Scale-Invariant Empirical-Bayes Laplace Approximation for ReLU Networks

Shivam Pal [1] **Piyush Rai** [1 2]

Abstract

A post-hoc Bayesian procedure should be invariant to function-preserving reparameterizations. Standard empirical-Bayes Laplace violates this principle in ReLU networks. Positive homogeneity creates scale-equivalent parameterizations, but the isotropic evidence objective depends on the chosen representative through the prior norm and curvature log-determinant. We introduce γ-*canonicalization*, which fixes a canonical representative on each ReLU scale orbit, fits an isotropic Laplace approximation there, and transports the prior back to the original coordinates. The resulting precision field transforms compatibly with curvature, making the generalized evidence and Jacobian-linearized predictive distribution invariant for every $\gamma \in [0, 1]$. Selecting γ and the prior precision jointly by marginal likelihood yields an EB-Laplace procedure defined on ReLU scale-equivalence classes. Empirically, standard EB-Laplace varies across functionally identical rescalings, while γ-canonicalized Laplace removes this variation to zero.

1. Introduction

Laplace's approximation (MacKay, 1992) provides a simple and scalable method for post-hoc Bayesian inference in neural networks. Recent advances have made it a practical and competitive choice for deep learning by introducing diagonal, Kronecker-factored, and last-layer approximations (Ritter et al., 2018; Daxberger et al., 2021).

Let $\hat{\mathbf{w}} \in \mathbb{R}^D$ denote a trained network parameter vector, and let $\mathbf{C}_{\hat{\mathbf{w}}}$ denote the curvature matrix used by Laplace, e.g. a generalized Gauss–Newton (GGN) matrix or a compatible approximation. In post-hoc Laplace inference, the posterior around $\hat{\mathbf{w}}$ is approximated by a Gaussian $q(\mathbf{w}) = \mathcal{N}\big(\mathbf{w} \mid \hat{\mathbf{w}},\ (\mathbf{C}_{\hat{\mathbf{w}}} + \lambda \mathbf{I})^{-1}\big)$, where $\lambda > 0$ is the prior precision under an isotropic Gaussian prior $p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$. Thus the posterior precision is the sum of curvature and prior precision. This makes Laplace simple and scalable, and in practice λ is usually chosen after training by empirical Bayes. A standard choice is to maximize the marginal likelihood with respect to λ (Immer et al., 2021):

$$\begin{aligned}\mathcal{L}(\lambda; \hat{\mathbf{w}}) = \ & \log p(\mathcal{D} \mid \hat{\mathbf{w}}) - \frac{1}{2} \log |\mathbf{C}_{\hat{\mathbf{w}}} + \lambda \mathbf{I}| \\ & + \frac{D}{2} \log \lambda - \frac{\lambda}{2} \|\hat{\mathbf{w}}\|^2.\end{aligned} \tag{1}$$

This objective determines the prior precision

$$\lambda^*(\hat{\mathbf{w}}) = \arg\max_{\lambda > 0} \mathcal{L}(\lambda; \hat{\mathbf{w}}).$$

Daxberger et al., 2021 showed that the quality of uncertainty estimates is heavily dependent on the choice of Gaussian prior precision λ^*.

ReLU scale symmetry. ReLU network has scale symmetry (Neyshabur et al., 2015a), this creates a problem because the parameterization is not unique for marginal likelihood 1. A hidden ReLU unit can be rescaled without changing the function i.e. multiplying all incoming weights and bias of the unit by $\alpha > 0$ and dividing all outgoing weights by α leaves the network output unchanged for every input. Therefore many distinct parameter vectors represent exactly the same predictor. We refer to all such equivalent parameterizations as a *scale orbit*.

Inconsistency in Laplace approximation. The key issue is that the likelihood term in (1) depends only on the realized function, whereas the remaining terms depend on the parameter representative. If $\mathbf{v} = T_\alpha(\mathbf{w})$ is a rescaled version of $\mathbf{w}$ on the same ReLU scale orbit, then neural network $f(\cdot; \mathbf{v}) = f(\cdot; \mathbf{w})$, so the likelihood is unchanged. However, in general,

$$\|\mathbf{v}\|^2 \neq \|\mathbf{w}\|^2 \qquad \text{and} \qquad \mathbf{C}_{\mathbf{v}} \neq \mathbf{C}_{\mathbf{w}}.$$

As a result, both the quadratic prior term and the log-determinant term can vary along a scale orbit, even though the underlying network function is unchanged. Therefore

[1]Department of Computer Science and Engineering, IIT Kanpur [2]Department of Intelligent Systems, IIT Kanpur. Correspondence to: Shivam Pal <pshivam@cse.iitk.ac.in>, Piyush Rai <piyush@cse.iitk.ac.in>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

the evidence objective $\mathcal{L}(\lambda; \mathbf{w})$, and hence the selected prior precision $\lambda^*(\mathbf{w})$, need not be invariant across equivalent ReLU parameterizations (Antoran et al., 2022b).

Proposition 1.1 (Toy Example)**.** *Consider the scalar ReLU regression model*

$$f(x; a, c) = c\,\mathrm{ReLU}(ax), \qquad a > 0,\ c > 0,$$

with Gaussian observation noise variance σ_y^2 and inputs $x_n > 0$. The rescaling

$$(a, c) \mapsto (\alpha a, c/\alpha), \qquad \alpha > 0,$$

preserves the realized predictor. Let

$$S = \sum_{n=1}^{N} x_n^2, \qquad \kappa = \frac{S}{\sigma_y^2}, \qquad q(\alpha) = (\alpha a)^2 + (c/\alpha)^2.$$

For standard isotropic EB-Laplace, the unique positive stationary point of the marginal likelihood (Eq. 1) along this scale orbit is

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

Since $q(\alpha)$ is non-constant along the scale orbit, the stationary point $\lambda^*(\alpha)$ is also non-constant. Hence equivalent parameterizations can yield different evidence optima and different selected prior precisions. Detailed derivations are given in Appendix 8.

2. Scale-consistent Laplace

The toy example above shows that standard isotropic empirical-Bayes Laplace is not well-defined on ReLU scale-equivalence classes. To obtain a well-defined post-hoc Bayesian procedure, the prior construction must transform consistently with the same ReLU rescaling symmetry that leaves the network function unchanged. This leads to a scale-consistent formulation of Laplace.

Curvature under ReLU rescaling. Let $\mathbf{w} \in \mathbb{R}^D$ denote network parameters, and let $T_\alpha(\mathbf{w})$ denote a ReLU rescaling of hidden units, where α may represent a collection of positive unitwise scale factors. Let $\mathbf{C}_\mathbf{w}$ denote the curvature matrix used by Laplace, taken here to be GGN-type curvature or a compatible approximation. Under ReLU rescaling, the network function is unchanged, while the parameter vector transforms by a positive diagonal map,

$$T_\alpha(\mathbf{w}) = \mathbf{D}_\alpha \mathbf{w},$$

for some diagonal matrix $\mathbf{D}_\alpha \succ 0$. The corresponding Jacobian transforms as

$$\mathbf{J}_{T_\alpha(\mathbf{w})}(x) = \mathbf{J}_\mathbf{w}(x)\mathbf{D}_\alpha^{-1},$$

and therefore GGN-type curvature transforms by the congruence rule

$$\mathbf{C}_{T_\alpha(\mathbf{w})} = \mathbf{D}_\alpha^{-1}\mathbf{C}_\mathbf{w}\mathbf{D}_\alpha^{-1}.$$

Thus the likelihood depends only on the realized function, while curvature transforms covariantly along the ReLU scale orbit.

Orbit-compatible precision fields. This suggests that the prior should not be fixed as a single isotropic matrix $\lambda\mathbf{I}$ in parameter space. Instead, we let the prior precision depend on the representative:

$$\mathbf{w} \mapsto \mathbf{\Lambda}(\mathbf{w}) \in \mathbb{S}_{++}^D.$$

We say that $\mathbf{\Lambda}(\mathbf{w})$ is *scale compatible* if it transforms under ReLU rescaling by the same rule as curvature,

$$\mathbf{\Lambda}(T_\alpha(\mathbf{w})) = \mathbf{D}_\alpha^{-1}\mathbf{\Lambda}(\mathbf{w})\mathbf{D}_\alpha^{-1}.$$

Given such a precision field, we define the generalized Laplace evidence

$$\begin{aligned}\mathcal{L}_{\mathrm{gen}}(\mathbf{w}) = &\ \log p(\mathcal{D} \mid \mathbf{w}) - \frac{1}{2}\log|\mathbf{C}_\mathbf{w} + \mathbf{\Lambda}(\mathbf{w})| \\ &+ \frac{1}{2}\log|\mathbf{\Lambda}(\mathbf{w})| - \frac{1}{2}\mathbf{w}^\top\mathbf{\Lambda}(\mathbf{w})\mathbf{w}.\end{aligned} \tag{2}$$

Theorem 2.1 (ReLU scale-orbit invariance)**.** *Assume the curvature matrix satisfies the ReLU rescaling congruence*

$$\mathbf{C}_{T_\alpha(\mathbf{w})} = \mathbf{D}_\alpha^{-1}\mathbf{C}_\mathbf{w}\mathbf{D}_\alpha^{-1},$$

and let the precision field satisfy

$$\mathbf{\Lambda}(T_\alpha(\mathbf{w})) = \mathbf{D}_\alpha^{-1}\mathbf{\Lambda}(\mathbf{w})\mathbf{D}_\alpha^{-1}.$$

Then the generalized Laplace evidence in (2) is invariant under ReLU rescaling:

$$\mathcal{L}_{\mathrm{gen}}(T_\alpha(\mathbf{w})) = \mathcal{L}_{\mathrm{gen}}(\mathbf{w}) \qquad \textit{for all } \alpha > 0.$$

The theorem shows that evidence-based hyperparameter selection becomes a property of the ReLU scale-equivalence class rather than of a particular parameter representative. Proof of the theorem 2.1 is given in Appendix 9. Moreover, linearized predictive covariance is also invariant along the ReLU scale orbit (see 9.1).

The remaining question is how to construct such a representative-dependent precision field in practice. A convenient way is to choose a canonical representative within each ReLU scale-equivalence class and transport a base prior from that representative back to the original parameters. The next section realizes this idea using γ-canonicalization.

3. γ-canonicalization and transported prior

The invariance theorem characterizes which precision fields are valid, but it remains to construct one in practice. We do so by choosing a canonical representative within each ReLU scale-equivalence class and transporting a base prior from that representative.

3.1. γ-canonical representative

We now define the canonical representative used to construct the transported prior. The goal is to choose, within each ReLU scale-equivalence class, a parameterization whose per-unit incoming and outgoing magnitudes satisfy a prescribed balance condition.

Consider hidden unit (ℓ, i) in a feedforward ReLU network. We use the convention

$$\mathbf{z}^{(\ell)} = \mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)},$$

so row i of $\mathbf{W}^{(\ell)}$ contains the incoming weights of hidden unit (ℓ, i), and column i of $\mathbf{W}^{(\ell+1)}$ contains its outgoing weights. We define the incoming and outgoing magnitudes

$$s_i^{(\ell)}(\mathbf{w}) = \left(\|\mathbf{W}^{(\ell)}[i,:]\|^2 + |\mathbf{b}^{(\ell)}[i]|^2\right)^{1/2},$$

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

Under ReLU rescaling of that unit by $\alpha > 0$,

$$s_i^{(\ell)} \mapsto \alpha s_i^{(\ell)}, \qquad t_i^{(\ell)} \mapsto \alpha^{-1} t_i^{(\ell)}.$$

Hence the product

$$s_i^{(\ell)}(\mathbf{w})\, t_i^{(\ell)}(\mathbf{w})$$

is invariant under unitwise ReLU rescaling.

γ-canonicalization. Given $\gamma \in [0,1]$, we define the γ-canonical representative

$$\bar{\mathbf{w}} = \mathcal{C}_\gamma(\mathbf{w})$$

as a representative on the same ReLU scale orbit satisfying

$$\left(s_i^{(\ell)}(\bar{\mathbf{w}})\right)^{1-\gamma} = \left(t_i^{(\ell)}(\bar{\mathbf{w}})\right)^{\gamma}, \qquad \forall(\ell, i). \tag{3}$$

This condition determines how scale is distributed between incoming and outgoing parameters.

Using the orbit-invariant quantity $s_i^{(\ell)}(\mathbf{w})t_i^{(\ell)}(\mathbf{w})$, the balancing relation implies

$$\begin{aligned} s_i^{(\ell)}(\bar{\mathbf{w}}) &= \left(s_i^{(\ell)}(\mathbf{w})t_i^{(\ell)}(\mathbf{w})\right)^{\gamma}, \\ t_i^{(\ell)}(\bar{\mathbf{w}}) &= \left(s_i^{(\ell)}(\mathbf{w})t_i^{(\ell)}(\mathbf{w})\right)^{1-\gamma}. \end{aligned} \tag{4}$$

Thus γ continuously interpolates between forward-normalized and backward-normalized representatives.

Three special cases are particularly natural. For

$$\gamma = 0, \qquad s_i^{(\ell)}(\bar{\mathbf{w}}) = 1,$$

the incoming parameters are normalized; we refer to this as *forward canonicalization*. For

$$\gamma = \tfrac{1}{2}, \qquad s_i^{(\ell)}(\bar{\mathbf{w}}) = t_i^{(\ell)}(\bar{\mathbf{w}}),$$

incoming and outgoing magnitudes are equalized. we refer to this as *balanced canonicalization* (Saul, 2023). For

$$\gamma = 1, \qquad t_i^{(\ell)}(\bar{\mathbf{w}}) = 1,$$

the outgoing parameters are normalized. We refer to this as *backward canonicalization*.

Constructive update. For a hidden unit (ℓ, i) with current statistics $s_i^{(\ell)}(\mathbf{w})$ and $t_i^{(\ell)}(\mathbf{w})$, define

$$\alpha_{\ell,i}(\mathbf{w};\gamma) = \left(t_i^{(\ell)}(\mathbf{w})\right)^{\gamma}\left(s_i^{(\ell)}(\mathbf{w})\right)^{-(1-\gamma)}. \tag{5}$$

Applying the exact ReLU symmetry update

$$\begin{aligned} \mathbf{W}^{(\ell)}[i,:] &\leftarrow \alpha_{\ell,i}\mathbf{W}^{(\ell)}[i,:], \\ \mathbf{b}^{(\ell)}[i] &\leftarrow \alpha_{\ell,i}\mathbf{b}^{(\ell)}[i], \\ \mathbf{W}^{(\ell+1)}[:,i] &\leftarrow \alpha_{\ell,i}^{-1}\mathbf{W}^{(\ell+1)}[:,i] \end{aligned} \tag{6}$$

enforces (3) for that unit. Because neighboring units are coupled through shared weight matrices, a single global pass need not satisfy (3) simultaneously for all units. We therefore sweep over hidden units, typically in reverse depth order, and repeat for a small number of iterations until the balancing residual is negligible. Endpoint choices $\gamma = 0$ and $\gamma = 1$ converge in single pass, while interior values $\gamma \in (0,1)$ may require multiple sweeps. Saul, 2023 proved the convergence of γ-canonicalization for $\gamma = 0.5$.

3.2. Transport map and induced prior

The iterative updates above not only produce the canonical representative $\mathcal{C}_\gamma(\mathbf{w})$, but also induce a diagonal transport in the flattened parameterization. Each update multiplies the coordinates belonging to an incoming row and bias by $\alpha_{\ell,i}$ and the coordinates belonging to the corresponding outgoing column by $\alpha_{\ell,i}^{-1}$. Accumulating these multiplicative factors over the full sequence of sweeps yields a positive diagonal matrix

$$\mathbf{D}(\mathbf{w};\gamma) \succ 0$$

such that

$$\mathbf{w} = \mathbf{D}(\mathbf{w};\gamma)\,\mathcal{C}_\gamma(\mathbf{w}).$$

Table 1. For a single trained predictor, we draw $K{=}20$ random ReLU-scale-preserving rescalings and fit Laplace on each representative. Columns report mean $\pm$ standard deviation *across orbit samples of the same predictor*, $\max \|\Delta p\|_\infty$ is the maximum absolute difference in predicted class probability between any two orbit samples, over all test inputs and classes. Standard EB-Laplace is orbit-dependent, selected prior precision, log marginal likelihood, and predictive distribution all vary substantially between equivalent parameterisations. For every γ, γ-canonicalised Laplace is orbit-invariant to floating-point precision, directly verifying the evidence-invariance and predictive-invariance theorems. The marginal likelihood $\log Z$ ranks γ in the same order as test NLL, ECE, and accuracy, selecting $\gamma^\star = 0.25$ on this model.

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}

Equivalently, $\mathbf{D}(\mathbf{w};\gamma)$ records for each parameter the total multiplicative factor relating the canonical representative to the original parameter vector. We now place a base isotropic Gaussian prior at the canonical representative,

$$p_\gamma(\bar{\mathbf{w}}) \propto \exp\Big(-\frac{\lambda}{2}\|\bar{\mathbf{w}}\|^2\Big), \qquad \bar{\mathbf{w}} = \mathcal{C}_\gamma(\mathbf{w}),$$

and transport this prior back to the original coordinates using the diagonal map $\mathbf{D}(\mathbf{w};\gamma)$. This yields the representative-dependent precision field

$$\mathbf{\Lambda}_\gamma(\mathbf{w};\lambda) = \lambda \mathbf{D}(\mathbf{w};\gamma)^{-2}. \tag{7}$$

By construction, $\mathbf{\Lambda}_\gamma(\mathbf{w};\lambda)$ transforms by the same congruence rule as the curvature. Therefore the generalized Laplace evidence is invariant across ReLU scale-equivalent parameterizations, and the empirical-Bayes choice of λ depends only on the learned predictor rather than on the specific representative returned by training.

3.3. Selecting the canonicalization parameter γ

The transported prior in (7) depends not only on the prior precision λ, but also on the canonicalization parameter γ. Since γ determines how scale is allocated between incoming and outgoing parameter groups, it acts as a genuine modeling hyperparameter rather than a purely numerical setting. A natural choice is therefore to select γ by the same generalized evidence principle used for λ. Let $\Gamma \subseteq [0,1]$ be a search set of candidate values. We define the joint evidence-based selection rule

$$(\lambda^*, \gamma^*) \in \arg\max_{\lambda>0,\ \gamma\in\Gamma} \mathcal{L}_{\text{gen}}(\mathbf{w};\lambda,\gamma), \tag{8}$$

where $\mathcal{L}_{\text{gen}}(\mathbf{w};\lambda,\gamma)$ denotes the generalized Laplace evidence obtained using the transported precision field. Because the generalized evidence is invariant across ReLU scale-equivalent parameterizations, the resulting choice of (λ^*, γ^*) is likewise orbit-consistent. In practice, we can use grid search over $[0,1]$.

In this work, we treat γ as a single global hyperparameter shared across the network. More flexible variants are also possible, such as layer-wise parameters $\gamma_\ell \in [0,1]$ or even unit-wise parameters $\gamma_{\ell,i} \in [0,1]$, which would allow different parts of the network to canonicalize at different balances. Such structured choices may provide additional flexibility, but would increase the hyperparameter space and are left for future work.

4. Experiments

We evaluate γ-canonicalization with two experiments. In the 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 rescalings. In the second, we use a CIFAR-10 ReLU ConvNet with KFAC-Laplace to examine how γ-canonicalization affects predictive uncertainty in a convolutional model. Although our theoretical guarantee is invariance rather than improved predictive performance, the second experiment tests whether γ-canonicalization can improve uncertainty estimates when the MAP representative is poorly scaled.

4.1. 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. All representatives preserve the original logits to numerical precision, with maximum test-set logit difference below 10^{-5}. 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 10.1. Tables 1 and 3 show that standard EB-Laplace is strongly orbit-dependent: functionally

Table 2. CIFAR-10 ConvNet with all-layer KFAC-Laplace. Mean ± 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

Table 3. **Orbit-dependence pathology of standard EB-Laplace**. Spread factor is $\exp(\max_k \log \lambda_k^\star - \min_k \log \lambda_k^\star)$, the multiplicative range of selected prior precisions across equivalent parameterisations. The strong negative correlation between $\log \lambda^\star$ and $\|w\|^2$ across orbit samples confirms the depdendence on arbitrary scale. standard EB implicitly couples $\lambda^\star$ to an arbitrary parameter norm. Under our construction every listed quantity is a function of the equivalence class, so the across-orbit statistics collapse to zero.

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	$< 2\times10^{-2}$ nats
$\mathrm{std}(\mathrm{NLL})$	0.0045	0
$\mathrm{std}(\mathrm{ECE})$	0.0059	0
$\max_{k,k'} \|p_k - p_{k'}\|_\infty$	2.62×10^{-1}	$\leq 1.07\times10^{-6}$
$\mathrm{Corr}(\log \lambda^\star, \|w\|^2)$	-0.971	—

identical representatives lead to different selected precisions, marginal likelihoods, and predictive probabilities. In contrast, for every tested 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 but is not selected, indicating that marginal likelihood is choosing a prior shape rather than merely 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.

4.2. 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 also satisfies the same congruence transformation law under ReLU rescaling, thus the scale-compatibility argument applies beyond diagonal curvature. Table 2 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 10.2.

5. Limitations

Our invariance guarantees are restricted to architectures with exact ReLU scale symmetries and curvature approximations that transform by the corresponding congruence law. The construction therefore does not directly apply to networks with normalization layers such as Batch Normalization or Layer Normalization. These layers change the scale-equivalence structure through normalization statistics and affine parameters, so the simple incoming/outgoing ReLU rescaling orbit used by γ-canonicalization is no longer the relevant quotient. Extending the method to normalized networks would require deriving the appropriate symmetry group, canonical section, and transported prior for that architecture. We leave such investigations to future work.

6. Conclusion

Standard empirical-Bayes Laplace is formulated in parameter coordinates, but ReLU networks represent the same predictor by an entire scale-equivalence class. This makes the standard isotropic construction gauge-dependent: equivalent parameterizations can yield different evidence values, selected prior precisions, posterior covariances, and linearized predictive uncertainties. We introduced γ-canonicalized Laplace, which selects a canonical representative on each scale orbit, places an isotropic Gaussian prior there, and transports the prior back to the original coordinates. The resulting precision field transforms compatibly with curvature, so the generalized evidence and Jacobian-linearized predictive distribution are invariant under ReLU rescaling, making EB-Laplace well defined.

Acknowledgments

SP acknowledges support from Visvesvaraya PhD fellowship. PR acknowledges support from Google Research and Satya & Rao Remala Foundation. SP also thanks the Research-I Foundation of the Department of CSE at IIT Kanpur for the generous travel support.

References

J. Antoran, J. U. Allingham, D. Janz, E. Daxberger, E. Nalisnick, and J. M. Hernández-Lobato. Linearised laplace inference in networks with normalisation layers and the neural g-prior. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022a. URL https:

//openreview.net/forum?id=uUH8x-h9zdB.

J. Antoran, D. Janz, J. U. Allingham, E. Daxberger, R. R. Barbano, E. Nalisnick, and J. M. Hernandez-Lobato. Adapting the linearised Laplace model evidence for modern deep learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. 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 2022b. URL https://proceedings.mlr.press/v162/antoran22a.html.

F. Bergamin, P. Moreno-Muñoz, S. Hauberg, and G. 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.

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

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In D. Precup and Y. W. 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.

Y. Gelberg, T. F. A. van der Ouderaa, M. van der Wilk, and Y. Gal. Variational inference failures under model symmetries: Permutation invariant posteriors for bayesian neural networks. In S. Vadgama, E. Bekkers, A. Pouplin, S.-O. Kaba, R. Walters, H. Lawrence, T. Emerson, H. Kvinge, J. Tomczak, and S. 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.

A. Immer, M. Bauer, V. Fortuin, G. Rätsch, and M. E. Khan. Scalable marginal likelihood estimation for model selection in deep learning. In *ICML*, pages 4563–4573, 2021. URL http://proceedings.mlr.press/v139/immer21a.html.

A. Kristiadi, F. Dangel, and P. 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.

O. Laurent, E. Aldea, and G. 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.

D. 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.

B. Neyshabur, R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *NIPS*, pages 2422–2430, 2015a.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In P. Grünwald, E. Hazan, and S. 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 2015b. PMLR. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

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

S. Rossi, A. Singh, and T. 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.

H. Roy, M. Miani, C. H. Ek, P. Hennig, M. Pförtner, L. Tatzel, and S. 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.

L. 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.

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

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

7. 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., 2021) and extended to linearized settings (Antoran et al., 2022b). 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.

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., 2015b;a). 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 improving optimization, they 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., 2022a) 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.

8. A toy example

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

$$\mathbf{G}(a, c) = \sum_{n=1}^{N} \frac{1}{\sigma_y^2} \mathbf{J}_n^\top \mathbf{J}_n = \frac{S}{\sigma_y^2} \begin{pmatrix} c^2 & ac \\ ac & a^2 \end{pmatrix}, \qquad S = \sum_{n=1}^{N} x_n^2. \tag{9}$$

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

$$\mathcal{L}_{\mathrm{std}}(\lambda; \alpha) = \frac{1}{2} \log \frac{\lambda}{\kappa q(\alpha) + \lambda} - \frac{\lambda}{2} q(\alpha). \tag{10}$$

Empirical-Bayes optimum. Differentiating (10) with respect to λ and setting the derivative to zero gives

$$\frac{1}{\lambda} - \frac{1}{\kappa q(\alpha) + \lambda} = q(\alpha). \tag{11}$$

Combining fractions,

$$\frac{\kappa}{\lambda(\kappa q(\alpha) + \lambda)} = 1. \tag{12}$$

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

$$\lambda^2 + \kappa q(\alpha)\lambda - \kappa = 0. \tag{13}$$

The unique positive root is

$$\lambda^\star(\alpha) = \frac{-\kappa q(\alpha) + \sqrt{\kappa^2 q(\alpha)^2 + 4\kappa}}{2}. \tag{14}$$

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. □

9. Proof of Theorem 2.1

Let $\mathbf{v} = T_\alpha(\mathbf{w}) = \mathbf{D}_\alpha \mathbf{w}$. Since ReLU rescaling leaves the network function unchanged,

$$\log p(\mathcal{D} \mid \mathbf{v}) = \log p(\mathcal{D} \mid \mathbf{w}).$$

Using the congruence assumptions,

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

Hence

$$|\mathbf{C}_\mathbf{v} + \mathbf{\Lambda}(\mathbf{v})| = |\mathbf{D}_\alpha^{-1}|^2\, |\mathbf{C}_\mathbf{w} + \mathbf{\Lambda}(\mathbf{w})|.$$

Likewise,

$$|\mathbf{\Lambda}(\mathbf{v})| = |\mathbf{D}_\alpha^{-1}|^2\, |\mathbf{\Lambda}(\mathbf{w})|.$$

Therefore the determinant factors introduced by rescaling cancel between the two log-determinant terms.

For the quadratic term,

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

Thus every term in $\mathcal{L}_{\text{gen}}$ is unchanged, so

$$\mathcal{L}_{\text{gen}}(\mathbf{v}) = \mathcal{L}_{\text{gen}}(\mathbf{w}).$$

□

Theorem 9.1 (Linearized predictive invariance)**.** *Let $\mathbf{v} = T_\alpha \mathbf{w} = \mathbf{D}_\alpha \mathbf{w}$ be a ReLU rescaling of $\mathbf{w}$ that leaves the network function unchanged. Assume the Jacobian and curvature transform as*

$$\mathbf{J}_\mathbf{v}(\mathbf{x}) = \mathbf{J}_\mathbf{w}(\mathbf{x})\mathbf{D}_\alpha^{-1}, \qquad \mathbf{C}_\mathbf{v} = \mathbf{D}_\alpha^{-1}\mathbf{C}_\mathbf{w}\mathbf{D}_\alpha^{-1}.$$

Let the prior precision field be scale compatible:

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

Define the Laplace posterior covariances

$$\mathbf{\Sigma}_\mathbf{w} = (\mathbf{C}_\mathbf{w} + \mathbf{\Lambda}(\mathbf{w}))^{-1}, \qquad \mathbf{\Sigma}_\mathbf{v} = (\mathbf{C}_\mathbf{v} + \mathbf{\Lambda}(\mathbf{v}))^{-1}.$$

Then the linearized predictive covariance is invariant along the ReLU scale orbit:

$$\mathbf{J}_\mathbf{v}(\mathbf{x})\mathbf{\Sigma}_\mathbf{v}\mathbf{J}_\mathbf{v}(\mathbf{x})^\top = \mathbf{J}_\mathbf{w}(\mathbf{x})\mathbf{\Sigma}_\mathbf{w}\mathbf{J}_\mathbf{w}(\mathbf{x})^\top.$$

Consequently, any linearized Gaussian predictive distribution whose mean is $f(\mathbf{x};\mathbf{w})$ and covariance is $\mathbf{J}_\mathbf{w}(\mathbf{x})\mathbf{\Sigma}_\mathbf{w}\mathbf{J}_\mathbf{w}(\mathbf{x})^\top$ is invariant under ReLU rescaling.

Proof. Since $\mathbf{v} = \mathbf{D}_\alpha \mathbf{w}$ is a ReLU rescaling, the deterministic network function is unchanged:

$$f(\mathbf{x};\mathbf{v}) = f(\mathbf{x};\mathbf{w}).$$

By the assumed transformation rules,

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

Taking inverses gives

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

Using the Jacobian transformation $\mathbf{J}_\mathbf{v}(\mathbf{x}) = \mathbf{J}_\mathbf{w}(\mathbf{x})\mathbf{D}_\alpha^{-1}$, we obtain

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

Thus both the linearized predictive mean and covariance are unchanged along the scale orbit. □

10. Experimental Details

10.1. Stress Orbit 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 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.

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. 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 $\exp\big(\max_k \log \lambda_k^\star - \min_k \log \lambda_k^\star\big)$.

10.2. 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 $\gamma = 0.5$.