \mathbb{E}[xx^\top] = \Sigma_x$, so the GN eigenbasis is the input-covariance eigenbasis and the GN diagonal is $\mathrm{diag}(\Sigma_x)$. Vanilla gradient descent on this objective evolves as $\Delta^{(t+1)} := \theta^{(t+1)} - \theta^* = (\mathbf{I} - \eta\,\Sigma_x)\,\Delta^{(t)}$. Under the eigenbasis, GN^{-1} is well known to be optimal: it converges in a single step at full batch and at the optimal linear rate stochastically (Section A.1). Below, we examine the three remaining quadrants in Table 1.

3.1. Full batch, identity basis: Adam "auto-tunes" to curvature

Heterogeneous curvature is a textbook motivation for preconditioning [13, 19, 24, 30, 31], and GN^{-1} in the eigenbasis is the textbook answer. But what happens when the basis is wrong? We construct an example where the misalignment hides the curvature from GN entirely, while Adam still adapts.

Consider the block-structured covariance

$$\Sigma_x = \begin{bmatrix} \mathbf{1}\mathbf{1}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \in \mathbb{R}^{2d \times 2d}, \tag{2}$$

where $\mathbf{1} \in \mathbb{R}^d$. The first block has top eigenvalue d (max stable LR $2/d$); the second has eigenvalues 1 (LR up to 2). The two blocks need very different step sizes, but each is symmetric within itself.

GN reduces to GD under the identity basis. Every diagonal entry of Σ_x equals 1, so under $U = \mathbf{I}$, $D_{ii}^{(\mathrm{GN})} = 1^p = 1$ uniformly. Both GN^{-1} and $\mathrm{GN}^{-1/2}$ degenerate to vanilla GD, and convergence is bottlenecked by the first block's curvature: any safe LR is $O(1/d)$, so reaching error ϵ requires $\Omega(d \log(1/\epsilon))$ steps.

Adam auto-tunes per-block. By symmetry, the first block sees a single rotated-gradient norm $\|g_0^{(t)}\|$ and the second block has independent per-coordinate norms $|g_i^{(t)}|$, giving updates $\Delta_0^{(t+1)} = (\mathbf{I} - \frac{\eta}{\|g_0^{(t)}\|}\Sigma_x)\Delta_0^{(t)}$ and $\Delta_i^{(t+1)} = (\mathbf{I} - \frac{\eta}{|g_i^{(t)}|}\Sigma_x)\Delta_i^{(t)}$. After a short burn-in, $\eta/\|g_0^{(t)}\|$ stabilizes near $2/d$ and $\eta/|g_i^{(t)}|$ near 2, simultaneously. Halving η thereafter reaches error ϵ in $O(\log(1/\epsilon))$ steps—a d-fold improvement over GN. This *auto-tuning* mirrors normalized gradient descent [22] and arises from the *mean* gradient in Adam's denominator at full batch; it vanishes stochastically,

where the denominator is dominated by gradient variance [23]. **Empirically:** on a 100-dim instance of (2) under the identity basis, Adam's loss drops by orders of magnitude in tens of steps while GN^{-1} tracks vanilla GD (Figure 2 (a); Section C.1).

Which power for GN? Even when GN preconditions in the wrong basis, the choice between $p = -1$ and $p = -1/2$ matters. The rate of preconditioned GD [9] depends on $\kappa(P^{1/2}\Sigma_x P^{1/2})$, and there exist Σ_x for which the half-power has the more favorable conditioning (Section A.3). Simulations on a 5-dim regression confirm both directions of the comparison: $\text{GN}^{-1/2} > \text{GN}^{-1}$ on a Σ_x engineered for it (Figure 2 (c)), and the opposite on its dual (Figure 1; Section A.3).

3.2. Stochastic regime: Adam $\approx \text{GN}^{-1/2}$ in any basis

We now ask the dual question: how does basis choice matter once gradient noise dominates? At single-sample batches, Adam's denominator no longer captures the population mean but the per-sample gradient norm. The next lemma shows that, under Gaussian inputs, this is exactly the Fisher up to a loss-dependent rescaling.

Lemma 1 *For linear regression with Gaussian inputs,* $\ell(\theta)\cdot\Sigma_x \preceq \frac{1}{2}\,\mathbb{E}[g(x)\,g(x)^\top] \preceq 3\,\ell(\theta)\cdot\Sigma_x$.

The empirical Fisher (Adam's target) is therefore equivalent to the Fisher (GN's target) up to a $\sqrt{\ell(\theta)}$ rescaling that is the same on every coordinate. Diagonalizing both sides in any basis U gives the per-step equivalence:

Corollary 2 *For single-sample updates, Adam and* $\text{GN}^{-1/2}$ *differ by a constant:* $\frac{1}{\sqrt{3\ell}}\,D^{(GN,-\frac{1}{2})} \preceq \frac{1}{2}\,D^{(A)} \preceq \frac{1}{\sqrt{\ell}}\,D^{(GN,-\frac{1}{2})}$.

A loss-dependent overall rescaling can always be absorbed into a learning-rate schedule, so this says Adam and $\text{GN}^{-1/2}$ are essentially interchangeable preconditioners under noise—in *any* orthonormal basis. To compare $\text{GN}^{-1/2}$ to its sister GN^{-1} in the same regime, we need a stochastic convergence rate. Define $\boldsymbol{A}(P) := P^{1/2}\Sigma_x P^{1/2}$.

Lemma 3 *For linear regression with stochastic Gaussian inputs and any PSD preconditioner* P,
$\mathbb{E}[\ell^{(t)}] \le O\left[\left(1 - \frac{\lambda_{\min}(\boldsymbol{A}(P))}{3\,\operatorname{Tr}(\boldsymbol{A}(P))}\right)^t \ell^{(0)}\right]$.

The bound depends on a stochastic-regime "condition number" $\kappa_s(\boldsymbol{A}(P)) := \operatorname{Tr}(\boldsymbol{A}(P))/\lambda_{\min}(\boldsymbol{A}(P)) \ge d$. In the eigenbasis, GN^{-1} minimizes this to d; $\text{GN}^{-1/2}$ achieves $\sum_i \sqrt{\lambda_i(\Sigma_x)/\lambda_{\min}(\Sigma_x)} \ge d$, with equality only when Σ_x is isotropic. So GN^{-1} remains optimal in the noise regime, with $\text{GN}^{-1/2}$ and (by Corollary 2) Adam matching it up to the spread of the spectrum. Proofs in Section A.2.

Experimental confirmation. The small-batch quadrants of the random-teacher MLP, sparse parity, staircase, and CIFAR-10 experiments (Figure 3 (a), Figures 4 to 6 in Section C.3) all show Adam and $\text{GN}^{-1/2}$ tracking each other in either basis, exactly as Corollary 2 predicts.

4. Logistic regression: Adam beats GN^{-1} even in the eigenbasis

The optimality of GN^{-1} in the eigenbasis is special to quadratics. Once non-convexity enters, GN^{-1}'s aggressive scaling can amplify rather than dampen the dynamics. We construct a logistic-regression example where, under full batches and the GN eigenbasis, Adam outperforms GN^{-1}.

Setup. Let x be uniform over $\{e_i\}_{i=1}^d$ with $\nu_i := \mathbf{Pr}(x = e_i)$, and $y \,|\, x = e_i \sim \mathrm{Bernoulli}(P_i)$ with $P_i \in [0.6, 0.8]$ (labels neither deterministic nor purely random; θ^* bounded). Consider the weight-tied two-layer linear network $q_i(\theta) := \mathbf{Pr}_\theta(y = 1 \mid x = e_i) = \sigma(\sum_j \theta_j^2 x_j) = \sigma(\theta_i^2)$, $\sigma(z) = 1/(1 + e^{-z})$. The squared parameterization is non-convex and structurally analogous to the key×query product in self-attention [28]. Define $\kappa(\nu) = \nu_{\max}/\nu_{\min}$ and target error ϵ. The eigenvectors of $H^{(\mathrm{GN})}(\theta)$ are the standard basis, so the identity and the GN eigenbasis coincide; both algorithms act coordinate-wise, and we analyze *local* convergence near θ^*.

Adam: $O(\log(1/\epsilon))$ steps. Adam reduces to sign-GD with per-coordinate step size η. Starting at $O(1)$ and halving every $O(1)$ steps yields $O(\log(1/\epsilon))$ local convergence—no dimension dependence.

GN^{-1}: $\widetilde{\Omega}(d^\delta \log(1/\epsilon))$ steps. The GN^{-1} update has a built-in tension: a small η keeps the iterate from blowing up far from θ^* (where $H^{(\mathrm{GN})}(\theta)$ vanishes), but a small $\eta^\infty := \lim_t \eta_t$ slows local convergence. Linearizing at θ^* with $H_*^{(\mathrm{GN})} := H^{(\mathrm{GN})}(\theta^*)$, the local contraction factor $\gamma(\eta^\infty, \alpha) := \|\mathbf{I} - \eta^\infty (H_*^{(\mathrm{GN})} + \alpha \mathbf{I})^{-1} H_*^{(\mathrm{GN})}\|_2$ governs the asymptotic rate.

Theorem 4 (Logistic lower bound) *Initialize $\theta^{(0)} = \frac{1}{\sqrt{d}}\mathbf{1}$. For any non-increasing $\{\eta_t\}$ and $\alpha \geq 0$, if $\theta^{(t)} \to \theta^*$, then $\gamma(\eta^\infty, \alpha) \geq 1 - c\sqrt{\log d}\, \max\left\{\frac{1}{\sqrt{d}}, \frac{\sqrt{d}}{\kappa(\nu)}\right\}$ for a universal constant c.*

Corollary 5 *For $\kappa(\nu) = \Omega(d^{1/2+\delta})$, $\delta \in [0, 1/2]$, GN^{-1} requires $t = \widetilde{\Omega}(d^\delta \log(1/\epsilon))$ local steps to reach $\|\theta^{(t)} - \theta^*\|_2 \leq \epsilon$.*

So GN^{-1} has a polynomial-in-d slowdown in the eigenbasis, exacted by the requirement to converge globally from $O(1/\sqrt{d})$ initialization. This is consistent with Abreu et al. [4]'s empirical findings for line-search GN, which violates the non-increasing-η assumption. Full proof in Section B. **Empirically:** a 2048-dim simulation confirms Adam $>$ GN^{-1} in the eigenbasis (Figure 2 (b)), and the same separation appears on a 1-layer Transformer whose attention mirrors this squared parameterization (Figure 3 (b); Section C.5).

5. Discussion

Our two-axis decomposition (Table 1) shows that Adam can outperform GN in the identity basis (linear regression) and even in the eigenbasis (logistic), while in the stochastic regime Adam aligns with $\mathrm{GN}^{-1/2}$ regardless of basis. The experiments inlined above are detailed in Section C, including additional non-convex MLPs (sparse parity, staircase, CIFAR-10) and a basis-interpolation sweep. The small-batch equivalence between Adam and $\mathrm{GN}^{-1/2}$ likely extends to large-scale training, where gradient variance dominates the mean; combined with Adam's auto-tuning at large batches, this suggests a promising direction: optimizers that retain auto-tuning even at small batches.

References

[1] E. Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. *Annual Conference Computational Learning Theory*, 2023. doi: 10.48550/arXiv.2302.11055.

[2] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. *arXiv preprint arXiv: 2202.08658*, 2022.

[3] Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31–60. PMLR, 2023. URL https://proceedings.mlr.press/v202/abbe23a.html.

[4] Natalie Abreu, Nikhil Vyas, Sham Kakade, and Depen Morwani. The potential of second-order optimization for llms: A study with full gauss-newton. *arXiv preprint arXiv: 2510.09378*, 2025.

[5] Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/884baf65392170763b27c914087bde01-Abstract-Conference.html.

[6] Frederik Benzing. Gradient descent on neurons and its link to approximate second-order optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1817–1853. PMLR, 2022. URL https://proceedings.mlr.press/v162/benzing22a.html.

[7] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SIGNSGD: compressed optimisation for non-convex problems. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 559–568. PMLR, 2018. URL http://proceedings.mlr.press/v80/bernstein18a.html.

[8] Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in transformers and their ability to learn sparse Boolean functions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5767–5791, Toronto,

Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.317. URL https://aclanthology.org/2023.acl-long.317.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9a39b4925e35cf447ccba8757137d84f-Abstract-Conference.html.

[11] Rudrajit Das, Naman Agarwal, Sujay Sanghavi, and Inderjit S. Dhillon. Towards quantifying the preconditioning effect of adam. *arXiv preprint arXiv: 2402.07114*, 2024.

[12] Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Pareto frontiers in neural feature learning: Data, compute, width, and luck. *arXiv preprint arXiv: 2309.03800*, 2023.

[13] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 2019. URL http://proceedings.mlr.press/v97/ghorbani19b.html.

[14] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1837–1845. PMLR, 2018. URL http://proceedings.mlr.press/v80/gupta18a.html.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

[16] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 and cifar-10 (canadian institute for advanced research), 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

[17] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information*

Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 4158–4169, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/46a558d97954d0692411c861cf78ef79-Abstract.html.

[18] Wu Lin, Felix Dangel, Runa Eschenhagen, Juhan Bae, Richard E. Turner, and Alireza Makhzani. Can we remove the square-root in adaptive gradient methods? A second-order perspective. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=vuMD71R20q.

[19] Hong Liu, Zhiyuan Li, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3xHDeA8Noi.

[20] James Martens and Roger B. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2408–2417. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/martens15.html.

[21] Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham M. Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=i9wDX850jR.

[22] Francesco Orabona. Normalized gradients for all, 2023. URL https://arxiv.org/abs/2308.05621.

[23] Vincent Roulet and Atish Agarwala. Per-example gradients: a new frontier for understanding and improving optimizers. *arXiv preprint arXiv: 2510.00236*, 2025.

[24] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv: 1611.07476*, 2016.

[25] Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N. Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 9481–9488. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17142.

[26] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July*

10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR, 2018. URL http://proceedings.mlr.press/v80/shazeer18a.html.

[27] Tijmen Tieleman and Geoffrey E. Hinton. Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning. Technical report, University of Toronto, 2012. Available at: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.5.pdf.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[29] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv: 2409.11321*, 2024.

[30] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. Pyhessian: Neural networks through the lens of the hessian. *arXiv preprint arXiv: 1912.07145*, 2019.

[31] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b05b57f6add810d3b7490866d74c0053-Abstract.html.

Contents

Appendix A. Theoretical results and omitted proofs

A.1. Optimality of GN^{-1} in the correct basis

For completeness, we provide proofs for the optimality of GN^{-1} for the quadratic loss under the correct eigenbasis.

Full batch. For GN^{-1}, the parameter estimation error evolves as

$$\begin{aligned}\theta^{(1)} - \theta^* &= (\theta^{(0)} - \theta^*) - \eta (H^{(\mathrm{GN})})^{-1} g^{(0)} \\ &= \theta^{(0)} - \eta \mathbb{E}[xx^\top]^{-1} \mathbb{E}[xx^\top(\theta - \theta^*)] = (1-\eta)(\theta^{(0)} - \theta^*).\end{aligned} \tag{3}$$

Hence GN^{-1} reaches the optimum in one step with $\eta = 1$.

Stochastic regime. Consider single-sample updates. Define $M^{(t)} := \mathbb{E}[(\theta^{(t)} - \theta^*)(\theta^{(t)} - \theta^*)^\top]$, the second moment of the distance to optimum. The GN^{-1} update with $g = (\theta - \theta^*)^\top x \cdot x$ gives

$$M^{(t+1)} = M^{(t)} - 2\eta\, M^{(t)} + \eta^2 \operatorname{Tr}(M^{(t)} \Sigma_x)\, \Sigma_x^{-1} + 2\eta^2 M^{(t)}. \tag{4}$$

Multiplying by Σ_x and taking the trace,

$$\mathbb{E}[\ell^{(t+1)}] = \big(1 - 2\eta + \eta^2(d+2)\big)\,\mathbb{E}[\ell^{(t)}]. \tag{5}$$

Setting $\eta = 1/(d+2)$ halves the expected error every $O(d)$ steps.

A.2. Equivalence of Adam and GN$^{-1/2}$ in the stochastic regime

We start with the equivalence between empirical and true Fisher (Lemma 1), used to derive the per-step equivalence (Corollary 2) and the loss-rate bound (Lemma 3).

A.2.1. PROOF OF LEMMA 1

Proof For a given θ, set $M := (\theta - \theta^*)(\theta - \theta^*)^\top$. Then

$$\mathbb{E}[g(x)g(x)^\top] = \mathbb{E}[x^\top M x \cdot xx^\top] = 2\Sigma_x M \Sigma_x + \operatorname{Tr}(\Sigma_x M)\,\Sigma_x \preceq 3\,\operatorname{Tr}(\Sigma_x M)\,\Sigma_x, \tag{6}$$

by Wick's theorem. The lemma follows since $\ell = \frac{1}{2}\mathbb{E}[x^\top M x] = \frac{1}{2}\operatorname{Tr}(\Sigma_x M)$. ■

A.2.2. PROOF OF COROLLARY 2

Proof Adam's preconditioner is based on $P^{(A)} := \mathbb{E}_{(x,y)}[g(x)g(x)^\top]$, so for an orthonormal basis U the diagonal entries satisfy $(D_{ii}^{(A)})^{-1} = \sqrt{u_i^\top P^{(A)} u_i}$. Combined with $(D_{ii}^{(\mathrm{GN},-1/2)})^{-1} = \sqrt{u_i^\top \Sigma_x u_i}$ and Lemma 1, the claim follows. ■

Single-sample $H^{(\mathrm{GN})}$. On single-sample batches, $g(\theta) = \ell_f' \nabla_\theta f$. With $\beta_1 = \beta_2 = 0$ and rotated gradient $\tilde{g} := U^\top g$,

$$D_{ii}^{(A)} = (|\tilde{g}_i|)^{-1} = \big((\ell_f')^2 (U^\top \nabla_\theta f_i)^2\big)^{-1/2}, \quad D_{ii}^{(\mathrm{GN})} = (H_{ii}^{(\mathrm{GN})})^{-1/2} = ((\ell_f')^2/\ell_f'')^{1/2}\, D_{ii}^{(A)}.$$

A.2.3. PROOF OF LEMMA 3

Proof Define $M^{(t)} = \mathbb{E}[(\theta^{(t)} - \theta^*)(\theta^{(t)} - \theta^*)^\top]$. For any P, the update $\theta^{(t+1)} - \theta^* = (\mathbf{I} - \eta P x x^\top)(\theta^{(t)} - \theta^*)$ yields

$$M^{(t+1)} = (\mathbf{I} - \eta P \Sigma_x)\, M^{(t)}\, (\mathbf{I} - \eta P \Sigma_x)^\top + \eta^2 P\big(\Sigma_x M^{(t)} \Sigma_x + \operatorname{Tr}(\Sigma_x M^{(t)})\Sigma_x\big) P. \tag{7}$$

Let $\widetilde{M}^{(t)} := \Sigma_x^{1/2} M^{(t)} \Sigma_x^{1/2}$ so $\mathbb{E}[\ell^{(t)}] = \mathbb{E}[\operatorname{Tr}(\widetilde{M}^{(t)})]$, and $\boldsymbol{A}(P) := \Sigma_x^{1/2} P \Sigma_x^{1/2}$. Then

$$\widetilde{M}^{(t+1)} \preceq (\mathbf{I} - \eta \boldsymbol{A}(P))\, \widetilde{M}^{(t)}\, (\mathbf{I} - \eta \boldsymbol{A}(P))^\top + 2\eta^2 \operatorname{Tr}(\widetilde{M}^{(t)})\, \boldsymbol{A}(P)^2. \tag{8}$$

Rotating into the eigenbasis of $\boldsymbol{A}(P)$ (preserving traces), let $v_t = \operatorname{diag}(\widetilde{M}^{(t)})$. Then

$$v_{t+1} = \big((\mathbf{I} - \eta \boldsymbol{A}(P))^2 + \eta^2 \boldsymbol{A}(P)^2 + \eta^2 \operatorname{diag}(\boldsymbol{A}(P)^2)\mathbf{1}^\top\big) v_t. \tag{9}$$

Taking the inner product with $\mathrm{diag}(\boldsymbol{A}(P)^{-1})$ on both sides, using $\mathbf{1}^\top v_t \geq \lambda_{\min}(\boldsymbol{A}(P))\,\mathrm{diag}(\boldsymbol{A}(P)^{-1})^\top v_t$, gives

$$\mathrm{diag}(\boldsymbol{A}(P)^{-1})^\top v_{t+1} \leq \big(1-\lambda_{\min}(\boldsymbol{A}(P))\,\eta(2-(2\lambda_{\max}+\mathrm{Tr}(\boldsymbol{A}(P)))\eta)\big)\,\mathrm{diag}(\boldsymbol{A}(P)^{-1})^\top v_t. \quad (10)$$

The optimal contraction is at $\eta = 1/(2\lambda_{\max} + \mathrm{Tr}(\boldsymbol{A}(P)))$, which is bounded above by $\big(1 - \frac{\lambda_{\min}(\boldsymbol{A}(P))}{3\,\mathrm{Tr}(\boldsymbol{A}(P))}\big)$. ■

A.3. Comparing GN powers

By Lemma 3, comparing GN^{-1} and $\mathrm{GN}^{-1/2}$ reduces to comparing the condition number of $\boldsymbol{A}(P) = P^{1/2}\Sigma_x P^{1/2}$ for $P \in \{\Sigma_x^{-1}, \Sigma_x^{-1/2}\}$. We claim:

Claim 6 *There exists Σ_x such that $\kappa(\boldsymbol{A}(diag(\Sigma_x)^{-1/2})) < \kappa(\boldsymbol{A}(diag(\Sigma_x)^{-1}))$.*

Empirically, by sampling random orthonormal U for fixed eigenvalue diagonals Λ and forming $\Sigma_x = U\Lambda U^\top$, we find both Σ_x such that $\mathrm{GN}^{-1/2}$ wins (Figure 2 (c)) and Σ_x such that GN^{-1} wins (Figure 1). Characterizing the favorable regime is left as future work.

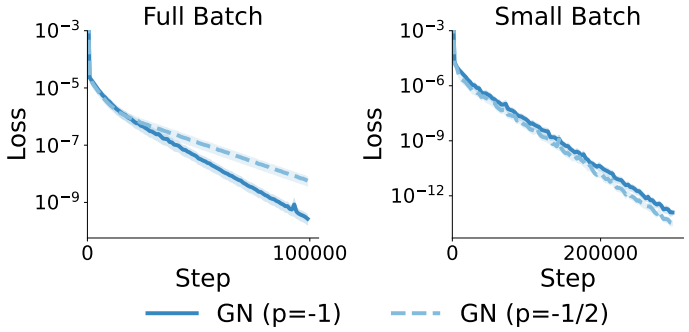


Figure 1. Comparing GN powers $p \in \{-\frac{1}{2}, -1\}$ on a covariance constructed so that GN^{-1} has the better condition number; here GN^{-1} converges as fast or faster than $\mathrm{GN}^{-1/2}$ at both small and large batches.

Appendix B. Proof of Theorem 4 (logistic lower bound)

The two-layer linear network with weight tying and squared parameterization satisfies $q_i(\theta) = \mathbf{Pr}_\theta(y = 1 \mid x = e_i) = \sigma\big(\sum_j \theta_j^2 x_j\big) = \sigma(\theta_i^2)$, $\sigma(z) = 1/(1 + e^{-z})$. For the logistic loss the population gradient and diagonal GN are

$$[g(\theta)]_i = 2\nu_i\theta_i(\sigma(\theta_i^2) - P_i), \quad [H^{(\mathrm{GN})}(\theta)]_{ii} = 4\nu_i\theta_i^2\sigma(\theta_i^2)(1 - \sigma(\theta_i^2)). \quad (11)$$

For step sizes $\{\eta_t\}$ and ridge $\alpha \geq 0$, the GN iteration $\theta^{(t+1)} = M_{\eta_t,\alpha}(\theta^{(t)})$ has coordinate update

$$[M_{\eta,\alpha}(\theta)]_i = \theta_i - \eta\,\frac{2\theta_i\nu_i(\sigma(\theta_i^2) - P_i)}{4\theta_i^2\nu_i\sigma(\theta_i^2)(1 - \sigma(\theta_i^2)) + \alpha}. \quad (12)$$

The proof hinges on a per-coordinate divergence threshold for the learning rate. Above it, the GN update diverges; requiring all coordinates to avoid divergence forces a small η^∞, which slows local convergence.

Lemma 7 (Divergence threshold) *For constants $\eta, \alpha > 0$, define the one-dim update map $M_{\eta,\alpha}(\theta) = \theta - \eta \frac{2\theta(\sigma(\theta^2)-P)}{4\theta^2\sigma(\theta^2)(1-\sigma(\theta^2))+\alpha}$. Consider $\theta^{(t+1)} = M_{\eta_t,\alpha}(\theta^{(t)})$. There is a universal constant c such that for any $P \in [0.6, 0.8]$ and any $\theta^{(0)} > 0$ with $\sigma((\theta^{(0)})^2) \le 0.55$, if the non-increasing $\{\eta_t\}$ satisfies $\eta_t \ge c\sqrt{\log(1/\theta^{(0)})}\big(\theta^{(0)} + \alpha/\theta^{(0)}\big)$ for all t, then $|\theta^{(t)}|$ diverges geometrically.*

Proof We split the proof into three parts.

Part 1: $\theta^{(1)}$ is large. $\sigma((\theta^{(0)})^2) \le 0.55$ implies $\theta^{(0)} \le 0.5$. Since $P \ge 0.6$, $\sigma((\theta^{(0)})^2) - P < 0$, so $\theta^{(1)} > \theta^{(0)}$ and

$$\theta^{(1)} \ge \theta^{(0)} + \eta_0 \tfrac{2\theta^{(0)}(0.6-0.55)}{4(\theta^{(0)})^2\cdot 0.5^2+\alpha} \ge \eta_0 \tfrac{0.1\theta^{(0)}}{(\theta^{(0)})^2+\alpha} = \tfrac{0.1\,\eta_0}{\theta^{(0)}+\alpha/\theta^{(0)}}.$$

Substituting the lower bound on η_0 from the lemma yields $\theta^{(1)} \ge 0.1\,c\,\sqrt{\log(1/\theta^{(0)})}$, which can be made arbitrarily large by choosing c large enough.

Part 2: key properties of $\theta^{(1)}$. We show that c can be chosen so that (i) $\sigma((\theta^{(1)})^2) \ge 0.9$ and (ii) $4(\theta^{(1)})^2(1-\sigma((\theta^{(1)})^2)) \le \theta^{(0)}$. (i) follows directly. For (ii), using $1-\sigma(z) \le e^{-z}$ and that $z(1-\sigma(z))$ is decreasing for $z \ge 2$, with $\theta^{(1)}_{\text{low}} := 0.1c\sqrt{\log(1/\theta^{(0)})}$,

$$\begin{aligned} 4(\theta^{(1)})^2(1-\sigma((\theta^{(1)})^2)) &\le 4(\theta^{(1)}_{\text{low}})^2 e^{-(\theta^{(1)}_{\text{low}})^2} = (0.04c^2)\log(1/\theta^{(0)})(\theta^{(0)})^{0.01c^2} \\ &= \big((0.04c^2)\log(1/\theta^{(0)})(\theta^{(0)})^{0.01c^2-1}\big)\,\theta^{(0)}. \end{aligned}$$

For sufficiently large c (e.g. $0.01c^2 > 2$), the parenthesized factor is < 1, since the polynomial decay dominates the logarithm.

Part 3: geometric divergence for $t \ge 1$. For any η above threshold and any θ with $\theta^2 \ge (\theta^{(1)})^2$, write $M_{\eta,\alpha}(\theta) = \theta(1-K(\theta))$ with $K(\theta) := \frac{2\eta(\sigma(\theta^2)-P)}{4\theta^2\sigma(\theta^2)(1-\sigma(\theta^2))+\alpha} > 0$. By (i), the numerator is $\ge 2\eta(0.9-0.8) = 0.2\eta$. By (ii) and $\theta^{(0)} \le 0.5$, the denominator is $\le \theta^{(0)} + \alpha/\theta^{(0)}$. Hence $K(\theta) \ge \frac{0.2\eta}{\theta^{(0)}+\alpha/\theta^{(0)}} \ge 0.2\,c\,\sqrt{\log(1/\theta^{(0)})}$. Since $\theta^{(0)} \le 0.5$, $\log(1/\theta^{(0)}) \ge \log 2$, and c can be chosen so that $0.2c\sqrt{\log 2} \ge 1+\sqrt{2}$. Thus $|1-K(\theta)| \ge \sqrt{2}$, so $|\theta^{(t+1)}| \ge \sqrt{2}|\theta^{(t)}|$. ■

Proof [Proof of Theorem 4] At θ^*, the diagonal entries of $H_*^{(\text{GN})} = H^{(\text{GN})}(\theta^*)$ are $\lambda_i(H_*^{(\text{GN})}) = 4(\theta_i^*)^2\nu_i\sigma((\theta_i^*)^2)(1-\sigma((\theta_i^*)^2))$. Since $P_i = \sigma((\theta_i^*)^2) \in [0.6, 0.8]$, $(\theta_i^*)^2$ is bounded, so $\lambda_i(H_*^{(\text{GN})})$ is proportional to ν_i (up to universal constants); in particular $\lambda_{\min}(H_*^{(\text{GN})}) \ge \nu_{\min}/c_1$.

The coordinate updates separate. Applying Lemma 7 per coordinate (with α/ν_i as the regularizer), since $\{\eta_t\}$ is non-increasing, η^∞ must satisfy $\eta^\infty \le c\sqrt{\log(1/\theta_i^{(0)})}\big(\theta_i^{(0)} + \alpha/(\nu_i\theta_i^{(0)})\big)$ for every i. With $\theta_i^{(0)} = 1/\sqrt{d}$, the tightest bound (at $\nu_i = \nu_{\max}$) gives $\eta^\infty \le c\sqrt{\log d}\,\big(1/\sqrt{d} + \sqrt{d}\alpha/\nu_{\max}\big)$. The local contraction factor satisfies $\gamma(\eta^\infty, \alpha) \ge 1 - \eta^\infty\,\lambda_{\min}(H_*^{(\text{GN})})/(\lambda_{\min}(H_*^{(\text{GN})}) + \alpha)$. Substituting and using $\lambda_{\min}/(\lambda_{\min}+\alpha) \le 1$ and $\alpha/(\lambda_{\min}+\alpha) \le 1$:

$$\gamma(\eta^\infty, \alpha) \ge 1 - c\sqrt{\log d}\big(1/\sqrt{d} + \sqrt{d}\lambda_{\min}/\nu_{\max}\big) \ge 1 - c'\sqrt{\log d}\big(1/\sqrt{d} + \sqrt{d}\nu_{\min}/\nu_{\max}\big),$$

which is the claimed bound. ■

Appendix C. Experiments

We complement the theory in §3–4 with experiments on (i) controlled simulations of the linear- and logistic-regression examples, (ii) non-convex MLPs on a random teacher, sparse parity, staircase, and CIFAR-10, and (iii) a 1-layer Transformer on a selection-based regression task. We use MSE unless otherwise noted, sweep η and α, and (for Adam) the schedule and β_2, fixing $\beta_1 = 0$ as in Das et al. [11]. Eigenbasis experiments use the Kronecker approximation [20, 29] of the GN eigenbasis. Curves report mean±SE over 10 seeds.

Hyperparameters, hardware, runtime. We sweep η first at factors of 3 (e.g. $0.01, 0.003, 0.001$) and then at factors of 2 around the optimum. We sweep α at factors of 10. For Adam we additionally sweep $\beta_2 \in \{0, 0.9, 0.95, 0.99\}$. When varying batch size we vary the gradient batch size and always use a large batch (4096) for $H^{(\mathrm{GN})}$ to ensure an accurate basis estimate. Runs use NVIDIA A100 GPUs. Simulations in §3.1 take $<$ 1min; GN-power simulations take $\sim$ 9min per 100k steps; parity/staircase $\sim$ 10min per 1k steps; CIFAR identity-basis runs $<$ 5min, Kronecker-eigenbasis runs $\sim$ 80min.

Kronecker factorization. For matrix-valued $W \in \mathbb{R}^{m\times n}$, let $g \in \mathbb{R}^{mn}$ be the flattened gradient and $G \in \mathbb{R}^{m\times n}$ the unflattened gradient. The $mn \times mn$ GN matrix is approximated by

$$H^{(\mathrm{GN})} := \mathbb{E}[gg^\top] \;\approx\; \mathbb{E}[GG^\top] \otimes \mathbb{E}[G^\top G]. \tag{13}$$

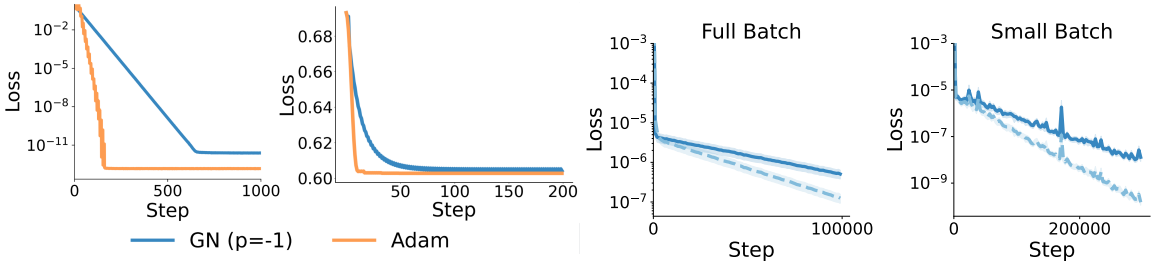
The eigenvectors of the Kronecker product are Kronecker products of those of the factors, so the eigenbasis of $H^{(\mathrm{GN})}$ is approximated by the eigenbases of the smaller $\mathbb{E}[GG^\top] \in \mathbb{R}^{m\times m}$ and $\mathbb{E}[G^\top G] \in \mathbb{R}^{n\times n}$. Without heuristic damping, we find the Kronecker approximation behaves similarly to the full eigenbasis while being substantially cheaper [6].

C.1. Simulations of the linear- and logistic-regression examples

Figure 2 reports simulations on three settings, with all three panels labeled inside the figure. *(a)* On the block-covariance example (2) ($d = 100$) with the identity basis, Adam converges quickly via auto-tuning while $\mathrm{GN}^{\pm 1}$ tracks vanilla GD. *(b)* On a 2048-dim instance of the logistic-regression example with the GN eigenbasis, Adam outpaces GN^{-1}, matching Theorem 4. *(c)* On a constructed Σ_x where $\mathrm{GN}^{-1/2}$ has the more favorable conditioning, $\mathrm{GN}^{-1/2}$ beats GN^{-1} at both small and large batches (further details in Section A.3).

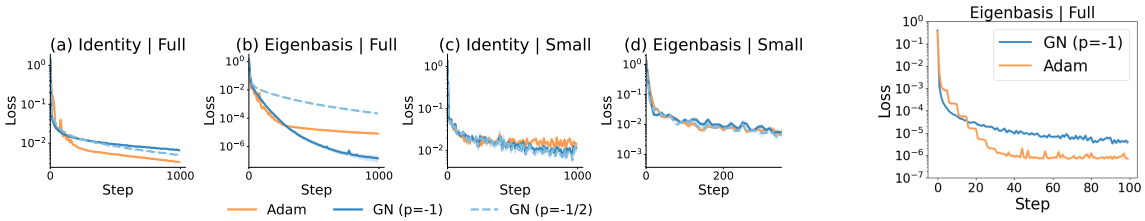
C.2. Random-teacher MLP and Selector-Transformer

Figure 3 summarizes our two flagship non-convex experiments. The random-teacher MLP *(a)* extends the linear-regression intuition to a smooth non-convex setting; the selector-Transformer *(b)* mirrors the squared parameterization of §4 via attention's query×key product. The two figures together cover both axes of Table 1: Adam tracks $\mathrm{GN}^{-1/2}$ at small batches in either basis (§3.2), and Adam matches or beats GN^{-1} when the basis is identity (§3.1) or the loss is logistic-like (§4).



(a) block-covariance LR (Id basis) (b) reparam. logistic (eigenbasis) (c) GN powers on Σ_{half} (Id basis)

Figure 2. Adam vs. GN simulations corroborating §3–4. *(a)* Block-covariance LR under the identity basis: Adam auto-tunes while $\text{GN}^{\pm 1}$ matches vanilla GD. *(b)* Reparameterized logistic regression under the GN eigenbasis: Adam beats GN^{-1} at full batches. *(c)* Σ_x constructed so that $\text{GN}^{-1/2}$ has the better condition number—$\text{GN}^{-1/2} > \text{GN}^{-1}$ at both small and large batches. (a)–(b) are the left and right halves of the left figure; (c) is the right figure.



(a) Random-teacher MLP (b) Selector-Transformer

Figure 3. Non-convex experiments. *(a)* Random-teacher MLP across the 2×2 grid of Table 1: Adam $\approx \text{GN}^{-1/2}$ at small batches in either basis, and Adam $\geq \text{GN}^{-1}$ in the identity basis. *(b)* Selector-Transformer: Adam outperforms GN^{-1} in the (Kronecker) eigenbasis with full batches, matching Theorem 4.

C.3. MLP experiments with squared loss

We learn each task with a single-hidden-layer MLP $\hat{y} = a \cdot \sigma(w^\top x + b)$, $w \in \mathbb{R}^{d \times m}$, $a, b \in \mathbb{R}^m$, with $\sigma = \text{ReLU}$.

- **Random teacher.** Inputs $x_i \sim \mathcal{N}(0, \Sigma_x)$ with random Σ_x. The teacher is itself a single-hidden-layer MLP; the student has hidden width $2\times$ the teacher's.
- **Sparse parity.** (d, k)-parity: $y = \prod_{i \in \mathcal{S}} x_i$ for an unknown support $\mathcal{S} \subseteq [d]$ [3, 5, 8, 12, 21]. We use $d = 20$, $k = 6$.
- **Staircase.** $y = \sum_{(s_i, e_i) \in \mathcal{P}} \prod_{j=s_i}^{e_i - 1} x_j$ for segments $\mathcal{P} = \{(s_i, e_i)\}_{i \in [k]}$ [1, 2]. We use $d = 21, k = 3, \mathcal{P} = \{(0, 7), (7, 14), (14, 21)\}$.
- **CIFAR-10** [16]. Inputs flattened to length 3072; labels are 10-dim one-hots. Each run uses 400 steps, sufficient for the large-batch eigenbasis runs to reach $\sim 47\%$ accuracy.

Figures 4 to 6 report the full 2×2 grid for each of these tasks, complementing Figure 3 (a) in the main text. Across all four, Adam $\approx \text{GN}^{-1/2}$ at small batches in either basis, and Adam matches or beats GN^{-1} in the identity basis.

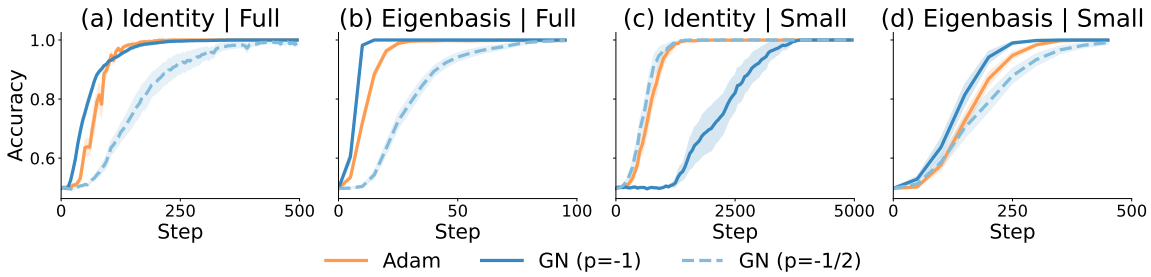


Figure 4. Sparse parity: Adam, GN^{-1}, and $\text{GN}^{-1/2}$ on the 2×2 grid (Table 1).

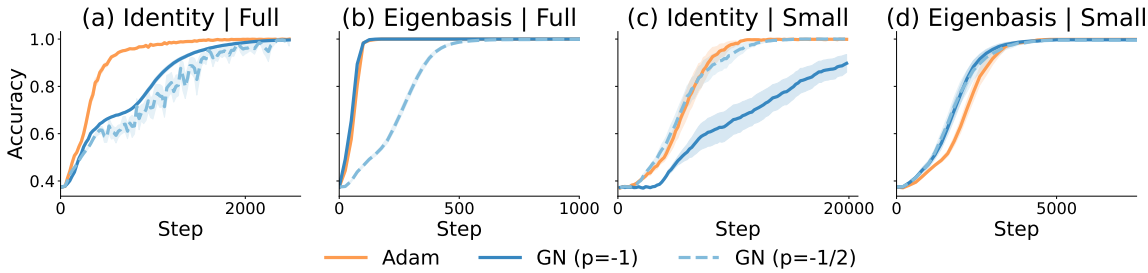


Figure 5. Staircase (a multi-feature generalization of sparse parity): same grid as Figure 4.

What about $p = -1$ for Adam? We define Adam's diagonal preconditioner with $p = -1/2$ in Section 2 for consistency with the standard algorithm. For completeness, Figure 7 compares Adam at $p \in \{-1/2, -1\}$ on sparse parity. Consistent with Lin et al. [18], the two powers behave comparably.

C.4. Interpolating between bases

To probe bases of intermediate quality, we interpolate between $\mathbf{I}$ and the GN eigenbasis U via geodesic interpolation (Algorithm 1), parameterized by $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$ (with $\gamma = 0$ giving $\mathbf{I}$ and $\gamma = 1$ giving U). Figure 8 shows the resulting curves on parity and staircase: Adam and $\text{GN}^{-1/2}$ track each other under the stochastic regime across all bases, as predicted by §3.2.

Algorithm 1: Geodesic interpolation between bases

Input: full GN basis U, interpolation factor $\gamma \in [0, 1]$.
  Compute matrix log $K := \text{logm}(U)$.
  Compute matrix exponent $\hat{U} := \exp(\gamma \cdot K)$.
  Take the real part $U_\gamma := \text{real}(\hat{U})$.
Output: U_γ.

C.5. Logistic experiment details

Simulations for §4. We follow the two-layer linear network of §4. Inputs are 2048-dim one-hots with $\nu_i \propto i^{-c}$, $c = 0.6$; we set $P_i = 0.75$ for all i.

Transformer experiments. We use a 1-layer 1-head Transformer of dimension 128. The attention block mirrors the squared parameterization in §4: query$\times$key matches the reparameterization, and

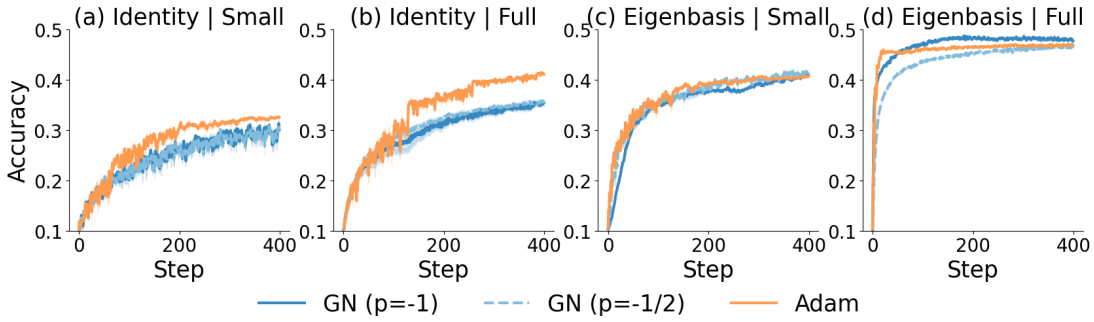


Figure 6. CIFAR-10: Adam, GN^{-1}, and $\mathrm{GN}^{-1/2}$ on the 2×2 grid.

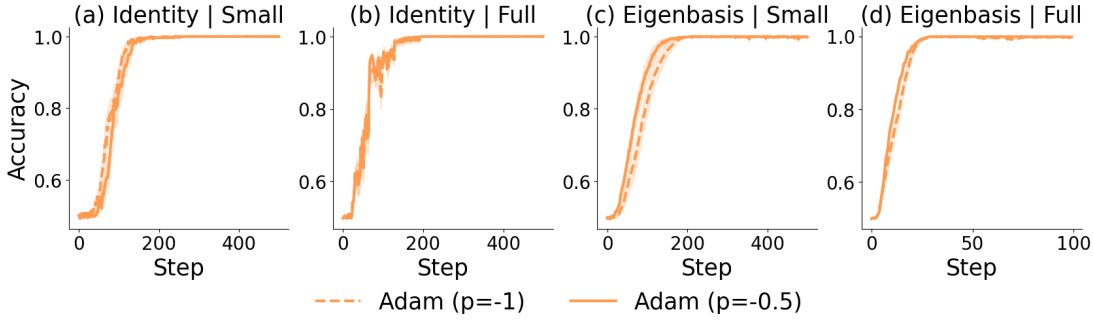
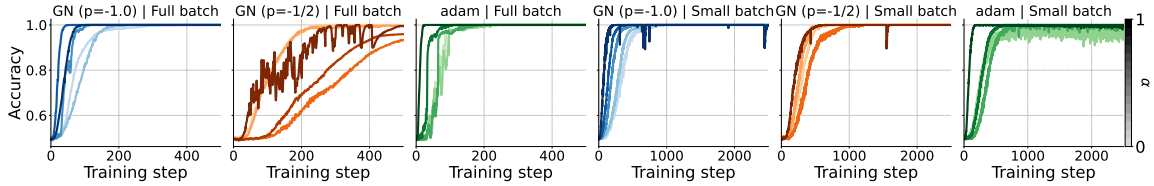
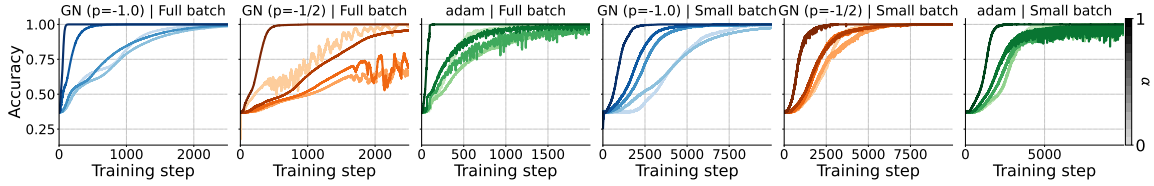


Figure 7. Sparse parity with Adam at power $p \in \{-1, -\frac{1}{2}\}$ on the 2×2 grid.

softmax matches the logistic. The selection task uses an input sequence of $T = 32$ Gaussian vectors followed by a length-d one-hot s ($d \geq T$), with target $y = \langle \theta_*, x_i \rangle$ when $s = e_i$. Figure 3 (b) compares GN and Adam under the Kronecker-approximated eigenbasis at batch size 16384. Results are aggregated over 10 seeds; each run takes ~ 90min.



(a) $(20, 6)$-Sparse parity



(b) $(21, 3)$-Staircase

Figure 8. Basis interpolation: GN^{-1}, $\mathrm{GN}^{-1/2}$, and Adam under bases interpolated geodesically between the eigenbasis (darker) and the identity (lighter), with $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$.