Scale-Invariant Empirical-Bayes Laplace Approximation for ReLU Networks

Shivam Pal[1] **and Piyush Rai**[1,2]
[1]Department of Computer Science and Engineering, IIT Kanpur
[2]Department of Intelligent Systems, IIT Kanpur
{pshivam,piyush}@cse.iitk.ac.in

Abstract

ReLU networks are well-known to have continuous scale symmetries; infinitely many reparameterized weight vectors that represent the same function. Dinh et al. [2017] have shown that, by reparameterizing, the geometry of the loss landscape (e.g., curvature) of these networks can be changed arbitrarily without affecting the underlying function. Thus, any quantity that depends on the network's geometry should be defined with care. In this paper, we show that the Empirical Bayes (EB) Laplace's approximation is not well-defined for ReLU networks, mainly due to Hessian dependence in computing the marginal likelihood, which is not invariant under ReLU scale symmetries. Thus, we can get different prior precisions, posterior covariances, and predictive uncertainties for a network having identical predictions under ReLU scaling. We define family of prior precision fields indexed by γ and introduce γ-*canonicalization*, which first converts the network to its *canonical representation* by balancing the ratio of incoming to outgoing weights, and then fits the Empirical Bayes (EB) Laplace to the canonical representation. This procedure is equivalent to transporting the isotropic Gaussian prior from the canonical representation to the original coordinates, by producing a ReLU-scale-compatible precision field. This makes the marginal likelihood invariant under ReLU scaling, while allowing γ to be selected by maximizing the marginal likelihood. The resulting EB Laplace approximation is well-defined on ReLU scale-equivalence classes and yields consistent prior precision selection and predictive uncertainties across functionally identical parameterizations.

1 Introduction

The Laplace's approximation is a simple post-hoc method for approximate Bayesian inference in deep neural networks. It fits a local Gaussian posterior whose mean is a maximum-a-posteriori (MAP) solution, and covariance is determined by curvature and prior precision [MacKay, 1992, Ritter et al., 2018, Daxberger et al., 2021]. The prior precision is often selected using marginal likelihood, which can be estimated solely from training data and without any validation data. This is also known as empirical Bayes (EB) or type-II maximum likelihood learning. Recent progress in structured curvature approximations, linearized prediction, and software support [Ritter et al., 2018, Daxberger et al., 2021, Immer et al., 2021a] has made the EB Laplace approximation an attractive method for post-hoc uncertainty estimation and Bayesian model selection.

In this paper, we study the transformation properties of marginal likelihood under ReLU scale symmetries. For ReLU networks, the marginal likelihood is not well-defined on the learned predictor. Due to the positive homogeneity property of ReLU, multiplying the incoming and outgoing weights of a hidden unit by $\alpha > 0$ and $1/\alpha$, respectively, makes the network function unchanged. These transformations generate ReLU *scale orbits*; continuous families of parameter vectors having the same

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

function [Neyshabur et al., 2015a,b, Dinh et al., 2017]. Along such an orbit, the marginal likelihood is not invariant mainly due to the log-determinant of the curvature and the norm of the chosen orbit which do not transform compatibly. Consequently, standard EB Laplace can give different marginal likelihoods, prior precisions, and linearized predictive uncertainties to the parameterizations that represent exactly the same predictor.

We fix this problem by introducing γ-*canonicalization* where $\gamma \in [0, 1]$ controls the choice of canonical representation. Our procedure first converts the network (MAP solution) to its canonical representation by balancing the ratio of incoming and outgoing weights of the hidden units of the network, then fitting the EB Laplace on the canonical representation. This generalizes the balanced and one-sided normalizations studied in scale-symmetric networks [Stock et al., 2019, Saul, 2023]. This procedure is equivalent to transporting the isotropic Gaussian prior from the canonical representation to the original coordinates, by producing a ReLU-scale-compatible precision field. We show that, for every γ, the resulting generalized marginal likelihood is invariant on the ReLU scale orbit. Thus, it makes the EB Laplace approximation explicitly depend on the learned function and not on arbitrary parametrization of the function that is obtained due to training artifacts. γ is selected jointly with prior precision using generalized marginal likelihood.

Empirically, we further show that the standard EB Laplace varies substantially across functionally identical ReLU rescaling, while γ-canonicalized EB Laplace becomes invariant. We also find that γ-canonicalized EB Laplace gives better uncertainty estimates than standard Laplace, although our main guarantees are for the invariance.

2 Empirical-Bayes Laplace Approximation

Let $\mathbf{f}(\cdot\,; \mathbf{w})\colon \mathcal{X} \to \mathbb{R}^C$ be a neural network with parameters $\mathbf{w} \in \mathbb{R}^D$, where C is the number of output classes and D is the number of parameters. Let $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ denote the training data. We assume that a MAP solution $\hat{\mathbf{w}}$ is obtained by minimizing the regularized training loss, and $\log p(\mathcal{D} \mid \mathbf{w})$ denotes the log-likelihood evaluated at $\mathbf{w}$.

Laplace Approximation. The Laplace approximation replaces the intractable posterior $p(\mathbf{w} \mid \mathcal{D})$ by a Gaussian centered at $\hat{\mathbf{w}}$. Given a likelihood-curvature approximation $\mathbf{C}_{\hat{\mathbf{w}}} \in \mathbb{S}_+^D$ and a prior precision $\lambda > 0$, the approximate posterior is $q(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \hat{\mathbf{w}}, \mathbf{\Sigma}_{\hat{\mathbf{w}}}), \mathbf{\Sigma}_{\hat{\mathbf{w}}} = (\mathbf{C}_{\hat{\mathbf{w}}} + \lambda \mathbf{I})^{-1}$. The exact Hessian is often not tractable, and also not guaranteed to be positive definite for neural networks. Common choices for curvature approximations are the generalized Gauss–Newton (GGN) matrix and the empirical Fisher matrix [MacKay, 1992, Ritter et al., 2018, Daxberger et al., 2021]. The GGN and Fisher are positive semi-definite by construction and are often preferred in practice. We will assume the GGN approximation of curvature throughout the paper.

Prior Precision and Empirical-Bayes Selection. The prior precision λ controls the scale of the Gaussian posterior and affects the uncertainty estimates. It can be selected using a validation set. However a common empirical-Bayes alternative is to select λ by maximizing the approximate marginal likelihood on the training data. Up to constants independent of λ, the log marginal likelihood is

$$\mathcal{L}_{\text{std}}(\lambda; \hat{\mathbf{w}}) = \log p(\mathcal{D} \mid \hat{\mathbf{w}}) - \tfrac{1}{2} \log |\mathbf{C}_{\hat{\mathbf{w}}} + \lambda \mathbf{I}| + \tfrac{D}{2} \log \lambda - \tfrac{\lambda}{2} \|\hat{\mathbf{w}}\|^2. \tag{1}$$

Note that for deep neural networks, the marginal likelihood cannot be computed exactly. Eq. 1 is only an approximation, widely used in Bayesian deep learning [MacKay, 1992, Immer et al., 2021a,b, Antoran et al., 2022a]. The prior precision is then given by

$$\lambda^\star(\hat{\mathbf{w}}) \in \arg\max_{\lambda > 0} \mathcal{L}_{\text{std}}(\lambda; \hat{\mathbf{w}}). \tag{2}$$

Linearized Predictive Distribution. For a test input x, let $\mathbf{J}_{\hat{\mathbf{w}}}(x) \in \mathbb{R}^{C \times D}$ denote the Jacobian of $\mathbf{f}(\cdot\,; \mathbf{w})$ with respect to $\mathbf{w}$, evaluated at $\hat{\mathbf{w}}$. The Jacobian-linearized distribution over network outputs is Gaussian with mean $\boldsymbol{\mu}(x) = \mathbf{f}(x; \hat{\mathbf{w}})$, and covariance $\mathbf{K}_\lambda(x) = \mathbf{J}_{\hat{\mathbf{w}}}(x)\, \mathbf{\Sigma}_{\hat{\mathbf{w}}}\, \mathbf{J}_{\hat{\mathbf{w}}}(x)^\top$.

3 Parameter-Space Symmetries in Neural Networks

A parameter-space symmetry of a neural network is a transformation of the network parameter that leaves the network function unchanged. Such symmetries matter for post-hoc Bayesian inference

because a quantity that changes under an exact function-preserving transformation is a property of the parameter representative, not of the learned function of the neural network.

Definition 3.1. A map $T\colon \mathbb{R}^D \to \mathbb{R}^D$ is a parameter-space symmetry of $\mathbf{f}(\cdot\,;\mathbf{w})$ if

$$\mathbf{f}(x; T(\mathbf{w})) = \mathbf{f}(x; \mathbf{w}) \qquad \text{for all } x \in \mathcal{X} \text{ and } \mathbf{w} \in \mathbb{R}^D.$$

Given a set of symmetries $\mathcal{G}$, a scalar-valued function $\phi(\mathbf{w})$ is *invariant* under $\mathcal{G}$ if $\phi(T(\mathbf{w})) = \phi(\mathbf{w})$ for all $T \in \mathcal{G}$. A matrix-valued function $\mathbf{\Phi}(\mathbf{w})$ is *equivariant* under $\mathcal{G}$ if it transforms predictably under each $T \in \mathcal{G}$, meaning there exists a known map ρ_T such that $\mathbf{\Phi}(T(\mathbf{w})) = \rho_T(\mathbf{\Phi}(\mathbf{w}))$. Invariance is the special case where ρ_T is the identity.

Neural networks have several exact parameter-space symmetries [Zhao et al., 2026]. For example, permutation symmetry, i.e., relabeling the hidden units of a layer and applying the corresponding permutation to adjacent weight matrices leaves the function unchanged [Hecht-Nielsen, 1990, Entezari et al., 2022, Ainsworth et al., 2023]. In this work, we focus only on continuous symmetry specific to ReLU networks.

3.1 ReLU Scale Orbits

ReLU networks have continuous exact symmetries arising due to their homogeneity property. For any $\alpha > 0$, $\mathrm{ReLU}(\alpha z) = \alpha\,\mathrm{ReLU}(z)$. Thus, the scale of a hidden unit can be redistributed between its incoming and outgoing weights without changing the network function. Consider a feedforward ReLU network with weight matrix $\mathbf{W}^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and bias vector $\mathbf{b}^{(\ell)} \in \mathbb{R}^{d_\ell}$ at layer ℓ. For a hidden unit (ℓ, i) and a scalar $\alpha > 0$, the rescaling

$$\mathbf{W}^{(\ell)}[i,:] \leftarrow \alpha\,\mathbf{W}^{(\ell)}[i,:], \qquad \mathbf{b}^{(\ell)}[i] \leftarrow \alpha\,\mathbf{b}^{(\ell)}[i], \qquad \mathbf{W}^{(\ell+1)}[:,i] \leftarrow \alpha^{-1}\mathbf{W}^{(\ell+1)}[:,i] \quad (3)$$

leaves the network function unchanged. Composing this over all hidden units creates a group $\mathcal{G}$ of exact function-preserving transformations. Let $\mathbf{w} \in \mathbb{R}^D$ be the flattened parameter and $\boldsymbol{\alpha} = (\alpha_{\ell,i}) > 0$ collects one positive scalar per hidden unit. Then, the corresponding transformation is $T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$. Here, $\mathbf{D}_{\boldsymbol{\alpha}} \in \mathbb{R}^{D\times D}$ is a diagonal matrix corresponding to $\boldsymbol{\alpha}$. The *ReLU scale orbit* of $\mathbf{w}$ is the equivalence class $[\mathbf{w}] = \{T_{\boldsymbol{\alpha}}(\mathbf{w}) \mid \boldsymbol{\alpha} > 0\}$. Every element of $[\mathbf{w}]$ represents exactly the same function.

3.2 Transformation Laws under ReLU Rescaling

Let $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$ be a ReLU rescaling. By symmetry, $\mathbf{f}(x;\mathbf{v}) = \mathbf{f}(x;\mathbf{w})$. Since $\boldsymbol{w} = \mathbf{D}_\alpha^{-1}\boldsymbol{v}$, and using the chain rule, the Jacobian transforms as follows.

$$\mathbf{J}_{\boldsymbol{v}}(x) = \frac{\partial \mathbf{f}}{\partial \boldsymbol{v}} = \frac{\partial \mathbf{f}}{\partial \boldsymbol{w}}\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{v}} = \mathbf{J}_{\boldsymbol{w}}(x)\,\mathbf{D}_\alpha^{-1}. \quad (4)$$

The generalized Gauss–Newton (GGN) curvature matrix can be written as $\mathbf{C}_{\boldsymbol{v}} = \sum_n \mathbf{J}_{\boldsymbol{v}}(x_n)^\top \mathbf{\Lambda}_n \mathbf{J}_{\boldsymbol{v}}(x_n)$. And this transforms by congruence,

$$\mathbf{C}_{\boldsymbol{v}} = \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{C}_{\boldsymbol{w}}\mathbf{D}_{\boldsymbol{\alpha}}^{-1}. \quad (5)$$

This congruence law also holds for standard KFAC curvature blocks under ReLU rescaling. We give the layerwise Kronecker-factor derivation in Appendix D.

4 Orbit Dependence of Standard Empirical-Bayes Laplace

Definition 4.1 (Orbit consistency)**.** An empirical-Bayes Laplace procedure is *orbit-consistent* if, for every ReLU rescaling $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w})$, the marginal likelihood, selected prior precision, and covariance of the predictive distribution are the same on the ReLU-scale orbit.

$$\mathcal{L}(\lambda;\mathbf{v}) = \mathcal{L}(\lambda;\mathbf{w}), \qquad \lambda^\star(\mathbf{v}) = \lambda^\star(\mathbf{w}), \qquad \mathbf{K}_\lambda(\mathbf{v};x) = \mathbf{K}_\lambda(\mathbf{w};x). \quad (6)$$

Standard EB Laplace approximation is not orbit-consistent. Let $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$ be a ReLU rescaling. A sufficient condition for $\mathcal{L}_{\mathrm{std}}$ to be orbit-consistent is that all four terms in (1) are invariant under this rescaling. The likelihood term satisfies this since the network function is unchanged,

$\log p(\mathcal{D} \mid \mathbf{v}) = \log p(\mathcal{D} \mid \mathbf{w})$. However, the prior quadratic term satisfies $\|\mathbf{v}\|^2 = \|\mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}\|^2 \neq \|\mathbf{w}\|^2$ in general. And the log-determinant term is also not invariant. Although the curvature transforms by congruence (5), the isotropic component $\lambda\mathbf{I}$ does not transform with it, so $\mathbf{C}_{\mathbf{v}} + \lambda\mathbf{I}$ is not a congruence transform of $\mathbf{C}_{\mathbf{w}} + \lambda\mathbf{I}$. Thus, $\mathcal{L}_{\mathrm{std}}(\lambda; \mathbf{w})$ is not constant on $[\mathbf{w}]$, and the prior precision $\lambda^\star(\mathbf{w})$ is selected by maximizing marginal likelihood. As we show below, the selected prior precision can also vary along the orbit. A detailed derivation of proposition 4.2 is provided in Appendix A.

Proposition 4.2 (Orbit dependence of standard EB-Laplace)**.** *Consider the scalar ReLU regression model $f(x;\, a, c) = c\,\mathrm{ReLU}(ax)$ with $a, c > 0$, Gaussian observation noise of variance σ_y^2, and inputs $x_n > 0$. The rescaling $(a, c) \mapsto (\alpha a,\, c/\alpha)$ for $\alpha > 0$ preserves the realized predictor. Let $S = \sum_{n=1}^{N} x_n^2$, $\kappa = S/\sigma_y^2$, and $q(\alpha) = (\alpha a)^2 + (c/\alpha)^2$. The unique positive stationary point of $\mathcal{L}_{\mathrm{std}}(\lambda; \alpha)$ in λ is*

$$\lambda^\star(\alpha) = \frac{-\kappa\, q(\alpha) + \sqrt{\kappa^2 q(\alpha)^2 + 4\kappa}}{2}. \tag{7}$$

Since $q(\alpha)$ is non-constant on the scale orbit, $\lambda^\star(\alpha)$ varies with α. Hence standard empirical-Bayes Laplace assigns different marginal likelihoods, selected prior precisions, and linearized predictive uncertainties to parameter vectors representing exactly the same function.

5 Scale-Consistent Laplace

The orbit-dependence of standard empirical-Bayes Laplace comes from a mismatch between curvature and the isotropic prior. Under a ReLU rescaling $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$, Jacobian-based curvature transforms by congruence (5), whereas the standard prior precision $\lambda\mathbf{I}$ is fixed in the original coordinates. In this section we identify the transformation law required for prior precision to be compatible with ReLU scale orbits, and prove the resulting invariance of marginal likelihood, selected prior precision and covariance of linearized prediction distribution.

5.1 Scale-Compatible Precision Fields

A scalar precision λ cannot make $\lambda\mathbf{I}$ transform like curvature under arbitrary hidden-unit rescalings. We therefore allow the prior precision to depend on the MAP solution.

Definition 5.1 (Scale-compatible precision field)**.** A map $\mathbf{w} \mapsto \boldsymbol{\Lambda}(\mathbf{w}) \in \mathbb{S}^D_{++}$ is *scale-compatible* if, for every ReLU rescaling $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$,

$$\boldsymbol{\Lambda}(\mathbf{v}) = \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\Lambda}(\mathbf{w})\mathbf{D}_{\boldsymbol{\alpha}}^{-1}. \tag{8}$$

Given such a field, define the generalized Laplace evidence

$$\mathcal{L}_{\mathrm{gen}}(\mathbf{w}) = \log p(\mathcal{D} \mid \mathbf{w}) - \tfrac{1}{2}\log|\mathbf{C}_{\mathbf{w}} + \boldsymbol{\Lambda}(\mathbf{w})| + \tfrac{1}{2}\log|\boldsymbol{\Lambda}(\mathbf{w})| - \tfrac{1}{2}\mathbf{w}^\top\boldsymbol{\Lambda}(\mathbf{w})\mathbf{w}. \tag{9}$$

When $\boldsymbol{\Lambda}(\mathbf{w}) = \lambda\mathbf{I}$, this reduces to the standard isotropic objective in (1).

Theorem 5.2 (Evidence invariance)**.** *Let $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$ be a ReLU rescaling. Suppose that the curvature satisfies* (5) *and that $\boldsymbol{\Lambda}$ is scale-compatible. Then $\mathcal{L}_{\mathrm{gen}}(\mathbf{v}) = \mathcal{L}_{\mathrm{gen}}(\mathbf{w})$.*

Proof of the theorem 5.2 is given in Appendix B. If $\{\boldsymbol{\Lambda}_\eta(\mathbf{w})\}_{\eta\in\mathcal{H}}$ is a family of scale-compatible precision fields, then $\mathcal{L}_{\mathrm{gen}}(\mathbf{v}; \eta) = \mathcal{L}_{\mathrm{gen}}(\mathbf{w}; \eta)$ for every η. Therefore, the set of hyperparameters that are obtained by maximizing the marginal likelihood will also be constant on ReLU scale orbit.

5.2 Posterior Equivariance and Predictive Invariance

Scale compatibility also gives the correct transformation rule for the posterior covariance and the covariance of the predictive distribution under the Laplace approximation. Let posterior covariance be $\boldsymbol{\Sigma}_{\mathbf{w}} = (\mathbf{C}_{\mathbf{w}} + \boldsymbol{\Lambda}(\mathbf{w}))^{-1}$, and $\boldsymbol{\Sigma}_{\mathbf{v}} = (\mathbf{C}_{\mathbf{v}} + \boldsymbol{\Lambda}(\mathbf{v}))^{-1}$. Then $\boldsymbol{\Sigma}_{\mathbf{v}} = \mathbf{D}_{\boldsymbol{\alpha}}\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{D}_{\boldsymbol{\alpha}}$. Thus, the posterior covariance is equivariant under scale-compatible precision field. This is the expected behavior under transformation. The induced Jacobian-linearized predictive covariance, however, is invariant.

Theorem 5.3 (Linearized predictive invariance)**.** *Let $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$. Suppose that the curvature satisfies* (5) *and that $\boldsymbol{\Lambda}$ is scale-compatible. Then, for every test input x,*

$$\mathbf{f}(x; \mathbf{v}) = \mathbf{f}(x; \mathbf{w}), \qquad \mathbf{J}_{\mathbf{v}}(x)\boldsymbol{\Sigma}_{\mathbf{v}}\mathbf{J}_{\mathbf{v}}(x)^\top = \mathbf{J}_{\mathbf{w}}(x)\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{J}_{\mathbf{w}}(x)^\top. \tag{10}$$

Proof. The mean is invariant because $\mathbf{v}$ and $\mathbf{w}$ represent the same network function. For the covariance, using the transformation rule for Jacobian from Eq. 4, $\mathbf{J}_{\mathbf{v}}(x) = \mathbf{J}_{\mathbf{w}}(x)\mathbf{D}_{\boldsymbol{\alpha}}^{-1}$

$$\begin{aligned}\mathbf{J}_{\mathbf{v}}(x)\boldsymbol{\Sigma}_{\mathbf{v}}\mathbf{J}_{\mathbf{v}}(x)^{\top} &= \mathbf{J}_{\mathbf{w}}(x)\mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{D}_{\boldsymbol{\alpha}}\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{D}_{\boldsymbol{\alpha}}\mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{J}_{\mathbf{w}}(x)^{\top} \\ &= \mathbf{J}_{\mathbf{w}}(x)\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{J}_{\mathbf{w}}(x)^{\top}.\end{aligned}$$

□

5.3 Transported Prior from a Canonical Section

We have now seen that, to make the EB Laplace scale-consistent, the precision field should be scale-compatible. Here we show how to construct such a precision field. Let's assume that $C_\gamma \colon \mathbb{R}^D \to \mathbb{R}^D$ selects a canonical representative $\bar{\mathbf{w}} = C_\gamma(\mathbf{w})$ from a ReLU scale orbit and satisfies the following property

$$C_\gamma(T_{\boldsymbol{\alpha}}(\mathbf{w})) = C_\gamma(\mathbf{w}) \text{ for every ReLU rescaling } T_{\boldsymbol{\alpha}}. \tag{11}$$

The above property Eq. 11 is also called a *section* property and C_γ is called *canonical section*. We assume that ReLU scale orbit is non-degenerate and the scale displacement between $\mathbf{w}$ and $C_\gamma(\mathbf{w})$ is well defined. Since $\mathbf{w}$ and $C_\gamma(\mathbf{w})$ lie on the same orbit, there exists a positive diagonal matrix $\mathbf{D}(\mathbf{w};\gamma)$ such that $\mathbf{w} = \mathbf{D}(\mathbf{w};\gamma)C_\gamma(\mathbf{w})$. This matrix records the scale displacement from the canonical representative to the original representative. Using an isotropic Gaussian prior with precision λ at the canonical representative,

$$p_\gamma(\bar{\mathbf{w}}) \propto \exp\!\left(-\tfrac{\lambda}{2}\|\bar{\mathbf{w}}\|^2\right), \qquad \bar{\mathbf{w}} = C_\gamma(\mathbf{w}). \tag{12}$$

Using $\bar{\mathbf{w}} = \mathbf{D}(\mathbf{w};\gamma)^{-1}\mathbf{w}$, the canonical quadratic penalty becomes

$$\tfrac{\lambda}{2}\|\bar{\mathbf{w}}\|^2 = \tfrac{1}{2}\mathbf{w}^{\top}\boldsymbol{\Lambda}_\gamma(\mathbf{w};\lambda)\mathbf{w}, \quad \text{where} \quad \boldsymbol{\Lambda}_\gamma(\mathbf{w};\lambda) = \lambda\,\mathbf{D}(\mathbf{w};\gamma)^{-2}. \tag{13}$$

Thus, quadratic penalty of an isotropic Gaussian prior in canonical coordinates is equivalent to that induced by a representative-dependent (transported), diagonal precision in the original coordinates.

Proposition 5.4 (Scale compatibility of the transported prior)**.** *Suppose C_γ satisfies the section property* (11)*. Then, for every $\lambda > 0$, the transported precision field $\boldsymbol{\Lambda}_\gamma(\mathbf{w};\lambda)$ defined by* (13) *is scale-compatible.*

Corollary 5.5 (Orbit-consistent transported Laplace)**.** *Suppose C_γ satisfies the section property* (11)*. Then, for every fixed γ and $\lambda > 0$, the generalized evidence and the Jacobian-linearized predictive distribution induced by $\boldsymbol{\Lambda}_\gamma(\mathbf{w};\lambda)$ are invariant under ReLU rescaling. Moreover, the evidence-maximizing set of prior precisions is the same for all representatives of a ReLU scale orbit.*

The proof of the Proposition 5.4 is given in Appendix C. Combining Proposition 5.4 with Theorems 5.2 and 5.3 gives the section-level results. The transported-prior framework reduces the remaining problem to constructing a section C_γ satisfying Eq. (11). Next we give such a construction via γ-canonicalization.

6 γ-Canonicalization

Any canonical section C_γ satisfying (11) makes empirical-Bayes Laplace orbit-consistent. Here, we introduce a practical algorithm for selecting a canonical section. The construction is based on a family of gauge conditions that balance incoming and outgoing weight norms of the hidden units. First, we state the assumptions required for this construction and then give the iterative procedure for selecting the canonical representation.

6.1 Canonical Gauge Condition

A canonical section should not depend on the arbitrary representative returned by MAP training. ReLU scale orbits provide a natural way to define such a section. For each hidden unit, the scale symmetry redistributes the magnitude between the incoming and outgoing weights. Fixing the relative placement of this magnitude gives a gauge condition on the orbit. The parameter $\gamma \in [0, 1]$ controls where the canonical representative lies between incoming-normalized, balanced, and outgoing-normalized representatives. This construction is closely related to equi-normalization and weight-balancing

Algorithm 1 γ-Canonicalization

Require: parameters $\mathbf{w}$, gauge index $\gamma \in [0, 1]$, tolerance $\varepsilon > 0$
Ensure: representative $\bar{\mathbf{w}} = C_\gamma(\mathbf{w})$

repeat
  for each hidden unit (ℓ, i) in sweep order **do**
    compute $s_i^{(\ell)}(\mathbf{w}) > 0$ and $t_i^{(\ell)}(\mathbf{w}) > 0$ ▷ Skip dead neuron
    $\alpha \leftarrow \big(t_i^{(\ell)}(\mathbf{w})\big)^{\gamma} \big(s_i^{(\ell)}(\mathbf{w})\big)^{-(1-\gamma)}$
    $\mathbf{W}^{(\ell)}[i, :] \leftarrow \alpha\, \mathbf{W}^{(\ell)}[i, :]$
    $b_i^{(\ell)} \leftarrow \alpha\, b_i^{(\ell)}$
    $\mathbf{W}^{(\ell+1)}[:, i] \leftarrow \alpha^{-1} \mathbf{W}^{(\ell+1)}[:, i]$
  end for
until $\max_{\ell,i} |\log \alpha_{\ell,i}| < \varepsilon$
return $\mathbf{w}$

methods for scale-symmetric networks [Stock et al., 2019, Saul, 2023]. Without loss of generality we assume a feedforward ReLU network without any normalization layers. Let us define the magnitude of incoming weight of hidden unit (ℓ, i) by

$$s_i^{(\ell)}(\mathbf{w}) = \Big(\big\| \mathbf{W}^{(\ell)}[i, :] \big\|^2 + \big| b_i^{(\ell)} \big|^2 \Big)^{1/2} \tag{14}$$

and the magnitude of outgoing weight by

$$t_i^{(\ell)}(\mathbf{w}) = \big\| \mathbf{W}^{(\ell+1)}[:, i] \big\|. \tag{15}$$

Under the elementary rescaling (3) of the same unit (ℓ, i) by $\alpha > 0$, with all other unit scales held fixed, these magnitudes transform as $s_i^{(\ell)} \mapsto \alpha s_i^{(\ell)}$, $t_i^{(\ell)} \mapsto \alpha^{-1} t_i^{(\ell)}$. Thus the elementary rescaling changes the ratio between incoming and outgoing magnitudes while preserving their product. Across a multilayer network, the gauge equations are coupled because rescaling the current hidden units changes the adjacent hidden unit's incoming and outgoing weight magnitudes. For $\gamma \in [0, 1]$, the γ-*canonical gauge condition* is given by

$$\big(s_i^{(\ell)}(\bar{\mathbf{w}})\big)^{1-\gamma} = \big(t_i^{(\ell)}(\bar{\mathbf{w}})\big)^{\gamma} \qquad \text{for every hidden unit } (\ell, i). \tag{16}$$

For $\gamma = 0$, the condition becomes $s(\bar{\mathbf{w}}) = 1$, giving an incoming-normalized representative. For $\gamma = 1$, it becomes $t(\bar{\mathbf{w}}) = 1$, giving an outgoing-normalized representative. For $\gamma = 0.5$, it becomes $s(\bar{\mathbf{w}}) = t(\bar{\mathbf{w}})$, giving a balanced or minimum norm representative.

Assumption 6.1 (Canonical section)**.** For each $\gamma \in [0, 1]$ and each non-degenerate parameter vector, there exists a deterministic representative $\bar{\mathbf{w}} = C_\gamma(\mathbf{w}) \in [\mathbf{w}]$ satisfying the gauge condition (16). The resulting map C_γ is a section of the ReLU scale quotient

$$C_\gamma(T_{\boldsymbol{\alpha}}(\mathbf{w})) = C_\gamma(\mathbf{w}) \qquad \text{for every ReLU rescaling } T_{\boldsymbol{\alpha}}. \tag{17}$$

Assumption 6.1 requires the existence of a canonical representative. In our implementation, we compute C_γ by the iterative algorithm 1.

6.2 Iterative Construction

To compute the canonical representative, we repeatedly apply elementary rescalings. Holding all other unit scales fixed, the rescaling factor that enforces (16) for unit (ℓ, i) is

$$\alpha_{\ell,i}(\mathbf{w}; \gamma) = \big(t_i^{(\ell)}(\mathbf{w})\big)^{\gamma} \big(s_i^{(\ell)}(\mathbf{w})\big)^{-(1-\gamma)}. \tag{18}$$

Applying (3) with this factor sets the local residual $r_{\ell,i}(\mathbf{w}; \gamma) = (1 - \gamma) \log s_i^{(\ell)}(\mathbf{w}) - \gamma \log t_i^{(\ell)}(\mathbf{w})$ to zero for that unit. Since the residuals are coupled across layers for interior values of γ, we sweep over hidden units and repeat until the maximum update size is below a tolerance. We do not provide a proof of convergence for the γ-canonicalization algorithm. It is left for future work. For $\gamma = 0.5$, the convergence proof is given by [Saul, 2023]. In the experiments, convergence is assessed by the maximum gauge residual and by verifying that canonicalization preserves the network outputs to numerical precision. We find the algorithm always converges within at most 200 sweeps. We use a first-to-last sweep order for $\gamma < 1$ and a last-to-first sweep order for $\gamma = 1$. The algorithm converges in a single sweep for $\gamma = 0$ and $\gamma = 1$.

6.3 Selecting γ

The parameter γ specifies where the isotropic prior is placed on the incoming and outgoing scale axis. It is therefore a prior-shape hyperparameter. Different values of γ correspond to different orbit-consistent priors. Since γ indexes a prior family, it can be selected by the same marginal likelihood criterion used to select the prior precision. For a finite candidate set $\Gamma \subseteq [0, 1]$, we choose

$$(\lambda^\star, \gamma^\star) \in \arg\max_{\lambda>0,\, \gamma\in\Gamma} \mathcal{L}_{\text{gen}}(\mathbf{w}; \lambda, \gamma). \tag{19}$$

By Corollary 5.5, under Assumption 6.1, the generalized evidence is invariant on $[\mathbf{w}]$ for every fixed (λ, γ). Hence the maximizing set of (λ, γ) values is the same for every representative of the ReLU scale orbit. The choice $\gamma = 0.5$ is geometrically natural and corresponds to the minimum norm canonical representation, However, it need not be optimal under the marginal likelihood. The evidence can prefer a prior placed closer to the incoming-normalized or outgoing-normalized representative. We therefore treat γ as a hyperparameter and select it from a grid search in the experiments.

7 Experiments

We evaluate γ-canonicalization on both small and large network architectures, including ResNets. First, we perform an orbit-stress test on a Fashion-MNIST ReLU MLP to test whether the evidence, selected precision, and predictive uncertainty remain invariant across functionally identical ReLU rescaling. Next, we use a CIFAR-10 ReLU ConvNet with KFAC-Laplace to examine how γ-canonicalization affects predictive uncertainty in a convolution model. Although our theoretical guarantee is about invariance, in the next experiment we test whether γ-canonicalization can improve uncertainty estimates when the MAP representative is poorly scaled.

Orbit-stress test. We first test orbit consistency by applying post-hoc Laplace to different parameterizations of the same trained ReLU predictor. We train a two-hidden-layer ReLU MLP on Fashion-MNIST and generate $K = 20$ function-preserving ReLU rescalings of the MAP solution by sampling unit-wise $\log \alpha_i \sim \mathcal{N}(0, 1)$ and applying the elementary rescaling in Eq. 3. All representatives preserve the original logits to numerical precision. We then fit full-network diagonal-GGN EB-Laplace to each representative, either in raw coordinates or after γ-canonicalization. Complete experimental details are given in Appendix E.

Table 1: Orbit stress test on a single trained ReLU predictor. We draw $K = 20$ function-preserving ReLU rescalings with $\log \alpha_i \sim \mathcal{N}(0, 1)$ and fit diagonal-GGN EB-Laplace to each representative. Values are mean $\pm$ standard deviation across orbit samples.

Method	$\log \lambda^\star$	NLL	ECE	Acc.	$\log Z/10^4$	$\|w\|^2$	$\max \|\Delta p\|_\infty$
Standard EB	2.69 ± 0.36	0.3906 ± 0.0045	0.0682 ± 0.0059	0.8718 ± 0.0008	-8.80 ± 0.89	2769 ± 1016	2.62×10^{-1}
$\gamma = 0.00$	5.07 ± 0	0.3646 ± 0	0.0387 ± 0	0.8725 ± 0	-4.72 ± 0	156	8.35×10^{-7}
$\gamma = 0.25$	5.18 ± 0	0.3644 ± 0	0.0389 ± 0	0.8726 ± 0	-4.70 ± 0	140	1.07×10^{-6}
$\gamma = 0.50$	5.29 ± 0	0.3672 ± 0	0.0428 ± 0	0.8724 ± 0	-4.95 ± 0	134	8.35×10^{-7}
$\gamma = 0.75$	5.35 ± 0	0.3723 ± 0	0.0495 ± 0	0.8721 ± 0	-5.42 ± 0	139	7.75×10^{-7}
$\gamma = 1.00$	5.32 ± 0	0.3776 ± 0	0.0562 ± 0	0.8723 ± 0	-5.94 ± 0	156	7.75×10^{-7}

Tables 1 and 2 show that standard EB-Laplace is strongly orbit-dependent. Functionally identical representatives give different selected precisions, marginal likelihoods, and predictive probabilities. In contrast, for every γ, canonicalized Laplace removes this variation to numerical precision. The evidence selects $\gamma = 0.25$, while the balanced representative $\gamma = 0.5$ gives the smallest parameter norm. However, it is not selected. It shows that generalized marginal likelihood is choosing a prior shape and not just minimizing the canonical parameter norm. Random function-preserving rescalings of the same MAP solution substantially change standard EB-Laplace, confirming that explicit canonicalization is needed for orbit-consistency rather than relying on the scale representative returned by training.

Table 2: Spread statistics of Orbit stress test.

Orbit statistic	Standard EB	γ-canonicalized EB
$\mathrm{std}(\log \lambda^\star)$	0.356	0
$\mathrm{range}(\log \lambda^\star)$	1.421	0
Spread factor of $\lambda^\star$	$4.14\times$	$1.00\times$
$\mathrm{range}(\log Z)$	3.47×10^4 nats	$< 2 \times 10^{-2}$ nats
std(NLL)	0.0045	0
std(ECE)	0.0059	0
$\max_{k,k'} \|p_k - p_{k'}\|_\infty$	2.62×10^{-1}	$\leq 1.07 \times 10^{-6}$
$\max_k \|f_k - f_0\|_\infty$	7.63×10^{-6}	$\leq 9.54 \times 10^{-6}$
$\mathrm{Corr}(\log \lambda^\star, \|w\|^2)$	-0.971	—

CIFAR-10 ConvNet with KFAC-Laplace. We also evaluate γ-canonicalization with all-layer KFAC-Laplace on a plain ReLU ConvNet trained on CIFAR-10. KFAC curvature satisfies the same congruence transformation law under ReLU rescaling, as shown in Appendix D, so the scale-compatibility argument applies beyond diagonal curvature. Table 3 shows the results. Over three independently trained MAP seeds, standard KFAC-Laplace is strongly underconfident. Accuracy is unchanged, but the mean predictive confidence is only 0.526 for an accuracy of 0.774. Canonicalizing before fitting KFAC-Laplace, with γ selected by marginal likelihood within the canonicalized grid, selects $\gamma = 0.5$ in all three runs and substantially improves predictive quality. NLL decreases from 0.877 to 0.734 and ECE from 0.248 to 0.130, with essentially unchanged accuracy. Full experimental details and additional metrics are given in Appendix E.2.

Table 3: CIFAR-10 ConvNet with all-layer KFAC-Laplace. Mean $\pm$ std over three MAP seeds.

Method	Acc.	NLL	ECE	Conf.
Standard KFAC-LA	0.774 ± 0.003	0.877 ± 0.027	0.248 ± 0.019	0.526 ± 0.019
γ-canon KFAC-LA	0.773 ± 0.002	0.734 ± 0.003	0.130 ± 0.006	0.643 ± 0.004

CIFAR-100 ResNet with last-layer Laplace. ResNet networks do not have full ReLU-scale symmetry due to batch normalization layers. Thus, γ-canonicalization is not directly applicable. However, if we consider only the last-layer Laplace approximation, we can apply γ-canonicalization to get the canonical representation. Table 4 shows the results for this. For these networks, the trained representative is already close to the canonical representative, so canonicalization does not improve the predictive performance. However, the marginal likelihood and prior precision still change substantially under rescaling, and canonicalization removes this variation to zero. While normalization layers help stabilize the scale in these networks during training, relying on them is not ideal. Relying on BatchNorm or specific weight decay values to accidentally balance the network is a fragile strategy. If a practitioner changes the optimizer, drops weight decay, or uses a novel architecture without standard normalization, standard EB-Laplace can fail silently, producing miscalibrated uncertainties. Canonicalization explicitly removes this fragility and guarantees that the Laplace fit is robust regardless of the training dynamics. These results further demonstrate that canonicalization does not modify the predictions of an already well-balanced ResNet. Canonicalization yields behavior close to standard EB Laplace when the model is already well-scaled, and corrects the calibration issues in poorly scaled networks, as shown by the CIFAR-10 experiments 3.

Table 4: Orbit-stress tests on CIFAR-100 with ResNet-18, ResNet-50, and ResNet-101 using last-layer EB-Laplace. For each MAP, we generate $K = 10$ function-preserving channel rescalings with independent $\log \alpha_i \sim \mathcal{N}(0, 1)$. Values are mean $\pm$ standard deviation across orbit samples.

Model	Method	$\log \lambda^\star$	$\log Z/10^3$	Acc.	NLL	ECE	Conf.
ResNet-18	MAP	–	–	0.7873	0.8773	0.0977	0.8849
	Standard EB (orig.)	3.92	-10.011	0.7794	0.7906	0.0314	0.7480
	Standard EB	2.89 ± 0.13	-27.060 ± 1.307	0.7720 ± 0.0010	1.1478 ± 0.0487	0.2711 ± 0.0260	0.5010 ± 0.0252
	$\gamma = 0.00$	2.54 ± 0	-10.087 ± 0	0.7793 ± 0	0.7918 ± 0	0.0327 ± 0	0.7466 ± 0
	$\gamma = 0.25$	3.04 ± 0	-10.049 ± 0	0.7794 ± 0	0.7912 ± 0	0.0321 ± 0	0.7473 ± 0
	$\gamma = 0.50$	3.54 ± 0	-10.032 ± 0	0.7794 ± 0	0.7909 ± 0	0.0318 ± 0	0.7476 ± 0
	$\gamma = 0.75$	4.04 ± 0	-10.034 ± 0	0.7794 ± 0	0.7910 ± 0	0.0318 ± 0	0.7476 ± 0
	$\gamma = 1.00$	4.53 ± 0	-10.055 ± 0	0.7793 ± 0	0.7912 ± 0	0.0319 ± 0	0.7474 ± 0
ResNet-50	MAP	–	–	0.7917	0.8801	0.1088	0.9004
	Standard EB (orig.)	4.18	-12.421	0.7870	0.7592	0.0341	0.8128
	Standard EB	3.33 ± 0.08	-37.230 ± 1.199	0.7799 ± 0.0005	1.0358 ± 0.0208	0.2169 ± 0.0128	0.5631 ± 0.0126
	$\gamma = 0.00$	4.22 ± 0	-12.715 ± 0	0.7871 ± 0	0.7598 ± 0	0.0318 ± 0	0.8096 ± 0
	$\gamma = 0.25$	3.97 ± 0	-12.507 ± 0	0.7870 ± 0	0.7594 ± 0	0.0336 ± 0	0.8117 ± 0
	$\gamma = 0.50$	3.71 ± 0	-12.432 ± 0	0.7870 ± 0	0.7593 ± 0	0.0338 ± 0	0.8125 ± 0
	$\gamma = 0.75$	3.45 ± 0	-12.485 ± 0	0.7871 ± 0	0.7595 ± 0	0.0338 ± 0	0.8121 ± 0
	$\gamma = 1.00$	3.19 ± 0	-12.668 ± 0	0.7871 ± 0	0.7599 ± 0	0.0325 ± 0	0.8103 ± 0
ResNet-101	MAP	–	–	0.8105	0.8285	0.1018	0.9122
	Standard EB (orig.)	4.17	-11.020	0.8082	0.7065	0.0195	0.8152
	Standard EB	3.34 ± 0.09	-33.887 ± 1.141	0.8014 ± 0.0006	1.0727 ± 0.0245	0.2753 ± 0.0135	0.5261 ± 0.0135
	$\gamma = 0.00$	4.31 ± 0	-11.279 ± 0	0.8083 ± 0	0.7076 ± 0	0.0187 ± 0	0.8119 ± 0
	$\gamma = 0.25$	3.98 ± 0	-11.075 ± 0	0.8082 ± 0	0.7067 ± 0	0.0192 ± 0	0.8145 ± 0
	$\gamma = 0.50$	3.65 ± 0	-11.019 ± 0	0.8083 ± 0	0.7065 ± 0	0.0196 ± 0	0.8151 ± 0
	$\gamma = 0.75$	3.31 ± 0	-11.105 ± 0	0.8081 ± 0	0.7071 ± 0	0.0188 ± 0	0.8140 ± 0
	$\gamma = 1.00$	2.96 ± 0	-11.342 ± 0	0.8079 ± 0	0.7085 ± 0	0.0177 ± 0	0.8110 ± 0

Table 5: Standard EB model selection on CIFAR-10. Evidence values are mean ± standard deviation over 20 rescalings. Selection rates are computed over all 400 pairs.

	Standard EB evidence (10^3 nats)			Width 128 selected (%)	
σ	Width 64	Width 128	Gap	Both at σ	Width 128 at 0.25
0.25	-110.6 ± 0.8	-121.7 ± 0.7	11.1	0.0	0.0
0.50	-119.3 ± 2.9	-136.4 ± 2.7	17.1	0.0	19.8
1.00	-160.6 ± 11.9	-214.3 ± 16.9	53.7	0.2	100.0

Effect on model selection. We compare two MLPs trained on CIFAR-10, with widths 64 and 128. Marginal likelihood computed using canonical representation favors the width-64 model by 12.4×10^3 nats. For each ReLU rescaling strength $\sigma \in \{0.25, 0.5, 1.0\}$ (scale $\log \alpha \sim \mathcal{N}(0, \sigma^2)$), we create 20 versions of each model without changing its network predictions and refit standard EB-Laplace, and compare all 400 pairs. As Table 5 shows, the mean evidence gap changes from 11.1×10^3 to 53.7×10^3 nats. Using the same rescaling strength for both models rarely changes their ranking, because the wider model loses more evidence in this example. However, when the width-128 model is kept at $\sigma = 0.25$, increasing the rescaling strength of the width-64 model reverses the canonical ranking in 19.8% of comparisons at $\sigma = 0.5$ and 100% at $\sigma = 1.0$. Thus, standard EB can change its model selection even though neither network's predictions change. Canonicalization keeps the evidence gap and model ranking fixed across these rescalings.

8 Related Work

Laplace approximation and empirical-Bayes prior selection. The Laplace approximation is a classical tool for Bayesian model comparison, with early evidence-based treatments established in the literature [MacKay, 1992]. Subsequent work makes it practical in neural networks through structured curvature approximations and scalable software [Ritter et al., 2018, Daxberger et al., 2021]. Empirical-Bayes selection of prior precision by marginal-likelihood maximisation has also been previously developed [Immer et al., 2021a] and extended to linearized settings [Antoran et al., 2022a]. Our work does not propose a new curvature approximation or scalable implementation. It identifies a structural defect in the standard empirical-Bayes objective. It depends on the arbitrary parameter representative. We construct a corrected version that is well-defined on ReLU scale-equivalence classes.

Classical Laplace theory and Singular models. Standard Laplace and Bayesian information criterion (BIC) asymptotics assume local identifiability and nonsingular quadratic expansion around an isolated mode [Tierney and Kadane, 1986, Kass et al., 1990]. ReLU networks violate these likelihood-level assumptions because of scale symmetry. Moreover, for deep neural networks, the Hessian need not be positive definite. A Gaussian prior can make the posterior curvature positive definite, allowing a finite local Laplace approximation to be computed. However, the usual regular-model asymptotic interpretation of such approximation is no longer valid. MacKay [1998] has shown that the accuracy of a Laplace approximation can depend on the chosen parameter basis. For non-identifiable models, the regular BIC expansion is invalid. The asymptotic expansion of the log marginal likelihood instead involves the log canonical threshold and its multiplicity, which replace the regular complexity term and may introduce an additional contribution [Watanabe, 2009]. Criteria such as wBIC [Watanabe, 2013] and sBIC [Drton and Plummer, 2017] are designed to account for this singular asymptotic behavior. We study a separate finite-sample defect of the empirical-Bayes Laplace surrogate used in practice. Our method does not correct the full singular asymptotics of neural-network marginal likelihoods. Rather, it makes that surrogate well-defined on ReLU scale orbit.

Scale symmetries and canonical representatives. Positive homogeneity of ReLU activations generates continuous families of parameter vectors representing the same function. These symmetries have been studied extensively in the context of norm-based capacity control and path-normalized optimization [Neyshabur et al., 2015a,b]. Furthermore, it has been shown that ReLU reparameterizations change sharpness measures without changing the represented function [Dinh et al., 2017], establishing that coordinate-dependent quantities are not intrinsic properties of the learned predictor. Our paper applies the same principle to empirical-Bayes Laplace. The evidence and selected prior

precision should not depend on an arbitrary representative. To construct a canonical representative, we build on iterative balancing procedures that minimize the weight norm over a scale orbit [Stock et al., 2019, Saul, 2023]. Our γ-canonicalization generalizes these procedures. Rather than using balancing to improve optimization, our method uses it to define where an isotropic Gaussian prior is placed on each orbit, turning balancing into a mechanism for scale-consistent empirical-Bayes inference.

Reparameterization invariance in approximate Bayesian inference. The most directly related work studies reparameterization invariance in approximate Bayesian inference, showing that linearized Laplace behaves differently from sampled Laplace in directions that do not change the function [Roy et al., 2024]. A related argument suggests that Hessian-based quantities become inconsistent under reparameterization when the relevant metric structure is ignored [Kristiadi et al., 2023]. This issue has been addressed by proposing a Riemannian Laplace approximation that adapts the posterior geometry to the local shape of the log-posterior [Bergamin et al., 2023]. Our contribution is narrower and more specific than these geometric approaches. We focus exclusively on the exact continuous scale symmetry of ReLU networks and derive the precise transformation law the empirical-Bayes prior precision must satisfy for both the evidence and the linearized predictive distribution to be invariant on scale orbits. Antoran et al. [2022b] study Laplace inference in networks with normalisation layers and show that BatchNorm and related layers break the posterior-volume interpretation of the standard evidence, proposing a neural g-prior to address heterogeneous Jacobian scales. Our work addresses a different invariance failure in plain ReLU networks, where the empirical-Bayes evidence and selected prior precision depend on the arbitrary representative of a ReLU scale orbit.

Weight-space symmetries in Bayesian neural networks. Discrete weight-space symmetries, such as hidden-unit permutations, have received attention in Bayesian neural networks. Previous studies analyze how permutation symmetries affect posteriors and propose constructions that respect permutation invariance [Rossi et al., 2023, Gelberg et al., 2024]. Other works explore both permutation and scale symmetries in weight-space posteriors [Laurent et al., 2024]. Our work is complementary. We address the continuous ReLU scale symmetry rather than discrete permutation symmetry, and focus specifically on the empirical-Bayes evidence objective and prior precision selection rather than on the posterior geometry itself.

9 Limitations and future work

γ-canonicalization is only applicable to neural networks having activation functions with positive homogeneity property such as ReLU, leaky ReLU, PReLU, etc. Moreover, normalization layers such as batch normalization or layer normalization break the scale symmetry. Thus, γ-canonicalization will not work on the network having normalization layers. Extending our method to work on normalized networks would require deriving the appropriate symmetry group, canonical section, and transported prior for that architecture. We apply γ-canonicalization post-hoc after the network has been trained. It would be interesting to see how training dynamics change if we apply γ-canonicalization during training. We leave such investigations to future work.

10 Conclusion

Standard EB Laplace approximation for ReLU networks is not invariant under ReLU scale symmetry. It gives different marginal likelihoods, prior precisions, and predictive distributions for function-preserving ReLU rescalings. We introduce γ-canonicalized EB Laplace approximation that fixes the ReLU scale dependence issue of standard EB Laplace. It selects a canonical representation and places an isotropic Gaussian prior on that, which is equivalent to transporting the isotropic Gaussian prior from the canonical representation to the original coordinates by producing a precision field. This precision field transforms compatibly with curvature, which makes the generalized marginal likelihood, prior precision, and predictive distribution invariant under ReLU rescaling. Moreover, Invariance of the marginal likelihood makes model selection via EB Laplace well-defined. Although our guarantee is for invariance, if the MAP solution has poor scale, γ-canonicalization can also give improved uncertainty estimates, as shown in our experiments.

Acknowledgments and Disclosure of Funding

SP acknowledges support from Visvesvaraya PhD fellowship. PR acknowledges support from Google Research and Satya & Rao Remala Foundation.

References

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 05 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.415. URL https://doi.org/10.1162/neco.1992.4.3.415.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux - effortless bayesian deep learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=gDcaUj4Myhn.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Mohammad Emtiyaz Khan. Scalable marginal likelihood estimation for model selection in deep learning. In *ICML*, pages 4563–4573, 2021a. URL http://proceedings.mlr.press/v139/immer21a.html.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401, Paris, France, 03–06 Jul 2015a. PMLR. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *NIPS*, pages 2422–2430, 2015b. URL http://papers.nips.cc/paper/5797-path-sgd-path-normalized-optimization-in-deep-neural-networks.

Pierre Stock, Benjamin Graham, Rémi Gribonval, and Hervé Jégou. Equi-normalization of neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1gEqiC9FX.

Lawrence K. Saul. Weight-balancing fixes and flows for deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uaHyXxyp2r.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 703–711. PMLR, 13–15 Apr 2021b. URL https://proceedings.mlr.press/v130/immer21a.html.

Javier Antoran, David Janz, James U Allingham, Erik Daxberger, Riccardo Rb Barbano, Eric Nalisnick, and Jose Miguel Hernandez-Lobato. Adapting the linearised Laplace model evidence for modern deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 796–821. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/antoran22a.html.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=jLpWq5QY6I.

Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. In Rolf ECKMILLER, editor, *Advanced Neural Computers*, pages 129–135. North-Holland, Amsterdam, 1990. ISBN 978-0-444-88400-8. doi: https://doi.org/10.1016/B978-0-444-88400-8.50019-4. URL https://www.sciencedirect.com/science/article/pii/B9780444884008500194.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dNigytemkL.

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986. doi: 10.1080/01621459.1986.10478240. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240.

Robert E. Kass, Luke Tierney, and Joseph B. Kadane. The validity of posterior expansions based on laplace's method. In S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, editors, *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, pages 473–488. North-Holland, Amsterdam, 1990.

David J. C. MacKay. Choice of basis for laplace approximation. *Machine Learning*, 33(1):77–86, 1998. doi: 10.1023/A:1007558615313. URL https://doi.org.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, USA, 2009. ISBN 0521864674.

Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(27):867–897, 2013. URL http://jmlr.org/papers/v14/watanabe13a.html.

Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):323–380, 03 2017. ISSN 1369-7412. doi: 10.1111/rssb.12187. URL https://doi.org/10.1111/rssb.12187.

Hrittik Roy, Marco Miani, Carl Henrik Ek, Philipp Hennig, Marvin Pförtner, Lukas Tatzel, and Søren Hauberg. Reparameterization invariance in approximate bayesian inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=204YOrDHny.

Agustinus Kristiadi, Felix Dangel, and Philipp Hennig. The geometry of neural nets' parameter spaces under reparametrization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vtLNwa6uX0.

Federico Bergamin, Pablo Moreno-Muñoz, Søren Hauberg, and Georgios Arvanitidis. Riemannian laplace approximations for bayesian neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bHS7qjLOAy.

Javier Antoran, James Urquhart Allingham, David Janz, Erik Daxberger, Eric Nalisnick, and José Miguel Hernández-Lobato. Linearised laplace inference in networks with normalisation layers and the neural g-prior. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022b. URL https://openreview.net/forum?id=uUH8x-h9zdB.

Simone Rossi, Ankit Singh, and Thomas Hannagan. On permutation symmetries in bayesian neural network posteriors: a variational perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=kj33zJ9Vue.

Yoav Gelberg, Tycho F. A. van der Ouderaa, Mark van der Wilk, and Yarin Gal. Variational inference failures under model symmetries: Permutation invariant posteriors for bayesian neural networks. In Sharvaree Vadgama, Erik Bekkers, Alison Pouplin, Sekou-Oumar Kaba, Robin Walters, Hannah Lawrence, Tegan Emerson, Henry Kvinge, Jakub Tomczak, and Stephanie Jegelka, editors, *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, pages 233–248. PMLR, 29 Jul 2024. URL https://proceedings.mlr.press/v251/gelberg24a.html.

Olivier Laurent, Emanuel Aldea, and Gianni Franchi. A symmetry-aware exploration of bayesian neural network posteriors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FOSBQuXgAq.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

A Proof of Proposition 4.2(Scalar Orbit Dependence)

Model. Consider $f(x;\, a, c) = c\,\mathrm{ReLU}(ax)$ with $a, c > 0$, Gaussian observation noise of variance σ_y^2, and inputs $x_n > 0$. Since $a > 0$, $\mathrm{ReLU}(ax_n) = ax_n$, so $f(x_n; a, c) = acx_n$. The rescaling $(a, c) \mapsto (\alpha a,\, c/\alpha)$ preserves the realized predictor on the dataset for every $\alpha > 0$.

Curvature. The Jacobian at (a, c) for a single input x_n is $\mathbf{J}_n(a, c) = (cx_n,\, ax_n)$. For Gaussian regression with noise variance σ_y^2, the GGN matrix is

$$\mathbf{G}(a, c) = \sum_{n=1}^{N} \frac{1}{\sigma_y^2} \mathbf{J}_n^\top \mathbf{J}_n = \frac{S}{\sigma_y^2} \begin{pmatrix} c^2 & ac \\ ac & a^2 \end{pmatrix}, \qquad S = \sum_{n=1}^{N} x_n^2. \tag{20}$$

This matrix has rank one with nonzero eigenvalue $\mu(\alpha) = \kappa q(\alpha)$, where $\kappa = S/\sigma_y^2$ and $q(\alpha) = (\alpha a)^2 + (c/\alpha)^2$.

Evidence along the orbit. Along the scale orbit, the likelihood is constant and $\|(a(\alpha), c(\alpha))\|^2 = q(\alpha)$. The eigenvalues of $\mathbf{G}(\alpha) + \lambda\mathbf{I}$ are λ and $\mu(\alpha) + \lambda$, so $|\mathbf{G}(\alpha) + \lambda\mathbf{I}| = \lambda(\mu(\alpha) + \lambda)$. Up to constants independent of λ and α, the standard empirical-Bayes objective reduces to

$$\mathcal{L}_{\mathrm{std}}(\lambda; \alpha) = \frac{1}{2} \log \frac{\lambda}{\kappa q(\alpha) + \lambda} - \frac{\lambda}{2} q(\alpha). \tag{21}$$

Empirical-Bayes optimum. Differentiating (21) with respect to λ and setting the derivative to zero gives

$$\frac{1}{\lambda} - \frac{1}{\kappa q(\alpha) + \lambda} = q(\alpha). \tag{22}$$

Combining fractions,

$$\frac{\kappa}{\lambda(\kappa q(\alpha) + \lambda)} = 1. \tag{23}$$

Since $q(\alpha) > 0$, this simplifies to $\lambda(\kappa q(\alpha) + \lambda) = \kappa$, or equivalently

$$\lambda^2 + \kappa q(\alpha)\lambda - \kappa = 0. \tag{24}$$

The unique positive root is

$$\lambda^\star(\alpha) = \frac{-\kappa q(\alpha) + \sqrt{\kappa^2 q(\alpha)^2 + 4\kappa}}{2}. \tag{25}$$

Non-invariance. The function $q(\alpha) = (\alpha a)^2 + (c/\alpha)^2$ is strictly convex in $\log \alpha$, minimized at $\alpha^2 = c/a$, and unbounded as $\alpha \to 0$ or $\alpha \to \infty$. Since $\lambda^\star$ is a strictly decreasing function of q (its derivative with respect to q is negative for all $q > 0$), $\lambda^\star(\alpha)$ is non-constant along the orbit. Hence standard empirical-Bayes Laplace assigns different marginal likelihoods, selected prior precisions, and linearized predictive uncertainties to parameter vectors representing exactly the same predictor. □

B Proof of theorem 5.2 (Evidence invariance)

Proof. The likelihood term is invariant because $\mathbf{v}$ and $\mathbf{w}$ represent the same predictor. For the posterior precision,

$$\mathbf{C}_\mathbf{v} + \mathbf{\Lambda}(\mathbf{v}) = \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\big(\mathbf{C}_\mathbf{w} + \mathbf{\Lambda}(\mathbf{w})\big)\mathbf{D}_{\boldsymbol{\alpha}}^{-1}.$$

Hence

$$|\mathbf{C}_\mathbf{v} + \mathbf{\Lambda}(\mathbf{v})| = |\mathbf{D}_{\boldsymbol{\alpha}}|^{-2} \,|\mathbf{C}_\mathbf{w} + \mathbf{\Lambda}(\mathbf{w})|\,.$$

Similarly,

$$|\mathbf{\Lambda}(\mathbf{v})| = |\mathbf{D}_{\boldsymbol{\alpha}}|^{-2} |\mathbf{\Lambda}(\mathbf{w})|.$$

The two log-determinant changes therefore cancel in (9). Finally,

$$\begin{aligned} \mathbf{v}^\top \mathbf{\Lambda}(\mathbf{v})\mathbf{v} &= (\mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w})^\top \mathbf{D}_{\boldsymbol{\alpha}}^{-1} \mathbf{\Lambda}(\mathbf{w}) \mathbf{D}_{\boldsymbol{\alpha}}^{-1} (\mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}) \\ &= \mathbf{w}^\top \mathbf{\Lambda}(\mathbf{w})\mathbf{w}. \end{aligned}$$

□

C Proof of Proposition 5.4 (Scale Compatibility of the Transported Prior)

Proof. Let $\mathbf{v} = T_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}$. By the section property, $C_\gamma(\mathbf{v}) = C_\gamma(\mathbf{w})$. Using the construction in Section (5.3),

$$\mathbf{v} = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w} = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{D}(\mathbf{w};\gamma)C_\gamma(\mathbf{v}).$$

Therefore

$$\mathbf{D}(\mathbf{v};\gamma) = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{D}(\mathbf{w};\gamma).$$

Since these matrices are diagonal,

$$\begin{aligned}\boldsymbol{\Lambda}_\gamma(\mathbf{v};\lambda) &= \lambda\mathbf{D}(\mathbf{v};\gamma)^{-2} \\ &= \lambda\big(\mathbf{D}_{\boldsymbol{\alpha}}\mathbf{D}(\mathbf{w};\gamma)\big)^{-2} \\ &= \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\big(\lambda\mathbf{D}(\mathbf{w};\gamma)^{-2}\big)\mathbf{D}_{\boldsymbol{\alpha}}^{-1} \\ &= \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\Lambda}_\gamma(\mathbf{w};\lambda)\mathbf{D}_{\boldsymbol{\alpha}}^{-1}. \end{aligned} \tag{26}$$

□

D KFAC Congruence under ReLU Rescaling

We verify that standard KFAC curvature blocks obey the same congruence law as Jacobian-based curvature under the ReLU rescalings of Section 3.1. Consider an affine layer

$$\mathbf{z}_\ell = \mathbf{W}_\ell\mathbf{a}_{\ell-1} + \mathbf{b}_\ell, \qquad \mathbf{a}_\ell = \phi(\mathbf{z}_\ell).$$

Let

$$\mathbf{D}_\ell = \mathrm{diag}(\alpha_{\ell,1},\dots,\alpha_{\ell,d_\ell})$$

collect the positive hidden-unit scaling factors at layer ℓ, with $\mathbf{D}_0 = \mathbf{D}_L = \mathbf{I}$ because input and output units are not rescaled. The ReLU parameter rescaling acts layerwise as

$$\mathbf{W}'_\ell = \mathbf{D}_\ell\mathbf{W}_\ell\mathbf{D}_{\ell-1}^{-1}, \qquad \mathbf{b}'_\ell = \mathbf{D}_\ell\mathbf{b}_\ell.$$

This is the layerwise matrix form of the elementary hidden-unit rescaling in Eq. 3. With column-major vectorization,

$$\mathrm{vec}(\mathbf{W}'_\ell) = \mathbf{S}_\ell\,\mathrm{vec}(\mathbf{W}_\ell), \qquad \mathbf{S}_\ell = \mathbf{D}_{\ell-1}^{-1}\otimes\mathbf{D}_\ell,$$

so $\mathbf{S}_\ell$ is the corresponding parameter-space block of the global rescaling matrix $\mathbf{D}_{\boldsymbol{\alpha}}$.

The parameter rescaling induces the activation transformation

$$\mathbf{z}'_\ell = \mathbf{D}_\ell\mathbf{z}_\ell, \qquad \mathbf{a}'_\ell = \mathbf{D}_\ell\mathbf{a}_\ell,$$

where the second identity follows from positive homogeneity, $\phi(\mathbf{D}_\ell\mathbf{z}_\ell) = \mathbf{D}_\ell\phi(\mathbf{z}_\ell)$. For the loss derivative

$$\boldsymbol{\delta}_\ell = \partial\mathcal{L}/\partial\mathbf{z}_\ell,$$

the chain rule gives

$$\boldsymbol{\delta}'_\ell = \mathbf{D}_\ell^{-1}\boldsymbol{\delta}_\ell.$$

With column-major vectorization, the KFAC approximation to the weight block is

$$\mathbf{C}_\ell^{\mathrm{KFAC}} = \mathbf{A}_{\ell-1}\otimes\mathbf{G}_\ell, \qquad \mathbf{A}_{\ell-1} = \mathbb{E}\big[\mathbf{a}_{\ell-1}\mathbf{a}_{\ell-1}^\top\big], \qquad \mathbf{G}_\ell = \mathbb{E}\Big[\boldsymbol{\delta}_\ell\boldsymbol{\delta}_\ell^\top\Big].$$

The transformations above imply

$$\mathbf{A}'_{\ell-1} = \mathbf{D}_{\ell-1}\mathbf{A}_{\ell-1}\mathbf{D}_{\ell-1}, \qquad \mathbf{G}'_\ell = \mathbf{D}_\ell^{-1}\mathbf{G}_\ell\mathbf{D}_\ell^{-1}.$$

Therefore, by the mixed-product property of the Kronecker product,

$$\mathbf{C}_\ell^{\mathrm{KFAC}\prime} = \mathbf{A}'_{\ell-1}\otimes\mathbf{G}'_\ell = (\mathbf{D}_{\ell-1}\otimes\mathbf{D}_\ell^{-1})\mathbf{C}_\ell^{\mathrm{KFAC}}(\mathbf{D}_{\ell-1}\otimes\mathbf{D}_\ell^{-1}) = \mathbf{S}_\ell^{-1}\mathbf{C}_\ell^{\mathrm{KFAC}}\mathbf{S}_\ell^{-1}.$$

This is the layerwise form of the global congruence law

$$\mathbf{C}_v = \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{C}_w\mathbf{D}_{\boldsymbol{\alpha}}^{-1}.$$

The KFAC factorization is compatible with the rescaling because the factorized expectation commutes with the induced activation and derivative transformations:

$$\mathbb{E}\big[\mathbf{a}'_{\ell-1}\mathbf{a}'^{\top}_{\ell-1}\big] \otimes \mathbb{E}\Big[\boldsymbol{\delta}'_{\ell}\boldsymbol{\delta}'^{\top}_{\ell}\Big] = (\mathbf{D}_{\ell-1} \otimes \mathbf{D}_{\ell}^{-1}) \Big(\mathbb{E}\big[\mathbf{a}_{\ell-1}\mathbf{a}_{\ell-1}^{\top}\big] \otimes \mathbb{E}\Big[\boldsymbol{\delta}_{\ell}\boldsymbol{\delta}_{\ell}^{\top}\Big]\Big) (\mathbf{D}_{\ell-1} \otimes \mathbf{D}_{\ell}^{-1}).$$

Biases are handled by augmenting the activation vector,

$$\bar{\mathbf{a}}_{\ell-1} = (\mathbf{a}_{\ell-1}^{\top}, 1)^{\top}, \qquad \bar{\mathbf{D}}_{\ell-1} = \operatorname{diag}(\mathbf{D}_{\ell-1}, 1).$$

The appended constant coordinate is left unscaled because it is not a hidden unit in the ReLU scale group, while the bias column itself scales with $\mathbf{D}_\ell$. Equivalently, for the augmented weight matrix $\bar{\mathbf{W}}_\ell = [\mathbf{W}_\ell \ \mathbf{b}_\ell]$,

$$\bar{\mathbf{W}}'_{\ell} = \mathbf{D}_{\ell}\bar{\mathbf{W}}_{\ell}\bar{\mathbf{D}}_{\ell-1}^{-1}.$$

The same congruence argument then applies to the augmented weight-bias block.

For convolutional layers in the standard unfolded-patch KFAC formulation, the proof carries over after replacing $\mathbf{a}_{\ell-1}$ by the unfolded input patches and applying the corresponding channel scaling to each spatial patch coordinate.

E Experimental Details

Our experiments are fully reproducible. All the experiments are conducted on a single NVIDIA A30 24GB GPU. Code is available in the supplementary material.

E.1 Orbit Stress Test

Dataset and model We use Fashion-MNIST Xiao et al. [2017], with $60{,}000$ training images and $10{,}000$ test images across 10 classes. Images are normalized using the training-set mean 0.2860 and standard deviation 0.3530, and are flattened to vectors of dimension 784. The model is a two-hidden-layer ReLU MLP with hidden dimensions $[64, 64]$ and a linear output layer. It has no normalization layers.

MAP training The MAP solution is obtained by minimizing cross-entropy with ℓ_2 weight decay 10^{-4} using Adam with learning rate 10^{-3}, batch size 256, for 10 epochs. The resulting MAP model achieves 87.67% test accuracy.

Laplace approximation We use `laplace-torch` Daxberger et al. [2021]. All Laplace fits use all parameters, diagonal GGN curvature, and a scalar prior precision λ. The prior precision is initialized at 1.0 and optimized by 100 Adam steps with learning rate 0.1 on the Laplace marginal likelihood. For prediction, we use the Jacobian-linearized predictive distribution with the probit link approximation. Model computations use `float32`; γ-canonicalization is performed in `float64` and cast back to `float32` before fitting Laplace.

Orbit generation Starting from the MAP solution $\hat{w}$, we generate $K = 20$ function-preserving representatives. For each representative, we sample independent hidden-unit log-scales

$$\log \alpha_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \sigma^2), \quad \sigma = 1.0,$$

and apply the elementary ReLU rescaling in Eq. 3 to every hidden unit. For every generated representative v_k, we verify

$$\max_{x \in \mathcal{D}_{\text{test}}} \big\| f(x; \hat{w}) - f(x; v_k) \big\|_{\infty} < 10^{-5}.$$

Thus all reported variation across orbit samples is due to the parameterization supplied to the Laplace approximation, not to any change in the predictor.

γ-canonicalization For each representative v_k and each $\gamma \in \{0,\, 0.25,\, 0.50,\, 0.75,\, 1.0\}$, we apply Algorithm 1 with tolerance $\varepsilon = 10^{-6}$ and a maximum budget of 200 sweeps. For interior values, all runs converged in fewer than 50 sweeps under the stopping criterion $\max_{\ell,i} |\log \alpha_{\ell,i}| < 10^{-6}$. The endpoints are computed by a single directional pass: forward incoming-normalization for $\gamma = 0$ and backward outgoing-normalization for $\gamma = 1$. The canonicalization cost is much smaller than the cost of fitting the Laplace approximation.

Table 6: Standard deviations under rescaling and retraining. Rescaling preserves the network function while retraining produces different learned functions. Predictive metrics use standard EB-Laplace.

Quantity	Rescaling	Retraining
Log prior precision $\log \lambda^\star$	0.356	0.016
Log evidence $\log Z$	8900	884
NLL	0.0045	0.0088
ECE	0.0059	0.0046
Accuracy	0.0008	0.0040
Squared parameter norm $\|w\|_2^2$	1016.0	4.7

Metrics We report the following quantities. NLL is the mean negative log-likelihood of the Laplace predictive distribution on the test set. ECE is the expected calibration error with 20 equal-width confidence bins Guo et al. [2017]. The quantity $\log Z$ is the Laplace marginal likelihood at the selected $\lambda^\star$. The norm $\|w\|^2$ is the squared Euclidean norm of the parameter vector supplied to Laplace; for standard EB this is the raw (possibly rescaled) representative, and for γ-canonicalized EB it is the canonical representative. The statistic $\max \|\Delta p\|_\infty$ is the maximum absolute difference in predicted class probability between any two orbit samples, computed over all $\binom{K}{2}$ pairs, all test inputs, and all classes. The spread factor is $\exp\big(\max_k \log \lambda_k^\star - \min_k \log \lambda_k^\star\big)$.

E.2 CIFAR-10 ConvNet with KFAC-Laplace

Model and training. We use a plain ReLU ConvNet without normalization layers. The channel widths are $(64, 128, 256, 256)$. The network consists of four convolutional blocks followed by global average pooling and a linear classifier. Each convolutional block contains a 3×3 convolution followed by a ReLU nonlinearity; spatial downsampling is performed by avg pooling. The model is trained on CIFAR-10 using Adam with learning rate 10^{-3} and weight decay 10^{-5} for 50 epochs. We use a cosine learning-rate schedule.

Laplace approximation. We fit all-layer empirical-Bayes Laplace with KFAC curvature using `laplace-torch`. The prior precision is optimized by marginal likelihood using `method="marglik"`. Predictions use the Jacobian-linearized predictive distribution with the probit link approximation. For the canonicalized variant, we first apply γ-canonicalization to the trained MAP solution and then fit the same KFAC-Laplace approximation. We evaluate $\gamma \in \{0, 0.1, 0.2, \ldots, 1.0\}$ and select γ by marginal likelihood within the canonicalized family. In all three seeds, the selected value is $\gamma = 0.5$.

F Within-orbit vs. between-solution variation

We compare variation between 20 function-preserving rescalings of one trained network ($\sigma = 1$), and 20 independently trained networks with the same architecture and training procedure. We fit standard EB-Laplace to every rescaled or independently trained network. Table 6 shows that the standard deviation of $\log \lambda^\star$ is approximately 22 times larger under rescaling than retraining. For $\log Z$, it is approximately 10 times larger. Predictive metrics also vary under rescaling, although retraining produces larger variation in NLL and accuracy. Variation across training runs can reflect different learned functions. In contrast, variation under function-preserving rescaling arises entirely from the parameter scales supplied to the Laplace fit. These results show that arbitrary scale choices can affect the selected prior precision and evidence more than independent retraining.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Read the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 9.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 4, 5, 6, Appendix A, B, and C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the experiments are completely reproducible. All the details is given in The appendix E. We also provide the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the complete code to reproduce the experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See the result Table 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is foundational.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: NA

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.