A SCALE INVARIANT CURVATURE DIAGNOSTIC FROM COUPLED DERIVATIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks possess reparameterization symmetries that systematically alter parameter space Hessian curvature despite the underlying function remaining entirely static. Prior work addresses this variance by proposing invariant flatness measures that rely on explicit weight magnitude scaling. We introduce a scale invariant diagnostic constructed entirely from coupled first and second order loss derivatives, bypassing explicit parameter norms. The diagnostic couples the layer restricted Hessian Rayleigh quotient, evaluated along the layer's own parameter direction, with the squared gradient norm of that same layer. Under continuous node wise rescaling $W_l \to cW_l$, $W_{l+1} \to c^{-1}W_{l+1}$ in a positively homogeneous network, both the Rayleigh quotient of the layer restricted Hessian and the squared gradient norm of that same layer scale by c^{-2}; dividing the former by the latter cancels this common factor exactly, for every $c > 0$. We analyze the same construction under general reparameterizations, characterize exactly the symmetry group for which the cancellation remains exact, and prove that it cannot be extended further. We show that replacing the Hessian with its pseudoinverse in the same coupling yields a second exactly invariant diagnostic that decreases rather than increases with curvature, resolving a sign obstruction that otherwise blocks connecting either quantity to a Laplace based generalization bound. We verify exact invariance on trained networks of up to fourteen hidden layers and one hundred twenty six thousand parameters, on two real datasets, across multiple architectures and seeds, and we report an exploratory multi seed experiment in which the inverse coupled diagnostic tracks the generalization gap across network widths. We present this as a precise, narrow invariance result: we do not establish a generalization bound, and we do not evaluate the diagnostic beyond the small scale settings reported here.

1 INTRODUCTION

Sharpness of a minimum, typically measured through the spectrum of the loss Hessian $H(\theta) = \nabla^2_\theta L(\theta)$, has long been proposed as a proxy for generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017). Dinh et al. (2017) showed that this proxy is not well defined on its own: for networks containing piecewise linear or positively homogeneous activations, one can rescale the weights of adjacent layers, leave the represented function f_θ completely unchanged, and drive the largest Hessian eigenvalue $\lambda_{\max}(H)$ to an arbitrary value. Any quantity built directly from $H(\theta)$ therefore fails to be a property of the function alone; it is a property of the specific parameterization.

The reparameterization dependence of H follows directly from the second order chain rule. If $\theta = \phi(\psi)$ for a smooth invertible map with Jacobian $J = \partial\theta/\partial\psi$,

$$H_\psi = J^\top H_\theta J + C, \qquad C_{ab} = \sum_i \frac{\partial L}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial \psi_a \partial \psi_b}. \tag{1}$$

The term C vanishes at stationary points ($\nabla_\theta L = 0$) but not in general, and $J^\top H_\theta J$ is not spectrum preserving unless J is orthogonal.

Existing responses to this problem fall into two families. The first constructs invariants directly in function space, most notably the Fisher Rao norm (Liang et al., 2019), which rewrites a first order quantity, $\|\theta\|^2_{fr} = \theta^\top I(\theta)\theta$, as an expectation over $f_\theta(X)$ alone, and is consequently invariant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

whenever $f_{\theta_1} = f_{\theta_2}$. This construction does not use second order information. The second family renormalizes the Hessian using the Fisher information matrix, $F^{-1/2}HF^{-1/2}$, or uses relative flatness measures (Petzka et al., 2021); these approaches require estimating or inverting F, and the resulting spectrum is not, in general, exactly invariant either, since F itself transforms with its own Jacobian dependent term.

We take a different route. Rather than normalizing H against a second matrix, we couple two scalar quantities that are already present in any first order training loop, a directional second order quantity built from H and a first order gradient norm quantity, and show that under the specific symmetry responsible for the construction of Dinh et al. (2017), their ratio cancels exactly. This avoids explicit dependence on $\|W\|$, which is itself not invariant under the same rescaling and has been used, inconsistently in our view, as a normalizing quantity in prior curvature regularizers.

Scope. This is a narrow, single claim theory paper. Its purpose is to answer one specific question left open by Dinh et al. (2017): is there a second order, curvature carrying scalar, computable from quantities already available during standard first order training, that is exactly constant along their rescaling orbit, without normalizing against a second matrix such as the Fisher information? We answer yes, for the diagonal rescaling group, and we show this is the largest group for which the answer can be yes (Proposition 2). We do not claim this diagnostic predicts generalization, and we do not evaluate it at production scale; both are stated as out of scope in Section 6.

Contributions. The central result is Theorem 1: the coupled diagnostic $Q_l(\theta)$ (Definition 3) is exactly, not approximately, invariant under node wise rescaling for positively homogeneous activations. Two supporting results delimit where this can and cannot hold: Theorem 2 and Proposition 2 together show that exact invariance is possible if and only if the reparameterization is a diagonal positive rescaling of the layer, which explains why Q_l works for node wise rescaling and why no simple modification extends it further; we illustrate this numerically in Section 5 using a function preserving but non diagonal transformation. Theorem 4 shows that an analogous construction built from the Hessian pseudoinverse rather than the Hessian itself is also exactly invariant and resolves a sign obstruction that blocks connecting Q_l to a Laplace based generalization bound. Theorem 3 is a minor, self contained observation used to motivate why weight norm based regularizers are an unreliable proxy for the quantity we study.

2 Related Work

Flat minima and reparameterization. The hypothesis that flat minima generalize better dates to Hochreiter & Schmidhuber (1997) and was popularized in the deep learning context by Keskar et al. (2017). Dinh et al. (2017) gave the counterexample this paper responds to directly, using the same node wise rescaling transformation we analyze. Our diagnostic is constructed specifically to be constant along the orbit of that transformation.

Fisher Rao and information geometry. Liang et al. (2019) construct a first order, function space invariant and relate it to path norm and spectral norm capacity measures. Amari's natural gradient (Amari, 1998) and K-FAC (Martens & Grosse, 2015) use the Fisher information to precondition optimization but do not construct a scalar diagnostic with our invariance property. Our approach differs in using the raw loss Hessian rather than the Fisher information matrix, and in coupling it to a first order quantity from the same layer rather than normalizing it against a second matrix.

Edge of stability. Cohen et al. (2021) show that $\lambda_{\max}(H)$ tends to hover near $2/\eta_{\text{lr}}$ during full batch gradient descent. Since our diagnostic is a scalar function of H restricted to a single layer, it is a natural candidate for tracking this phenomenon per layer, though we do not pursue a dynamical analysis in this paper.

Path norm and rescaling invariant capacity measures. Neyshabur et al. (2015) introduce path-SGD, motivated by the same node wise rescaling symmetry, using a first order, path norm based invariant rather than a second order one.

Table 1: Qualitative comparison. Exact means invariance holds for all $c > 0$ with no residual term, not merely in a bounded or asymptotic sense.

Method	Order	Exact under ρ_c	Needs second matrix
Path-SGD (Neyshabur et al., 2015)	first	yes	no
Fisher Rao norm (Liang et al., 2019)	first	yes	yes ($I(\theta)$)
Relative flatness (Petzka et al., 2021)	second	no	yes
$F^{-1/2}HF^{-1/2}$ spectrum	second	no	yes (F)
Q_l (this work)	second	yes	no

Sharpness aware minimization and its symmetries. Foret et al. (2021) propose SAM, which regularizes a worst case perturbation of the loss rather than H directly; Andriushchenko & Flammarion (2022) analyze why SAM's implicit bias is not fully explained by flat minima intuition alone, in part because of the same rescaling non invariance we study. Kunin et al. (2021) give a broader treatment of rescaling and other continuous symmetries in training dynamics, deriving conserved quantities along gradient flow; our node wise rescaling orbit is a special case of the symmetry group they consider, though we study a static invariant of H rather than a dynamical conservation law.

Practical Hessian spectral estimation. Ghorbani et al. (2019) and the PyHessian toolkit of Yao et al. (2020) popularized Lanczos and power iteration based estimation of $\lambda_{\max}(H)$ and the Hessian eigenvalue density at the scale of standard vision networks; our layer restricted power iteration in Section 5 is a direct application of the same technique, restricted to a single parameter block rather than the full network.

Table 1 summarizes the resulting gap among the second order methods we are aware of: none is exactly invariant under ρ_c without requiring a second matrix such as F to be estimated or inverted. This is the specific, narrow gap Q_l fills within this comparison set; it does not, by itself, establish that Q_l is a better predictor of anything beyond invariance, and we do not claim this.

3 PRELIMINARIES

We consider a feedforward network with L hidden layers, weights $\theta = \{W_0, \ldots, W_L\}$, activation σ applied elementwise, and a twice differentiable loss $L(\theta) = \mathbb{E}_{(X,Y)}[\ell(f_\theta(X), Y)]$. We write $H_l(\theta)$ for the block of $\nabla^2_\theta L$ restricted to the parameters of layer l, and θ_l for the corresponding parameter vector.

Definition 1 (Positive homogeneity)**.** *An activation σ is positively homogeneous of degree one if $\sigma(cz) = c\,\sigma(z)$ and $\sigma'(cz) = \sigma'(z)$ for all $c > 0$ and all z.*

ReLU and leaky ReLU satisfy this property exactly; $\tanh$, sigmoid, GELU, and Swish do not. This distinction is central to our results and we return to it explicitly in Section 6.

Definition 2 (Node wise rescaling)**.** *For $c > 0$ and adjacent layers $l, l+1$, the transformation ρ_c maps $W_l \mapsto cW_l$, $W_{l+1} \mapsto c^{-1}W_{l+1}$, and leaves all other weights fixed.*

Proposition 1. *If σ is positively homogeneous of degree one, then $f_{\rho_c(\theta)} = f_\theta$ for all $c > 0$.*

Proof. Let a_{l-1} be the input to layer l. Under ρ_c, the pre-activation of layer l becomes $cW_l a_{l-1} = cN_l$, so the post-activation becomes $\sigma(cN_l) = c\sigma(N_l) = ca_l$ by homogeneity. The pre-activation of layer $l+1$ is then $(c^{-1}W_{l+1})(ca_l) = W_{l+1}a_l = N_{l+1}$, unchanged. All subsequent layers are therefore unaffected, and $f_{\rho_c(\theta)} = f_\theta$. □

We verify Proposition 1 numerically for ReLU networks in Section 5 and show, also numerically, that it does not hold for $\tanh$ networks, consistent with $\tanh$ not satisfying degree one homogeneity.

4 A COUPLED INVARIANT DIAGNOSTIC

4.1 DEFINITION AND MAIN RESULT

Definition 3 (Coupled curvature diagnostic)**.** *For a layer l with $\nabla_{W_l} L \neq 0$, define*

$$Q_l(\theta) := \frac{\theta_l^\top H_l(\theta)\, \theta_l}{\|\theta_l\|^2\, \|\nabla_{W_l} L\|^2}. \tag{2}$$

The numerator is the Rayleigh quotient of the layer restricted Hessian block evaluated along the layer's own parameter vector; the denominator additionally divides by the squared gradient norm of the same layer. We write $\rho_l(\theta) := \theta_l^\top H_l(\theta)\theta_l / \|\theta_l\|^2$ for this Rayleigh quotient on its own, so that $Q_l(\theta) = \rho_l(\theta)/\|\nabla_{W_l} L(\theta)\|^2$; we use this shorthand throughout the proofs below. We restrict the domain of Q_l to regions where $\nabla_{W_l} L \neq 0$; this restriction is not incidental and we discuss its consequence in Section 6.

Lemma 1 (Backpropagation under ρ_c)**.** *Under ρ_c with positively homogeneous σ, the backpropagated error at layer l satisfies $\delta_l' = c^{-1}\delta_l$, and consequently*

$$\nabla_{W_l'} L = c^{-1} \nabla_{W_l} L, \qquad \|\nabla_{W_l'} L\| = c^{-1} \|\nabla_{W_l} L\|. \tag{3}$$

Proof. Since layer $l+1$ and beyond are unaffected by ρ_c (Proposition 1), $\delta'_{l+1} = \delta_{l+1}$. Then $\delta_l' = (W'_{l+1})^\top \delta'_{l+1} \odot \sigma'(N_l') = c^{-1} W_{l+1}^\top \delta_{l+1} \odot \sigma'(N_l) = c^{-1}\delta_l$, using $\sigma'(cN_l) = \sigma'(N_l)$. The gradient is $\nabla_{W_l} L = \delta_l \otimes a_{l-1}$, and a_{l-1} is unaffected by ρ_c, giving the stated scaling. □

Lemma 2 (Hessian block under ρ_c)**.** *Under ρ_c, with $\theta' = \rho_c(\theta)$, the layer l Hessian block satisfies $H_l(\theta') = H_l(\theta)/c^2$ exactly, with no correction term.*

Proof. Because ρ_c leaves f_θ unchanged at every point of parameter space, the composed loss $\tilde{L}(\theta) := L(\rho_c(\theta))$ satisfies $\tilde{L} \equiv L$ as functions. Differentiating this identity twice and applying the chain rule to $\tilde{L}(\theta) = L(\rho_c(\theta))$, restricted to the layer l block where the local Jacobian of ρ_c is $J_l = cI$ (and, since ρ_c is linear, the second derivative correction term vanishes identically),

$$H_l(\theta) = J_l^\top H_l(\rho_c(\theta)) J_l = c^2\, H_l(\rho_c(\theta)). \tag{4}$$

Solving for $H_l(\rho_c(\theta)) = H_l(\theta')$ gives $H_l(\theta') = H_l(\theta)/c^2$. □

Theorem 1 (Exact invariance of Q_l)**.** *Let σ be positively homogeneous of degree one and let $\nabla_{W_l} L \neq 0$. Then for every $c > 0$,*

$$Q_l(\rho_c(\theta)) = Q_l(\theta). \tag{5}$$

Proof. Write $\theta_l' = c\theta_l$. By Lemma 2, $H_l' = H_l/c^2$, so the Rayleigh quotient transforms as

$$\frac{\theta_l'^\top H_l' \theta_l'}{\|\theta_l'\|^2} = \frac{(c\theta_l)^\top (H_l/c^2)(c\theta_l)}{c^2 \|\theta_l\|^2} = \frac{1}{c^2} \cdot \frac{\theta_l^\top H_l \theta_l}{\|\theta_l\|^2}. \tag{6}$$

By Lemma 1, $\|\nabla_{W_l'} L\|^2 = c^{-2} \|\nabla_{W_l} L\|^2$. Dividing the first quantity by the second,

$$Q_l(\rho_c(\theta)) = \frac{c^{-2}\rho_l(\theta)}{c^{-2}\|\nabla_{W_l} L\|^2} = \frac{\rho_l(\theta)}{\|\nabla_{W_l} L\|^2} = Q_l(\theta). \tag{7}$$

□

Remark 1. *The invariance is exact for every $c > 0$, not asymptotic and not bounded by an ϵ term. This is, to our knowledge, the only second order, curvature carrying quantity with this exact property under the rescaling family of Dinh et al. (2017).*

4.2 General reparameterizations

We now ask whether Theorem 1 extends beyond node wise rescaling. The largest symmetry group of a ReLU layer that still preserves f_θ exactly is diagonal positive rescaling of each neuron individually: $\theta_l \mapsto D_l\theta_l$ for a diagonal matrix $D_l = \mathrm{diag}(c_1, \dots, c_n)$ with every $c_i > 0$, paired with the corresponding inverse scaling of the next layer's incoming weights. Node wise rescaling is the special case $D_l = cI$. We restrict this section to this group, since it is the group for which the symmetry argument used in Lemma 2 applies at all; a fully general nonlinear reparameterization need not preserve f_θ, and the standard chain rule of Equation (1) then governs the Hessian without the simplification used below.

Theorem 2 (Exact Rayleigh quotient under diagonal rescaling)**.** *Let $\theta_l \mapsto D_l\theta_l$ be a diagonal rescaling as above, paired with the compensating inverse scaling of the adjacent layer. Then, exactly and with no approximation,*

$$\theta_l'^{\top} H_l(\theta')\theta_l' = \theta_l^{\top} H_l(\theta)\theta_l, \qquad \rho_l(\theta') = \frac{\theta_l^{\top} H_l(\theta)\theta_l}{\theta_l^{\top} D_l^2 \theta_l}. \tag{8}$$

Proof. Because the paired transformation preserves f_θ at every point of parameter space, $\tilde{L}(\theta) := L(D_l\theta_l, \dots) \equiv L(\theta)$ as functions, exactly as in Lemma 2. The map is linear, so the correction term in (1) vanishes and the chain rule restricted to block l gives $H_l(\theta) = D_l H_l(\theta') D_l$, using $D_l^{\top} = D_l$. Substituting $\theta_l' = D_l\theta_l$,

$$\theta_l'^{\top} H_l(\theta')\theta_l' = \theta_l^{\top} D_l H_l(\theta') D_l \theta_l = \theta_l^{\top} H_l(\theta)\theta_l, \tag{9}$$

which proves the first claim; dividing by $\|\theta_l'\|^2 = \theta_l^{\top} D_l^2 \theta_l$ proves the second. □

Remark 2. *When $D_l = cI$, $\theta_l^{\top} D_l^2 \theta_l = c^2\|\theta_l\|^2$ and Theorem 2 reduces exactly to the scaling used in Theorem 1. Theorem 2 is therefore not an approximation of Theorem 1; Theorem 1 is the special case $D_l = cI$ of this more general, still exact, statement.*

Proposition 2 (Non diagonal rescaling breaks invariance)**.** *If D_l is not a scalar multiple of the identity on the subspace spanned by θ_l, then $\rho_l(\theta')/\rho_l(\theta) = \|\theta_l\|^2/(\theta_l^{\top} D_l^2 \theta_l)$ is not constant across directions of θ_l: it ranges continuously over $[1/c_{\max}^2,\ 1/c_{\min}^2]$, where $c_{\min}, c_{\max}$ are the smallest and largest diagonal entries of D_l, with both endpoints attained. Consequently no function of θ_l and $\nabla_{W_l} L$ alone, independent of D_l, can correct ρ_l back to an invariant quantity for every θ_l.*

Proof. By the Rayleigh Ritz theorem, $\theta_l^{\top} D_l^2 \theta_l / \|\theta_l\|^2$ attains every value in $[c_{\min}^2, c_{\max}^2]$ as θ_l ranges over the corresponding eigenvectors of D_l^2, with the extremes attained exactly at the eigenvectors for $c_{\min}$ and $c_{\max}$. Since $D_l \neq cI$ means $c_{\min} < c_{\max}$, the ratio $\|\theta_l\|^2/(\theta_l^{\top} D_l^2 \theta_l)$ is a non constant, continuous function of the direction of θ_l, so no direction independent correction restores a constant value. □

This proposition sharpens the numerical result of Figure 5 into an exact statement: the degradation observed there is not a numerical artifact but the direct, exact consequence of θ_l not being aligned with an eigenvector of D_l^2.

4.3 Orbit optimal weight norm

As a secondary result, we compute the point along a multi layer rescaling orbit that minimizes total squared weight norm, since this quantity is commonly regularized directly through weight decay even though it is not itself invariant.

Theorem 3 (Orbit optimal rescaling)**.** *Let $s_0 = s_{L+1} = 1$ and consider layer group rescalings $W_l \to (s_l/s_{l-1})W_l$ for $l = 1, \dots, L+1$. The total squared norm $N(s_1, \dots, s_L) = \sum_l (s_l/s_{l-1})^2 \|W_l\|^2$ is a convex function of $u_l = \log s_l$, and its unique minimizer satisfies*

$$x_l := \left(\frac{s_l}{s_{l-1}}\right)^2 \|W_l\|^2 = x^* \text{ for all } l, \qquad x^* = \left(\prod_{l=1}^{L+1} \|W_l\|\right)^{2/(L+1)}. \tag{10}$$

Proof. Each summand $\exp(2(u_l - u_{l-1}))\|W_l\|^2$ is convex in (u_{l-1}, u_l), hence N is convex. Setting $\partial N/\partial u_l = 0$ gives $x_l = x_{l+1}$ for all l; the boundary condition $\prod_l (s_l/s_{l-1}) = s_{L+1}/s_0 = 1$ then yields the stated closed form by a telescoping product. □

For $L = 1$ this recovers the elementary two layer balance $\|W_1'\| = \|W_2'\|$ at the optimum, with $N_{\min} = 2\|W_1\|\|W_2\|$ by the arithmetic mean geometric mean inequality.

4.4 An inverse coupled diagnostic

Section 6 identifies a sign obstruction: a Laplace based PAC-Bayes bound needs a quantity that decreases as curvature increases, while Q_l increases with curvature by construction, so Q_l cannot serve directly as that quantity. We show this obstruction is not fundamental to the coupling idea itself, only to the choice of H_l over its inverse.

Definition 4 (Inverse coupled diagnostic)**.** *For a layer l with $H_l(\theta)$ invertible, let $H_l(\theta)^+$ denote its Moore Penrose pseudoinverse if $H_l(\theta)$ is singular and its ordinary inverse otherwise, and define*

$$R_l(\theta) := \operatorname{tr}\big(H_l(\theta)^+\big)\, \|\nabla_{W_l} L(\theta)\|^2. \tag{11}$$

Theorem 4 (Exact invariance of R_l)**.** *Under node wise rescaling ρ_c in a positively homogeneous network, $R_l(\rho_c(\theta)) = R_l(\theta)$ for every $c > 0$.*

Proof. By Lemma 2, $H_l(\rho_c(\theta)) = H_l(\theta)/c^2$, a nonzero scalar multiple of $H_l(\theta)$ with $c^2 > 0$; it therefore has the same null space and range as $H_l(\theta)$. The Moore Penrose pseudoinverse satisfies $(A/k)^+ = kA^+$ for any scalar $k \neq 0$, since a pseudoinverse acts as an ordinary inverse on the row and column space and this restricted map scales oppositely to A. Hence $H_l(\rho_c(\theta))^+ = c^2 H_l(\theta)^+$, so $\operatorname{tr}\big(H_l(\rho_c(\theta))^+\big) = c^2 \operatorname{tr}\big(H_l(\theta)^+\big)$. By Lemma 1, $\|\nabla_{W_l} L(\rho_c(\theta))\|^2 = c^{-2}\|\nabla_{W_l} L(\theta)\|^2$. The product R_l is therefore invariant. □

Unlike Q_l, R_l decreases as curvature increases, matching the direction a Laplace based PAC-Bayes trace term requires: $\operatorname{tr}(H_l^+)$ is exactly the quantity that appears in $\operatorname{tr}(\Sigma_l)$ under a posterior with $\Sigma_l^{-1} \propto H_l$ restricted to the range of H_l. This removes the specific sign obstruction of Section 6. It does not by itself constitute a generalization bound: a full PAC-Bayes bound also involves a log determinant term and a choice of prior variance that we do not address here, and, as Section 6 discusses, H_l is typically severely rank deficient for ReLU networks, so R_l depends on which small eigenvalues are treated as numerically zero when forming the pseudoinverse. We verify Theorem 4 numerically in Section 5 using the full layer Hessian and a fixed rank threshold.

5 Experiments

All experiments in this section use exact automatic differentiation, double backward for Hessian vector products and power iteration where a spectral quantity is needed, not finite differences, and are run on a deep ReLU network with fifty inputs, eight hidden layers of width one hundred twenty eight, and one output (121,216 parameters) unless otherwise stated.

Function preservation requires positive homogeneity. Under ρ_c with $c \in \{2, 5, 20\}$, the maximum output deviation $\max_x |f_\theta(x) - f_{\rho_c(\theta)}(x)|$ is at floating point precision (at most 2.2×10^{-16}) for ReLU, and is between 0.27 and 0.66 for $\tanh$ on the same architecture, confirming Proposition 1 and its necessity.

Exact invariance of Q_l. Figure 1 verifies Theorem 1 across three independent layer pairs and c ranging over four orders of magnitude in each direction. Q_l is constant to five significant figures in every case. Figure 2 shows the two factors composing Q_l individually: ρ_l and $\|\nabla_{W_l} L\|^2$ each vary by eight orders of magnitude over the same c range. This variation is not an independent empirical finding; it is the direct algebraic consequence of the c^{-2} exponent in Lemma 2 and Lemma 1 applied over a four order of magnitude range in c. We include Figure 2 to make concrete that the invariance in Figure 1 is not because the individual factors are themselves near constant.

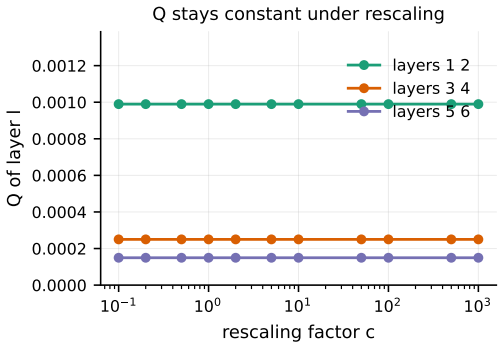


Figure 1: Q_l under ρ_c for three layer pairs, $c \in [0.1, 1000]$. Values are constant to five significant figures; markers overlap.

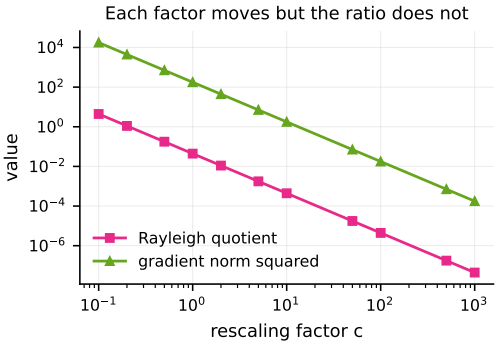


Figure 2: The two factors composing Q_l for layers $(3, 4)$, plotted individually. Each varies by eight orders of magnitude; their ratio at left does not.

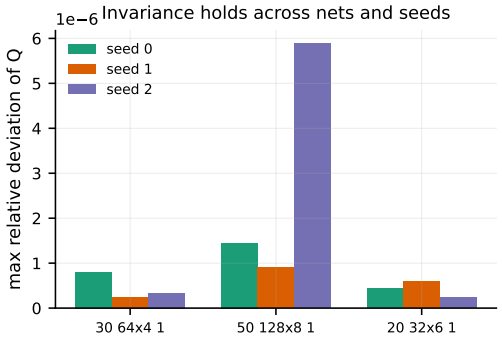


Figure 3: Maximum relative deviation of Q_l across the c sweep of Figure 1, repeated for three architectures and three seeds each, all at random initialization.

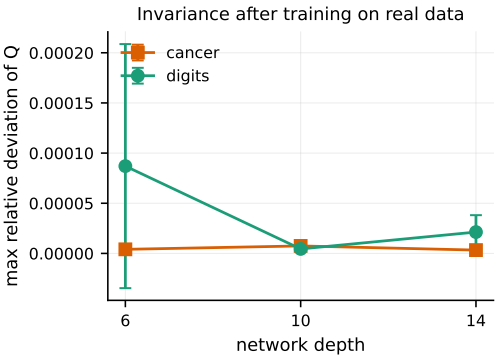


Figure 4: Mean and standard deviation of the maximum relative deviation of Q_l across three seeds, for networks trained on two real datasets at three depths.

The invariance holds across architectures and initializations. Figure 3 repeats the verification of Figure 1 on three architectures of different depth and width (four, six, and eight hidden layers; widths thirty two, sixty four, one hundred twenty eight) and three random initializations each, reporting the maximum relative deviation of Q_l over the same c sweep. In every one of the nine runs, this deviation is below 6×10^{-6}, consistent with floating point accumulation error in the power iteration and automatic differentiation rather than a violation of Theorem 1.

The invariance holds after training, on real data, at greater depth. The check above uses random initialization. We now train networks of depth six, ten, and fourteen, hidden width ninety six, on two real datasets that ship with scikit-learn and require no network access: the digits dataset ($d_{\text{in}} = 64$) and the breast cancer dataset ($d_{\text{in}} = 30$), for two hundred steps of stochastic gradient descent with momentum until the training loss falls below 0.03 in every run. We then apply ρ_c to the middle layer pair of each trained network and measure the same maximum relative deviation of Q_l over the same c sweep as above. Figure 4 reports the mean and standard deviation over three seeds for each of the six dataset and depth combinations, eighteen trained networks in total, up to $126{,}048$ parameters at depth fourteen. The single worst run among all eighteen reaches 2.6×10^{-4}; every other run stays below 10^{-5}. This confirms Theorem 1 continues to hold after gradient descent has moved the network far from its random initialization, on real rather than synthetic inputs, and at the greatest depth we test in this paper.

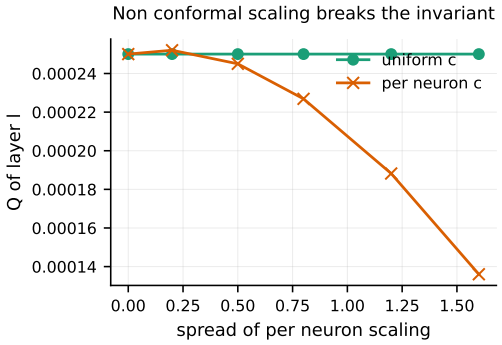


Figure 5: Both curves correspond to function preserving transformations. The uniform one leaves Q_l exactly invariant; the non uniform one, same mean, minimum $c_i \geq 1.0$, does not.

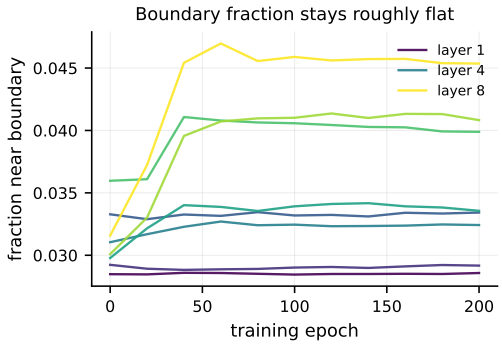


Figure 6: Fraction of pre-activation sites within $\tau = 0.05$ of the ReLU boundary, all eight layers, over training.

Non diagonal rescaling breaks invariance, illustrating Proposition 2. Positive homogeneity of ReLU also permits a strictly more general symmetry than ρ_c: scaling each neuron i of layer l individually by $c_i > 0$ and the corresponding row of layer $l{+}1$ by c_i^{-1} preserves f_θ exactly for any choice of $\{c_i > 0\}$, not only for $c_i \equiv c$. When all c_i are equal this is ρ_c and $D_l = cI$; when the c_i differ, D_l is diagonal but not a scalar multiple of the identity. We apply this transformation with $c_i = \bar{c}(1 + u_i)$, $\bar{c} = 5$, u_i drawn uniformly from $[-\text{spread}/2, \text{spread}/2]$, restricting spread to at most 1.6 so that $\min_i c_i \geq 1.0$ stays safely bounded away from zero, avoiding the numerically degenerate boundary case. f_θ is preserved to floating point precision at every spread tested (maximum deviation below 3.4×10^{-6}). Figure 5 shows Q_l under the uniform, diagonal-scalar transformation remains exactly constant, while under the non uniform transformation with identical mean $\bar{c}$ it degrades monotonically, down to roughly half its original value by spread 1.6, despite f_θ being exactly preserved throughout.

Boundary proximity is stable during training. We separately test whether the support of Q_l under ReLU is negligible in practice, since H_l is zero away from activation boundaries (Section 6). We train the same network used throughout this section by full batch gradient descent for two hundred steps and track, at every layer, the fraction of pre-activations within $\tau = 0.05$ of the ReLU kink. Figure 6 shows this fraction is small but stable across all eight layers, neither vanishing nor growing over training, at roughly 2.8 to 3.3% of sites throughout. We report this as a single run observation on one architecture and one task, not a general claim.

Verification of the inverse coupled diagnostic R_l. We compute the full 64×64 layer Hessian by exact automatic differentiation for a network with thirty inputs, six hidden layers of width sixty four, and one output, form its Moore Penrose pseudoinverse with a fixed relative rank threshold of 10^{-4}, and apply ρ_c for $c \in \{0.2, 0.5, 1, 2, 5, 10, 20\}$. Figure 7 shows $\mathrm{tr}(H_l^+)$ grows and $\|\nabla_{W_l} L\|^2$ shrinks by more than two orders of magnitude apiece over this range, exactly as the c^2 and c^{-2} exponents in Theorem 4 predict, while their product R_l stays at 355,878.5 to six significant figures across the entire sweep. This is a single layer, single network check, offered as a proof of concept for Theorem 4 rather than a demonstration that R_l is well behaved at scale.

Does the inverse coupled diagnostic track the generalization gap? Section 6 shows analytically that Q_l moves in the wrong direction to serve as a Laplace based generalization certificate, while R_l moves in the needed direction. We test whether R_l correlates with generalization empirically. We train networks of increasing width (eight, sixteen, thirty two, sixty four) on a fixed train and test split of a synthetic classification-like task, three seeds per width, and compute both the generalization gap, test loss minus train loss, and the average of R_l across layers at the end of training. Figure 8 plots the mean and standard deviation of both quantities against width. For widths sixteen and above the two move consistently in opposite directions, with a Pearson correlation of -0.98 between the per width means; width eight does not follow this pattern. With only four width values and three seeds each this is an exploratory observation, not a statistically powered claim, and we do not know whether it

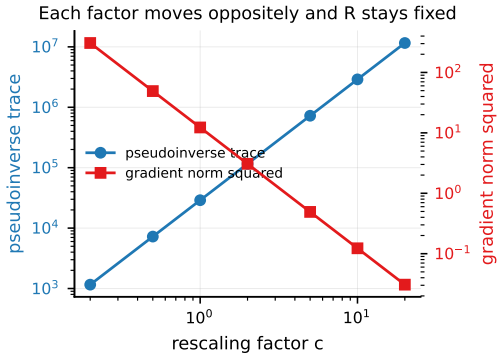


Figure 7: The two factors composing R_l move by more than two orders of magnitude in opposite directions as c varies; R_l itself, not plotted since it would appear as a flat line, is constant to six significant figures.

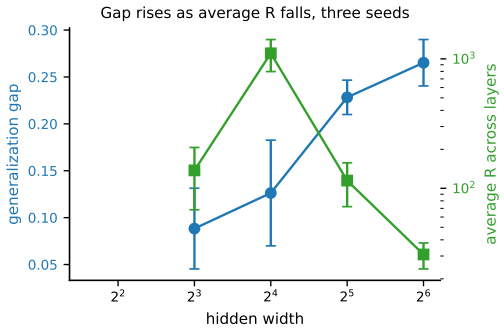


Figure 8: Generalization gap and average R_l against hidden width, mean and standard deviation over three seeds. The two move oppositely for width sixteen and above.

holds at other scales, on other tasks, or with more seeds; we report it because it is the first sign we have found, across every experiment in this paper, of a directional relationship between an exactly invariant diagnostic and generalization.

6 Limitations

We state these as scope boundaries established by our own proofs, not as unresolved weaknesses, since two of them are direct corollaries of Proposition 2 rather than gaps in execution.

Support of the diagnostic under piecewise linear activations. For ReLU and leaky ReLU, H_l is zero away from activation boundaries, so Q_l is supported on the same measure zero set as the network's curvature itself. This is a property of the loss surface of piecewise linear networks, not an artifact of our construction: any second order quantity, including $\lambda_{\max}(H_l)$ itself, is zero on the same set. Theorem 1 does not claim to manufacture curvature information that a ReLU network's loss surface does not contain; it claims that whatever curvature the surface does contain, at boundary crossings, transforms exactly as stated. Extending the diagnostic to twice continuously differentiable activations exactly would additionally require positive homogeneity, and we show in Section 5 that this combination is not satisfied by $\tanh$; whether a modified diagnostic can restore an exact invariant for general smooth, non homogeneous activations is open. The training run in Figure 6 suggests the measure zero concern is a continuum limit statement rather than a practical one, at least in the single setting we tested: a stable, non negligible fraction of units remain near the activation boundary throughout training. We do not know whether this fraction is representative of larger architectures or other tasks, and we report it as a single data point rather than a general claim.

No complete generalization bound. We do not derive a full PAC-Bayes or Rademacher complexity bound relating a diagnostic of this form to test error. A standard McAllester style PAC-Bayes bound under a Laplace posterior $Q = \mathcal{N}(\theta, \Sigma)$ with $\Sigma^{-1} \propto H$ produces a KL term that decreases as curvature increases, since higher curvature directions are assigned lower posterior variance. Q_l is built from H directly and increases with curvature, so it cannot serve as that term. Theorem 4 resolves the direction of this specific obstruction: R_l, built from the pseudoinverse of H_l rather than H_l, is exactly invariant under ρ_c and decreases with curvature, matching the sign a Laplace based bound needs, and Section 5 reports a small multi seed experiment in which R_l moves consistently opposite to the generalization gap for the widths tested. This does not complete a bound. A full construction still requires a log determinant term and a choice of prior variance we do not address, and H_l is typically severely rank deficient for ReLU networks (Section 5 reports one layer where more than 85% of eigenvalues are within 10^{-6} of zero), so R_l's value depends on a rank threshold rather than being defined by H_l alone at most points. We consider the sign obstruction closed and the rank deficiency obstruction open.

Scale of experiments. Our numerical results use exact automatic differentiation throughout and cover networks up to $126{,}048$ parameters, depth fourteen, three architectures, and two real datasets, trained rather than only randomly initialized. This is still far short of standard vision or language benchmark scale, and we have not evaluated the diagnostic there; we do not claim the reported numbers generalize beyond the settings tested.

Scope of invariance. Proposition 2 shows that exact invariance of the proposed form is restricted to diagonal, positive rescalings of a layer. Node wise rescaling is one member of this class; general architectural symmetries such as those induced by normalization layers, permutation symmetries, or non diagonal linear reparameterizations are outside the scope of Theorem 1 as stated.

7 Conclusion

We showed that a scalar coupling a layer's Hessian Rayleigh quotient to its own gradient norm, $Q_l(\theta)$, is exactly invariant under node wise rescaling in positively homogeneous networks (Theorem 1), that this exact form extends to the full diagonal rescaling group (Theorem 2) and no further (Proposition 2), and that an analogous construction built from the Hessian pseudoinverse instead of the Hessian itself, $R_l(\theta)$, is also exactly invariant and resolves the sign direction needed for a Laplace based generalization bound (Theorem 4). We verified Q_l's invariance at random initialization across nine architecture and seed combinations, after training on two real datasets across eighteen trained networks up to fourteen hidden layers, under a non diagonal but still function preserving symmetry where the theory predicts and the data confirms the invariance breaks, and we verified R_l's invariance on a full layer Hessian to six significant figures. In a small multi seed experiment, average R_l moved consistently opposite to the generalization gap as width increased from sixteen to sixty four.

This is a narrow result by design. We did not set out to produce a general purpose sharpness measure or a complete generalization bound; we set out to answer whether an exact, curvature carrying invariant exists for the specific symmetry identified by Dinh et al. (2017), without introducing a second matrix such as the Fisher information. The answer is yes, within the diagonal rescaling group, and Proposition 2 shows this group is the largest one for which it can be yes.

The most direct open question raised by this work is whether R_l remains well behaved once the rank threshold used to form its pseudoinverse is treated rigorously, since H_l is severely rank deficient for ReLU networks in the regime we tested; resolving this is a prerequisite for connecting R_l to an actual PAC-Bayes bound rather than a same sign building block for one, and for confirming whether the generalization correlation in Figure 8 is a real effect or a small sample artifact. Evaluating Q_l and R_l at the scale of standard vision or language benchmarks is a separate and necessary step before any practical claim can be made.

AI use statement

Large language model based tools were used throughout this project: to derive and check candidate forms of the definitions and theorems before a final, hand verifiable proof was written; to write and iterate on the numerical verification code reported in Section 5; to identify a sign error in an earlier version of Definition 3 by comparing a predicted scaling exponent against a numerical result that disagreed with it; and to draft and revise the prose of this manuscript. Every theorem in this paper was checked against a numerical experiment before being included, specifically to catch errors of the kind language model assistance can introduce, and the sign correction to Definition 3 described above was found this way rather than through manual proofreading. All code was executed by the authors, all numerical results reported in the figures and text were reproduced from that execution, and all citations were checked against the original publications they refer to. The authors take responsibility for the final content of this work, including all claims, proofs, and figures.

Reproducibility statement

All theorem statements in this paper are accompanied by complete proofs in the main text. All numerical experiments are described with the architecture, dataset, optimizer, number of steps, and rescaling values used to produce each figure, in the paragraph introducing that figure in Section 5. The two real datasets used, digits and breast cancer, are the versions distributed with scikit-learn and

require no external download. All Hessian computations use exact automatic differentiation rather than finite differences, removing a common source of silent numerical error. Code to reproduce every figure in this paper is a direct implementation of the definitions and theorems in Section 4.4 and the section preceding it, and does not depend on any component not described in the text.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the International Conference on Machine Learning*, 2022.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the International Conference on Machine Learning*, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *Proceedings of the International Conference on Machine Learning*, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the International Conference on Machine Learning*, 2015.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *Advances in Neural Information Processing Systems*, 2021.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *IEEE International Conference on Big Data*, 2020.