On the Interplay of Priors and Overparametrization in Bayesian Neural Network Posteriors

Julius Kobialka* LMU Munich, MCML

Emanuel Sommer* LMU Munich, MCML

Chris Kolb LMU Munich, MCML

Juntae Kwon LMU Munich

Daniel Dold HTWG Konstanz

David Rügamer LMU Munich, MCML

Abstract

Bayesian neural network (BNN) posteriors are often considered impractical for inference, as symmetries fragment them, non-identifiabilities inflate dimensionality, and weight-space priors are seen as meaningless. In this work, we study how overparametrization and priors together reshape BNN posteriors and derive implications allowing us to better understand their interplay. We show that redundancy introduces three key phenomena that fundamentally reshape the posterior geometry: balancedness, weight reallocation on equal-probability manifolds, and prior conformity. We validate our findings through extensive experiments with posterior sampling budgets that far exceed those of earlier works, and demonstrate how overparametrization induces structured, prior-aligned weight posterior distributions.

1 INTRODUCTION

Bayesian neural networks (BNNs) have been a central tool for facilitating uncertainty quantification in deep learning by placing distributions over weights and approximating their posterior. As the posterior of such networks is usually intractable, one has to rely on simplifying variational assumptions (e.g., Blundell et al., 2015) or other approximation techniques (e.g., Daxberger et al., 2021) to estimate these distributions. However, these approximations are rather crude and often insufficient, raising doubts about whether BNNs are truly appropriate for uncertainty quantification.

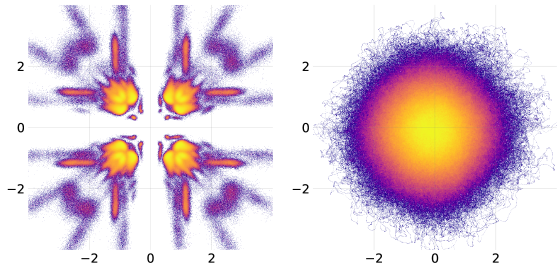


Figure 1: Sampled marginal bivariate posterior densities of two weights (axes, lighter means higher density) for an underparametrized (left) and an overparametrized model (right) on the same task using 10 million posterior samples.

Sampling-based inference for BNNs, an alternative approach that estimates the posterior without restrictive distributional assumptions, has gained traction thanks to recent advances. These include methods to scale to large datasets (Springenberg et al., 2016) and large models (Sommer et al., 2025), as well as approaches improving acceptance rates (e.g., Papamarkou, 2023), prior understanding (e.g., Vladimirova et al., 2019), initialization (Sommer et al., 2024), and scalable software implementations (Duffield et al., 2025). Several of these works report superior performance compared to variational methods and other approximations.

Nevertheless, criticisms in the community remain (see, e.g., Papamarkou et al., 2024), questioning the adequacy of (sampling-based) inference for BNNs: *Why does the posterior landscape appear less and less fragmented (cf. Figure 1) in overparametrized models? How does the non-identifiability affect posterior inference? Are there meaningful priors in weight space? What can be said about the individual weight distributions?*

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. *Authors contributed equally.

In this paper, we make important advances toward answering these questions. In particular, we identify *overparametrization* as the common property that shapes BNN posteriors and helps to explain the recent success of sampling.

Our Contributions In this work, we theoretically and empirically investigate how overparametrization and priors shape the posterior and how this affects sampling-based inference. Specifically, we

- translate properties from the optimization literature into statements about prior choices and posterior shape;
- show how overparametrization affects the posterior by creating equal-probability manifolds;
- show that overparametrization induces what we call *prior conformity*, yielding an easy-to-encode a priori understanding of weight distributions;
- conduct extensive experimental evaluations to analyze the effect of changing posterior shapes on sampling-based inference, with sampling budgets largely exceeding existing work.

2 RELATED LITERATURE

Sampling-based inference for BNNs typically relies on Markov Chain Monte Carlo (MCMC) methods. It is often considered a gold standard in the Bayesian deep learning (BDL) community (see, e.g., Farquhar et al., 2020), theoretically allowing to sample from the true posterior (in the limit). While devising MCMC methods that can handle the high dimensionality and intricate structure of parameter spaces in modern neural networks has remained a challenge (Papamarkou et al., 2022; Wenzel et al., 2020), recent advances have made progress toward more efficient mixing of Markov chains (Sen et al., 2024), scaling samplers for larger datasets (Paulin et al., 2025; Chen et al., 2014; Zhang et al., 2020), and parameter spaces (Sommer et al., 2024, 2025, 2026).

Non-Identifiabilities and Symmetries A particular challenge for sampling-based BNN inference lies in the fact that contemporary neural networks contain a vast number of neurons and connections that induce a high degree of redundancy in model parameters (Papamarkou et al., 2022). As such, the function learned by a neural network is invariant to a range of parameter transformations, which subsequently induces loss invariance (Ziyin, 2024). In the literature, these invariant transformations are also viewed as symmetries, with the most common examples introduced in Section 3. In the BNN literature, symmetries have been repeatedly discussed with the general premise of being harmful in posterior inference as they introduce redundancy and, thus, cause inefficiencies in posterior approximation (Papamarkou et al., 2022; Wiese et al., 2023; Laurent et al., 2024).

Proposals to deal with symmetries in neural networks include bias sorting (Pourzanjani et al., 2017), assigning individual parameter offsets (Ziyin et al., 2025b), or skip connections (Kurle et al., 2021) to remove permutation invariances; using invariant networks (e.g., Maron et al., 2019; Navon et al., 2023), removing scaling symmetries via regularization (Laurent et al., 2024), or computing a model average over the elements of a symmetry orbit (Gelberg et al., 2024). While symmetries are also known to slow down sampling (Nalisnick, 2018; Papamarkou et al., 2022; Wiese et al., 2023), only a few papers have studied symmetries in sampling-based inference systematically.

Bayesian Neural Networks Likewise, there are only a few works that elucidate the connection between posterior structure, symmetries, and overparametrization in BNNs. Earlier large-scale empirical studies of sampling-based posterior approximations were able to uncover posterior mode connectivity that can be exploited by samplers (Izmailov et al., 2021), in line with the mode connectivity phenomenon in deep NNs (Garipov et al., 2018; Draxler et al., 2018) and BDL methods exploiting this phenomenon, such as subspace inference (Izmailov et al., 2020; Dold et al., 2024, 2025). In contrast to the optimization community, where the influence of overparametrization is hypothesized to make the loss landscape more benign, work such as Trippe and Turner (2017) expressed concern that a potential pathology for variational approximations of overparametrized BNN posteriors can paradoxically lead to worse performance and induce conditional independence between parameters and data.

We refer the reader to Appendix A for further discussion of related literature.

3 THE INTERPLAY BETWEEN OVERPARAMETRIZATION AND SYMMETRIES

We start with an instructive example to explain the interplay between overparametrization, priors, and symmetries in a simplified setup and define our notation.

Setup and Notation First, let $f(\mathbf{x}; \mathbf{w})$ denote a neural network with input $\mathbf{x} \in \mathcal{X}$ and parameters $\mathbf{w} \in \mathcal{W} \equiv \mathbb{R}^d$, where $\mathbf{w}$ stacks all weights across all layers. Throughout the paper, we will work with various

architectures, but focus on homogeneous networks to better convey our findings and demonstrate the properties induced by network components. We will also use $f(\mathbf{w})$ if discussing a property that holds for all $x \in \mathcal{X}$. When considering fully-connected architectures, we will use "M_1-M_2-...-M_L" as a short form for a network with M_l neurons in layers $l \in [L] := \{1, \dots, L\}$ Given training data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, we define the empirical (unregularized) loss $\mathcal{L}(\mathbf{w}) := -\sum_{i=1}^n \ell(f(\mathbf{x}_i; \mathbf{w}), \mathbf{y}_i)$, where ℓ is a pointwise loss function. In this paper, we will study BNNs that have a corresponding regularized risk learning problem

$$\arg\min_{\mathbf{w}} \mathcal{L}_\lambda := \arg\min_{\mathbf{w}} \Big(\mathcal{L}(\mathbf{w}) + \lambda \mathcal{R}(\mathbf{w}) \Big),$$

with regularizer $\mathcal{R}(\mathbf{w})$ and tuning parameter $\lambda > 0$.

From a Bayesian perspective, we interpret $\mathcal{L}(\mathbf{w})$ as the negative log-likelihood of the data given the weights $\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^n \log p(\mathbf{y}_i \mid f(\mathbf{x}_i; \mathbf{w}))$ with f typically encoding the mean of the distribution, and assume a prior $p(\mathbf{w}) \propto \exp(-\lambda \mathcal{R}(\mathbf{w}))$. Then the posterior $p(\mathbf{w}|\mathcal{D}) \propto \exp(-\mathcal{L}_\lambda)$. A standard choice we will not only use throughout this paper but also advocate for is the isotropic Gaussian prior $p(\mathbf{w}) \propto \exp\big(-\frac{1}{2\tau^2}\|\mathbf{w}\|_2^2\big)$ with $\lambda = 1/(2\tau^2)$. If prior variances differ across layers $l \in [L]$ of an L-layer network, we will make this explicit by writing $\mathcal{L}_{\boldsymbol{\tau}}$ with $\boldsymbol{\tau} = \{\tau_l\}_{l \in [L]}$. We will use the Bayesian and regularized risk point of view often interchangeably when talking about properties of the weight space.

Overparametrization Given the recent success of warmstarted sampling-based inference (see, e.g., Sommer et al., 2025), we will investigate BNNs that have more hidden neurons than necessary to represent the underlying data-generating process. This is in line with other works studying overparametrization, such as Nguyen (2019) or Kim et al. (2025). An alternative layerwise property to define overparametrization is to assume that the activation matrix of the l-th layer has rank strictly smaller than its number of neurons. Formal definitions are given in Appendix B.

3.1 Instructive Overparametrization Example

In the following instructive example, we will provide some intuition on how overparametrization interacts with symmetries and introduce concepts required for our more general exposition in subsequent sections. As an extension of *single-neuron linear networks* (Kunin et al., 2024; Azulay et al., 2021), we consider a linear neural network for scalar inputs given by

$$f(x; \mathbf{w}) = \sum_{m=1}^M w_{1,m} w_{2,m} x, \quad x \in \mathbb{R}, \tag{1}$$

which can be interpreted as a shallow network with M hidden units (a *1-M-1* network). It is intuitively clear why f is *overparametrized*: the space of functions that can be represented by f is the hypothesis space $\{h : h(x) = x\beta, \beta \in \mathbb{R}, x \in \mathcal{X}\}$, i.e., a univariate linear model. Yet, f uses $2M$ instead of just one parameter. As a result, f admits various symmetries:

1. a (positive) rescaling symmetry as $\forall c > 0$ and $\forall m \in [M]$, we have that $f(\mathbf{w}) = f(\tilde{\mathbf{w}})$ for $\tilde{\mathbf{w}} = (w_{1,1}, w_{2,1}, \dots, c \cdot w_{1,m}, c^{-1} \cdot w_{2,m}, \dots, w_{1,M}, w_{2,M})$;
2. as a special case of the above, a sign-flip symmetry as $f(\mathbf{w}) = f(\tilde{\mathbf{w}})$ for $\tilde{\mathbf{w}}$ containing at least one pair of flipped signs, i.e., $\tilde{\mathbf{w}} = (\dots, -w_{1,m}, -w_{2,m}, \dots)$;
3. a permutation symmetry, since $f(\mathbf{P}\mathbf{w}) = f(\mathbf{w})$ for any block-permutation matrix $\mathbf{P}$ that permutes hidden units as pairs;
4. a rotation symmetry, since f only depends on inner products. In particular, for any orthogonal matrix $\mathbf{R}$, $f(\mathbf{w}) = \mathbf{w}_1^\top \mathbf{w}_2 x = (\mathbf{R}\mathbf{w}_1)^\top (\mathbf{R}\mathbf{w}_2) x = f(\tilde{\mathbf{w}})$ with $\mathbf{w}_i = (w_{i,1}, \dots, w_{i,M})$ and $\tilde{\mathbf{w}}$ being the vector $\mathbf{w}$ containing the rotated components.

The Bayesian Perspective From a Bayesian point of view, the existence of symmetries in f implies invariance to certain transformations. If the inferential target is merely f, symmetry-respecting priors are the principled baseline encoding knowledge about f. More specifically, the matching prior for rescaling symmetries must encode a Cartesian product of scale-invariant improper prior distributions over the product of weights for every weight combination that admits rescaling symmetries, e.g., $p(w_j, w_{j'}) \propto (w_j w_{j'})^{-1}$. This is a rather atypical choice in BNNs, and as a consequence, most posterior densities will not be scale invariant. In contrast, every symmetric prior admits sign-flip symmetries. If f is sign-flip invariant, we have $p(w_j|\mathcal{D}) = p(-w_j|\mathcal{D})$, and thus $\mathbb{E}(w_j|\mathcal{D}) = 0$ for every marginal posterior of w_j. Permutation symmetries imply that the (co-)variance structure of the permuted weights is identical. In particular, heterogeneous weight priors imply a permutation variant posterior density. Finally, rotations, which can be viewed as a continuous generalization of permutations, are challenging to encode in general. For specific examples as shown in the next section, however, an independent Gaussian distribution $\mathbf{w} \sim \mathcal{N}(0, \tau^2 \mathbf{I})$ preserves such symmetries.

3.2 The Influence of Regularization

The previously described conformity between symmetries and weights can also be regarded from a regularization perspective. While the *1-M-1* network might admit previously discussed symmetries, this is not necessarily the case for the loss function $\mathcal{L}_\lambda$, in particular for $\lambda > 0$ and rescaling symmetries. In contrast to

f, the regularizer $\mathcal{R}$ in $\mathcal{L}_\lambda$ is typically not invariant under rescaling. For L_2-regularization, we obtain for the example above that for any $\mathbf{w} \in \mathbb{R}^{2M}$ there exists a rescaled version $\tilde{\mathbf{w}} \in \mathbb{R}^{2M}$ for which $\mathcal{L}_\lambda(\tilde{\mathbf{w}}) \leq \mathcal{L}_\lambda(\mathbf{w})$. That is, the penalty breaks the rescaling invariance and prefers *balancedness*. For $M = 1$, this means that it must hold $w_{1,1}^2 = w_{2,1}^2 = |\beta|$ with $\beta = w_{1,1}w_{2,1}$, as the regularizer would otherwise be larger while f stays constant. This can be seen by applying the AM-GM inequality: $\mathcal{R}(\mathbf{w}) = \frac{1}{2}(w_{1,1}^2 + w_{2,1}^2) \geq \sqrt{w_{1,1}^2 w_{2,1}^2} = |\beta|$, where equality holds iff $|w_{1,1}| = |w_{2,1}| = \sqrt{|\beta|}$. In general, for any homogeneous unit, we obtain this balancedness at optimality (Parhi and Nowak, 2023).

For $M > 1$ and $\beta = \mathbf{w}_1^\top \mathbf{w}_2$, however, exact balancedness is only enforced between product factors, i.e., $|w_{1,m}| = |w_{2,m}| \forall m \in [M]$ and not across neurons. This can be understood as a result of the function's rotation symmetry, yielding the following global minimizers of $\mathcal{R}(\mathbf{w})$ for a fixed β:

$$\mathbf{w}_1 = \pm\sqrt{|\beta|} \cdot \boldsymbol{v},\ \mathbf{w}_2 = \pm \text{sign}(\beta)\sqrt{|\beta|} \cdot \boldsymbol{v}$$

with $\boldsymbol{v} \in \mathbb{S}^{M-1} := \{\boldsymbol{v} \in \mathbb{R}^M : ||\boldsymbol{v}||_2 = 1\}$, using $\text{sign}(\beta)$ for $\mathbf{w}_2$ without loss of generality.

The Bayesian Perspective The zero-centered isotropic Gaussian prior as the equivalent of an L_2 regularization does not preserve scaling symmetries. However, it preserves sign-flip symmetries (and rotation symmetry in this example). As we will see in the following sections, this trait and the balancedness among weight norms imply important properties.

3.3 The Influence of Overparametrization

In Figure 2, we plot the resulting posterior from a bivariate marginal point of view for two components $w_{1,m}, w_{2,m}$ for different M. This reveals an interesting effect: While the scaling and permutation symmetries are clearly visible for $M = 1$, producing a butterfly-shaped posterior when the penalty (prior) is not too strong, increasing M causes the rotation symmetry to dominate, pushes the posterior mode closer to zero, and decreases the anisotropy of the distribution. More specifically, under the constraint $\sum_m u_m := \sum_{m=1}^M w_{1,m}w_{2,m} = \beta$, i.e., an optimal likelihood, and using the prior $w_{j,m} \sim \mathcal{N}(0, \tau^2)$, we have $\mathbb{E}(w_{j,m} | \sum_m u_m = \beta) = 0$, $\text{Var}(w_{j,m} | \sum_m u_m = \beta) = \tau^2 + O(M^{-2})$, and

$$\text{Cov}(w_{j,m}, w_{j',m} | \textstyle\sum_m w_{1,m}w_{2,m} = \beta) = \beta/M.$$

We will formalize this result in the next section.

The Bayesian Perspective From a Bayesian perspective, increasing the overparametrization in this linear network case implies that when in high likelihood regions, posterior variances approach the prior variances, and the covariance between two weights vanishes, as can be nicely seen in Figure 2.

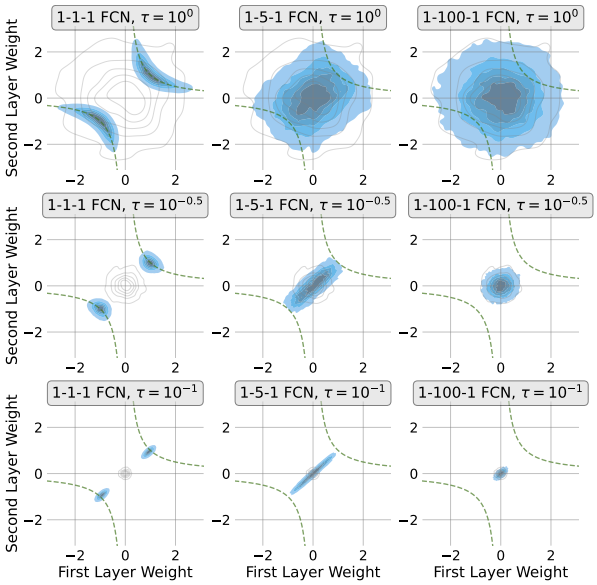


Figure 2: Posterior marginals of *1-M-1* fully-connected network (FCN) with zero-mean Gaussian priors and different variances τ^2 and data $y = x + \epsilon$, $\epsilon \sim N(0, 0.1)$.

4 GENERAL RELU NETWORKS

Moving from the previous instructive example to more general ReLU networks, we investigate the existence of these properties in more realistic architectures. As before, we analyze BNNs that obtain constant high likelihood values, which allows us to reason about their local posterior geometry in regions that ultimately matter the most, which is again in line with recently proposed warmstarting approaches. In contrast to the previous section, where this was formalized by assuming an underlying linear model, we here generalize the idea by contrasting an overparametrized model f with excessive hidden neurons with an interpolating network f^* that achieves minimum norm (i.e., the smallest model that can still fit the data and adheres to the prior). Due to non-identifiabilities, many models f can achieve the same loss as f^*, but not all are minimum norm solutions. We therefore define the following:

Definition 1 (Manifold of minimum norm interpolators)**.** Given an L-layer network f with weights $\mathbf{w}$ and interpolating L-layer network f^* with weights $\mathbf{w}^*$ attaining the minimum norm among all interpolants of L-layer networks, the *minimum norm manifold* is defined by

$$\mathcal{M} = \{\mathbf{w} \in \mathbb{R}^d : \mathcal{L}_\lambda(\mathbf{w}) = \mathcal{L}_\lambda(\mathbf{w}^*)\}, \tag{2}$$

where we assume $\mathcal{R}$ to be a (weighted) L_2-regularization (i.e., a diagonal Gaussian prior).

Overparametrization then means that each neuron $\varpi \in [M^*]$ of f^* can be replicated by a group $G_\varpi \subseteq [M]$ of $k_\varpi = |G_\varpi|$ neurons in f, as described by a surjective assignment map $\varsigma : [M] \to [M^*]$. The key quantity of interest becomes the vector of reallocation coefficients $\boldsymbol{\rho}^{(\varpi)}$, which describes how the norm of neuron ϖ in f^* is distributed among its copies in f, which is the higher-dimensional analogue of the equal spreading of weight norm observed in Section 3.3. That is, each neuron $m \in G_\varpi$ in f takes the form $\boldsymbol{\omega}_m = \sqrt{\rho_m}\,\boldsymbol{\omega}^*_\varpi$, where $\rho_m \geq 0$ denotes the fraction of the squared norm of neuron ϖ allocated to copy m, with $\sum_{m\in G_\varpi} \rho_m = 1$. Since ReLU is positive homogeneous, this rescaling preserves the function (see Lemma 1).

As we cannot condition the posterior on $\mathcal{M}$ in the usual way, we define an induced (conditional) law via shrinking tubes around $\mathcal{M}$ (an ε-neighborhood $\mathcal{M}^\varepsilon$). This yields a way to study the ambient space around the minimum-norm manifold $\mathcal{M}$.

Definition 2. For any $\varepsilon > 0$, define the ε-tube around a manifold $\mathcal{M} \subseteq \mathcal{W}$ as

$$\mathcal{M}^\varepsilon := \big\{\mathbf{w} \in \mathbb{R}^d : \operatorname{dist}(\mathbf{w}, \mathcal{M}) \leq \varepsilon\big\},$$

where the distance is measured in the Euclidean norm.

For each $\varepsilon > 0$, the tube has positive Lebesgue measure, admitting:

$$\mathbb{P}^\varepsilon_n(A) := \mathbb{P}_n\big(\mathbf{w} \in A \mid \mathbf{w} \in \mathcal{M}^\varepsilon\big), \qquad A \subseteq \mathcal{M},$$

where $\mathbb{P}_n(d\mathbf{w}) = p(\mathbf{w} \mid \mathcal{D}_n)\, d\mathbf{w}$ is the ordinary posterior given n observations.

4.1 Two-Layer ReLU Networks

The general shallow *p-M-1* ReLU network $f(\mathbf{x}) = \sum_{m=1}^M w_{2,m}^\top \phi(\mathbf{w}_{1,m}^\top \mathbf{x})$ with ReLU activation functions ϕ and multivariate input $\mathbf{x} \in \mathbb{R}^p$ has been intensively studied in the literature and admits useful properties such as convexifiability (see, e.g., Mishkin et al., 2022) and connectedness of solution sets in the overparametrized regime (Kim et al., 2025). To generalize previously made observations of balanced weight behavior as M increases, we obtain the following:

Theorem 1 (informal)**.** *Under Assumption 1 (an overparametrized p-M-1 model f and a minimal norm cost interpolant f^* with $M^* < M$ neurons), the coefficients $\boldsymbol{\rho}^{(\varpi)} \in \Delta^{k_\varpi - 1}$ of any norm-preserving reallocation $\varsigma : [M] \to [M^*]$ of the weight strength of neuron $\varpi \in [M^*]$ of f^* to a group of weights $m \in G_\varpi$ with $k_\varpi = |G_\varpi| \geq 1$ neurons from a model f on $\mathcal{M}$ follows a symmetric Dirichlet distribution.*

In order to move towards conclusions about the posterior of $\mathbf{w}$, we use Definition 2 to regard the ambient space of $\mathcal{M}$. For this, we first define $\boldsymbol{\omega}_m := (\mathbf{w}_{1,m}, w_{2,m}) \in \mathbb{R}^{p+1}$ to be the block of weights (in- and outgoing weights) corresponding to the m-th neuron in the hidden layer and $\{\boldsymbol{\omega}_m\}_{m\in G_\varpi}$ the block of neurons in G_ϖ that overparametrize neuron $\boldsymbol{\omega}_\varpi$.

Corollary 1. *Consider a k_ϖ-fold split of a neuron $\boldsymbol{\omega}^*_\varpi$, $\varpi \in [M^*]$ into k_ϖ neurons $\boldsymbol{\omega}_m = \sqrt{\rho_m}\,\boldsymbol{\omega}^*_\varpi$ with $\boldsymbol{\rho}^{(\varpi)} := (\rho_m)_{m\in G_\varpi} \in \Delta^{k_\varpi - 1}$. Then the induced distribution of $\boldsymbol{\rho}^{(\varpi)}$ on $\mathcal{M}^\epsilon$ is symmetric Dirichlet,*

$$\boldsymbol{\rho}^{(\varpi)} \sim \operatorname{Dirichlet}\Big(\tfrac{p+1}{2}, \ldots, \tfrac{p+1}{2}\Big),$$

for each $\varpi = 1, \ldots, M^*$ independently under tube-conditioning as in Definition 2.

Although the ε-tube $\mathcal{M}^\epsilon \subset \mathbb{R}^d$ resides within the ambient parameter space $\mathcal{W}$, the Dirichlet result strictly isolates the marginal distribution of the reallocation coefficients $\boldsymbol{\rho}^{(\varpi)}$ near $\mathcal{M}$.

If we assume the setup from Theorem 1 and consider the minimum-norm manifold $\mathcal{M}_\varsigma \subseteq \mathcal{M}$ for a specific assignment or reallocation map ς, we can define the induced posterior with ε-tube:

Corollary 2. *On $\mathcal{M}_\varsigma$, let $\mathbb{P}_n := \lim_{\varepsilon\downarrow 0} \mathbb{P}^\varepsilon_n$, assuming this weak limit exists. It holds (1) $\mathbb{E}_{\mathbb{P}_n}(w) = O(M^{-1/2})$, (2) $\operatorname{Cov}_{\mathbb{P}_n}(w, w') = O(M^{-2})$.*

These results follow from Theorem 1 and imply that the posterior around this manifold will move closer and closer to zero and induce increasingly less dependence (covariance) between weights the more the hidden layer (M) grows. This statement holds uniformly across assignments ς with comparable group sizes, together with the plausible assumption that the posterior places substantial mass near $\mathcal{M}$. Concurrently, because the likelihood remains nearly constant in the immediate neighborhood of the manifold, the L_2 regularization (Gaussian prior) induces approximately quadratic fluctuations strictly in the directions normal to the manifold. This may therefore explain why highly redundant, one-dimensional weight marginals resemble zero-centered Gaussian distributions, as can be seen in Figure 3. We emphasize, however, that this analysis is local to the neighborhood of $\mathcal{M}$ and does not by itself determine the shape of the full marginal posterior.

While the preceding analysis is specific to two-layer networks, the phenomena it predicts, namely prior-like marginals, vanishing covariance, and modes near zero, are clearly visible in deeper architectures as well (cf. Figures 3 and 4). The following section explains why these patterns persist beyond the shallow setting.

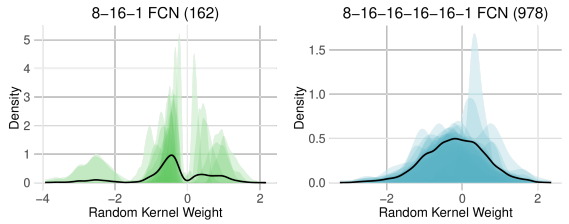


Figure 3: Univariate marginal distributions of selected first hidden layer weights for sampled BNNs and their change with varying network size (number of parameters in brackets) on the `concrete` regression task. The 20 chain-wise distributions are overlaid with transparency, and the aggregated joint marginal is shown as a solid black line. Despite similar downstream performance, the marginals differ markedly between the two network parameterizations.

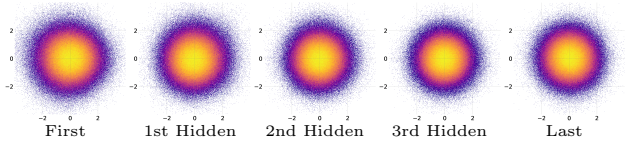


Figure 4: Bivariate marginal posterior densities of a 4-hidden layer BNN fitted on the `ionosphere` dataset. The grid visualizes the empirical densities of 8 million posterior samples obtained from 8,000 independent chains. Columns display representative densities of different layer weights.

4.2 Properties of Deeper Networks

As minimum-norm manifolds become harder to characterize explicitly, we will move from the tube-based analysis to a more general treatment of BNNs based directly on the posterior π.

Balancedness The following exposition assumes

$$f(\mathbf{x}) = \mathbf{W}_L \phi(\mathbf{W}_{L-1} \phi(\cdots(\mathbf{W}_1 \mathbf{x}))) \tag{3}$$

with ReLU activation functions ϕ and arbitrary depth L. Let $\mathbf{w} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$ collect the layer-wise weights and $d_l := |\mathrm{vec}(\mathbf{W}_l)|$, the posterior $\pi := p(\mathbf{w}|\mathcal{D}) \propto \exp(-\mathcal{L}_{\boldsymbol{\tau}}(\mathbf{w}))$, and assume zero-mean Gaussian priors with variances τ_l^2 for layer $l \in [L]$. One can view depth in such homogeneous networks as a form of *multiplicative* overparametrization: positive homogeneity of ReLU induces a positive rescaling symmetry, introducing one continuous degree of freedom per hidden neuron per layer boundary. Whereas Section 4.1 characterized how the Gaussian prior organizes *additive* redundancy, the following balancedness result relates to the analogous mechanism for multiplicative redundancy, where the isotropic Gaussian prior breaks the rescaling invariance by penalizing unequal norms across adjacent factors.

For this, we utilize a structural property of homogeneous networks derived in Du et al. (2018). In their analysis of gradient flow, the authors establish the following pointwise algebraic identity for all $\mathbf{w}$ where the gradient exists:

$$\langle \mathbf{W}_l, \nabla_{\mathbf{W}_l} \mathcal{L}(\mathbf{w}) \rangle_F = \langle \mathbf{W}_{l+1}, \nabla_{\mathbf{W}_{l+1}} \mathcal{L}(\mathbf{w}) \rangle_F.$$

Since this equality holds for all $\mathbf{w}$ due to the 1-homogeneity of the ReLU activation and the chain rule, it necessarily holds a.s. in expectation over the stationary distribution π assuming it exists and fulfills standard regularity conditions:

$$\mathbb{E}_\pi[\langle \mathbf{W}_l, \nabla_{\mathbf{W}_l} \mathcal{L}(\mathbf{w}) \rangle_F] = \mathbb{E}_\pi[\langle \mathbf{W}_{l+1}, \nabla_{\mathbf{W}_{l+1}} \mathcal{L}(\mathbf{w}) \rangle_F].$$

Remark 1. If the continuous-time weight vector $\mathbf{w}(t)$, e.g., follows the overdamped Langevin diffusion and is initialized at stationarity, then $\mathbf{w}(t) \sim \pi$ for all t and the above identities hold for expectations of observables evaluated at any fixed time. If, in addition, the diffusion is ergodic, then time averages along a single trajectory converge to the corresponding π-expectations.

Using the equality of expectations, we have the following balancedness result, which is also demonstrated visually for a 4-hidden layer BNN in Figure 4.

Theorem 2. *For f as defined in (3) and every adjacent pair $(l, l{+}1)$ it holds at stationarity that*

$$\tau_l^{-2}\, \mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] \;-\; \tau_{l+1}^{-2}\, \mathbb{E}_\pi\big[\|\mathbf{W}_{l+1}\|_F^2\big] \;=\; d_l - d_{l+1}.$$

As a direct consequence, we obtain the following.

Corollary 3. *If $\tau_l^2 \equiv \tau^2\, \forall l \in [L]$, then $\mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] - \mathbb{E}_\pi\big[\|\mathbf{W}_{l+1}\|_F^2\big] = \tau^2 (d_l - d_{l+1}) \;\; \forall l \in [L-1]$. In particular, if also $d_l = d_{l+1}$, then $\mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] = \mathbb{E}_\pi\big[\|\mathbf{W}_{l+1}\|_F^2\big]$. Furthermore, if $\mathbb{E}_\pi[\mathbf{W}_l] = \mathbb{E}_\pi[\mathbf{W}_{l+1}] = \mathbf{0}$, we have $\mathrm{tr}\big(\mathrm{Cov}_\pi(\mathrm{vec}(\mathbf{W}_l))\big) = \mathrm{tr}\big(\mathrm{Cov}_\pi(\mathrm{vec}(\mathbf{W}_{l+1}))\big)$.*

Even when not setting variances to the same value across layers, we obtain an expected balancedness across the network (cf. Corollary 5 in Appendix C).

Overparametrization and Prior Conformity A second phenomenon that arose in the previous exposition in single-hidden layer networks was what one could call *prior conformity*, i.e., the tendency for (marginal) weight distributions to resemble the (marginal) prior. We can try to explain this phenomenon by data-independent directions in the posterior using a local approximation, similar to recent results such as Roy et al. (2024). For this, assume a locally optimal reference parameter $\mathbf{w}^{\mathrm{ref}} \in \mathcal{W}$ and Laplace approximation based on a generalized Gauss-Newton precision matrix

$$p(\mathbf{w} \mid \mathcal{D}) \;\approx\; \mathcal{N}\big(\mathbf{w}^{\mathrm{ref}}, (\mathbf{H}^\star)^{-1}\big), \quad \mathbf{H}^\star := \mathbf{H}^\star_{\mathrm{data}} + 2\lambda \mathbf{I}.$$

Further, let $\mathbf{J} \in \mathbb{R}^{n \times d}$ be the Jacobian of the network outputs w.r.t. $\mathbf{w}$ at $\mathbf{w}^{\mathrm{ref}}$ (stacked over all n training

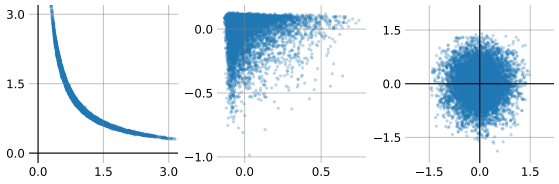


Figure 5: Samples from a 1-2-1 ReLU network on a linear dataset (left) and its separation into image (center) and null space of $\mathbf{J}$ (right).

samples), and $\mathbf{H}^\star \approx \mathbf{J}^\top \boldsymbol{\Upsilon} \mathbf{J} + 2\lambda \mathbf{I}$ for a positive semidefinite weight matrix $\boldsymbol{\Upsilon}$ (e.g., $\boldsymbol{\Upsilon} = \sigma^{-2}\mathbf{I}$ for L_2-loss). If layer l is overparametrized, then $\ker(\mathbf{J}) \neq \{\mathbf{0}\}$ and

$$(\mathbf{H}^\star)^{-1}\big|_{\ker(\mathbf{J})} = (2\lambda)^{-1}\mathbf{I} = \tau^2\mathbf{I}.$$

Hence, the posterior covariance equals the prior covariance along likelihood-flat directions. This can be interpreted as a *prior conformity* of the posterior as the Gaussian prior "fills in" singularities in redundant subspaces. Connected to this result are approximate rotational invariances.

Let $\mathcal{W} = \ker(\mathbf{J}) \oplus \text{img}(\mathbf{J})$ be the orthogonal decomposition at $\mathbf{w}^{\text{ref}}$ and $\mathbf{P}_{\ker(\mathbf{J})}$ the projection matrix projecting into $\ker(\mathbf{J})$. Further, let $\mathbf{Z}$ contain the orthonormal columns spanning $\ker(\mathbf{J})$, i.e., $\mathbf{P}_{\ker(\mathbf{J})} = \mathbf{Z}\mathbf{Z}^\top$. Then, it trivially holds

$$\mathbf{Z}^\top(\mathbf{w} - \mathbf{w}^{\text{ref}}) \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}),$$

independent of the data. To confirm these statements, we can compute $\mathbf{Z}$ and project all posterior samples onto $\ker(\mathbf{J})$ to check whether the distribution of the projected and centered weights aligns with the prior (as demonstrated in Figure 5).

When specialized to the shallow ReLU setting, $\ker(\mathbf{J})$ at a minimum-norm solution contains the tangent directions of $\mathcal{M}$ along which neurons can be rebalanced without changing the function, but also includes additional directions that preserve the function only to first order without necessarily preserving the norm. The Laplace argument thus captures a broader class of redundant directions than the Dirichlet results, though with stronger assumptions.

4.3 The Role of the Bias

In the previous sections, our exposition mainly focused on layers without a bias. When incorporating a bias into layers, e.g., $f(\mathbf{x}) = \mathbf{w}_2^\top \phi(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$, regularization effects induced by the prior still enforce $\|(\mathbf{W}_1)_{[i,:]}\|_2 = |w_{2i}|$ at optimality. This is, however, only the case if $\mathbf{b}_1$ is not regularized (Parhi and Nowak, 2023), i.e., when using an improper flat prior $p(\mathbf{b}_1) \propto 1$. In general, homogeneity is broken at the entry and exit points of the network: the first layer receives a shift in its input space, and the last layer produces an affine rather than homogeneous mapping, e.g., encoding $\mathbb{E}(y|\mathbf{x} = \mathbf{0})$ in a regression setting. As a consequence, the induced posterior distributions, typically in the first and last layers, deviate from the balanced patterns described above and display different marginal shapes (cf. Figure 6). With suitable activations in the first layer, however, bias expectations will be zero in subsequent layers, and the ReLU nonlinearity can continue to be positively homogeneous, implying the same rescaling invariances as before.

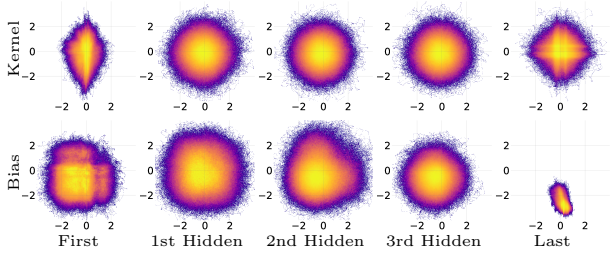


Figure 6: Bivariate marginal posterior densities of a 4-hidden layer BNN fitted on the `airfoil` dataset. The grid visualizes the empirical densities of 10M posterior samples obtained from 10k independent chains. The rows and columns display representative densities of randomly chosen layer weights.

5 OVERPARAMETRIZATION INFLUENCE ON SAMPLING-BASED INFERENCE

After studying the posterior from a more theoretical perspective, we now provide empirical evidence of how overparametrization affects sampling-based posterior approximations. While our theory focused mainly on fully-connected networks, we also provide results from experiments using other network architectures. In particular, Appendix E provides a variety of benchmarks to validate empirical findings in recent papers, confirming the good performance of sampling for BNN inference when sufficiently overparametrized, and analyzing their corresponding posterior shapes.

5.1 Under- vs. Overparametrized Models

We begin by analyzing the functional diversity recoverable by the sampler. As shown in Figure 7, for smaller architectures the sampler remains confined to distinct modes in parameter space, while larger parameter counts make the marginals more closely aligned with the prior. In Section 4.2, we plot the marginal distributions of two weights per chain as well as aggregated over chains in Figure 3. The plot implies that marginalizing over differently initialized chains in over-

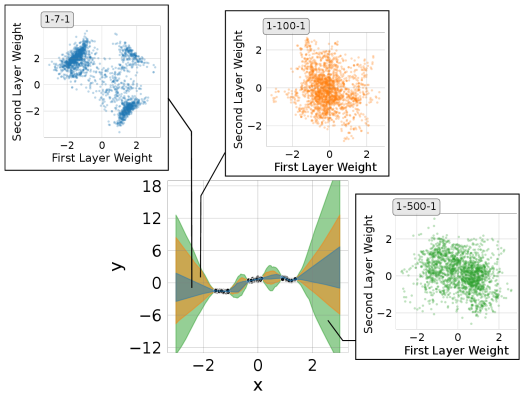


Figure 7: Empirical marginal posterior view for ReLU networks with increasing widths. As the marginal coverage of the posterior increases, the functional variance recovered by the sampler increases as well.

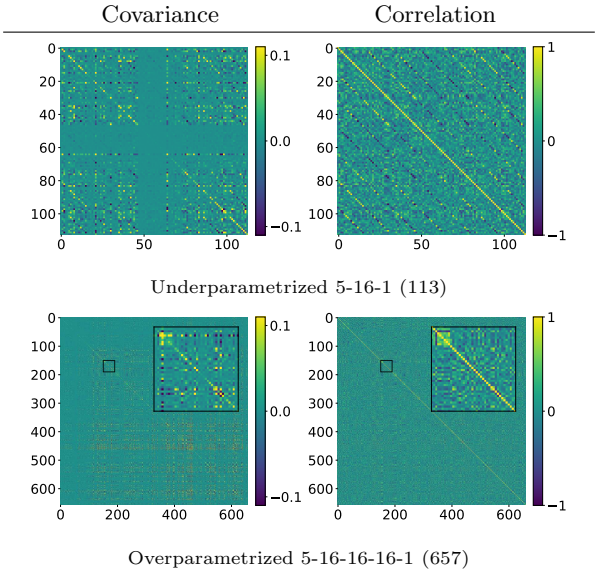


Figure 8: Empirical covariance and correlation of weights $\mathbf{w}$ across posterior samples on the `airfoil` dataset. The sampler is able to recover an intricate correlation structure in the parameter space.

parametrized models reveals the suggested marginal posterior pattern, which is not the case for the underparametrized model. A similar pattern can be found in Figure 1. This connectedness, alas, comes at the cost of having to sample in higher dimensions. It is therefore not immediately clear whether the smoother posterior landscape helps represent the functional variance of the model more truthfully or whether this is only a consequence of the increased degrees-of-freedom in the model's parameters.

While the answer also depends on the specific sampler employed, the setup in Figure 2, i.e., a *1-M-1* network, highlights one particular mechanism. In ReLU networks, samplers can become trapped in regions of the posterior landscape where the gradient vanishes (Sommer et al., 2024). For a positive input $x > 0$ w.l.o.g., and assuming a zero-centered symmetric posterior distribution of the first-layer weights $\mathbf{w}_1$, the probability that all hidden pre-activations are negative is 2^{-M}. Hence, this event becomes exponentially unlikely as the hidden width M increases. Combined with smoother marginal distributions in the overparametrized regime, this may improve the effectiveness of some samplers.

5.2 Balancedness Experiments

Figures 6 and 14 to 16 reveal more nuanced behavior. Observed patterns depend on two factors: the layer type (first, hidden, last) and the parameter type (bias vs. kernel). Specifically, weights in input and output layers exhibit distinct, multimodal or highly concentrated margins, whereas intermediate-layer weights yield margins closely matching their isotropic Gaussian priors. This backs our discussion of balancedness in networks with a bias in Section 4.3 as well as Sommer et al. (2024), which demonstrated enhanced sampler exploration in intermediate layers of BNNs.

5.3 Overparametrization Does Not Imply Data-Independency

Increasing the parameter count in deep networks can lead to problems in variational approximations of the posterior, potentially even inducing conditional independence of parameters and data (see, e.g., Trippe and Turner, 2017). Results from Section 4.1 and Figure 2 also suggest that the mean of the weights and their covariance decrease as the number of parameters increases, which is in line with this finding. In sampling-based inference, however, a zero-mean weight distribution does not imply a degenerate model, as it does not concentrate on a single solution as variational methods typically do. To investigate whether overparametrization can lead to diminishing covariance in larger (yet finite-width) neural networks, we run sampling-based inference for two networks of different sizes with good performance on the task. Figure 8 visualizes the empirical covariance and correlation matrices. By the intricate patterns visible in the plots, we conclude that the sampler is able to recover a complex correlation structure in parameter space, even when navigating a higher-dimensional parameter space. We further do not observe indications of data ignorance as suggested for variational methods (Coker et al., 2022). This suggests that while (some) marginals appear prior-like and most of them have zero mean (cf. Figure 6), there is still substantial correlation structure left, and prior conformity is much more likely a result of marginalization given the many directions with likelihood-flat regions.

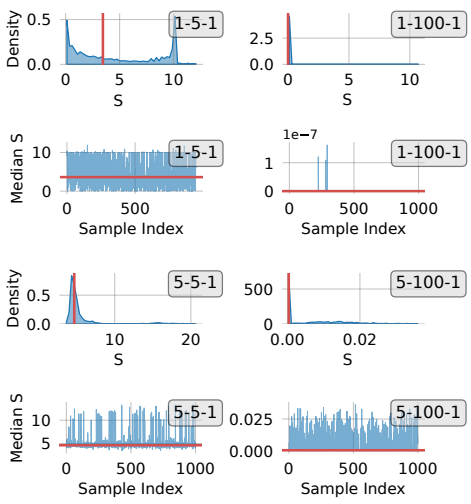


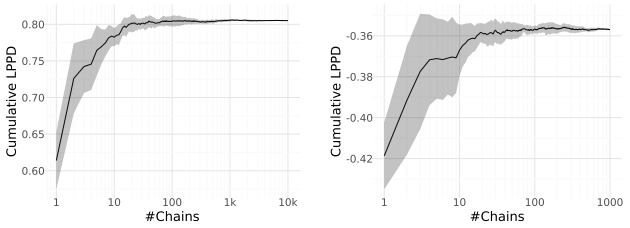
Figure 9: Minimal singular values S calculated for every hidden neuron cluster across all posterior samples, as well as median minimum singular value per neuron cluster per posterior sample $\mathbf{w}$. The median (first row) or the median of medians (second row) is indicated in red.

5.4 Flat Likelihood Directions and Prior Conformity

Following our discussion of *prior conformity* (Section 4.2), we probe for inter-neuron likelihood-flat directions in single-hidden layer ReLU networks by analyzing samples from a posterior chain, extending work like Ghorbani et al. (2019) beyond single parameter points. For each posterior sample $\mathbf{w}$, we cluster neurons with activation vectors on the training data, $\boldsymbol{\Xi}(\mathbf{w})$, that exhibit high cosine similarity. To identify invariant relationships, we project each cluster's activations onto the subspace orthogonal to the all-ones vector. A zero minimum singular value S implies a non-trivial, zero-sum combination of neuron activations that is constant across all data points. This signals a representational redundancy and thus a continuous manifold of non-identifiable parameters producing identical model outputs. We observe in Figure 9 that such redundancies, and corresponding likelihood-flat directions, become more prevalent as model size increases, appearing at nearly every posterior sample. This analysis is equivalent to examining $\ker(\mathbf{J})$ at each sample, as depicted in Figure 5 and detailed in Appendix C.

6 DISCUSSION

In this work, we showed that overparametrization reshapes BNN posteriors through the interplay of symmetries and priors. We demonstrated that network symmetries often align with specific weight priors, offering a principled notion of prior choice. Moreover, redundancy and homogeneity induce balancedness across layers, leading to equal-probability manifolds and prior conformity in redundant directions.



(a) A fully-connected BNN on `airfoil`.

(b) A convolutional BNN on `Fashion-MNIST`.

Figure 10: Cumulative LPPD increase over the number of chains (standard deviation across 5 random chain orderings) for different architectures and datasets.

Practical Considerations The absence of gaps in the marginal posterior of weights and their smoothness when averaging over a large number of chains suggests a smooth connected surface, making a full characterization of the BNN posterior practically possible. To this end, we compare different architectures and BNNs in terms of their cumulative log pointwise predictive density (LPPD, Gelman et al., 2014), examining the rate at which this metric saturates (details in Appendix D). While a scattered posterior with many different important regions separated by "loss barriers" might require a large number of chains or show no signs of saturation in uncertainty metrics, Figure 10 suggests convergence after around 10-20 chains. This result is reassuring as it shows that retrieving an approximate posterior via sampling is indeed practically feasible for BNNs.

Limitations A limitation of our work is its theoretical exposition's main focus on fully-connected ReLU networks. Yet, our empirical results indicate that the effects carry over to other architectures, including convolutional and residual networks. Another concern is that overparametrization increases dimensionality, which could, in principle, render sampling inefficient. However, recent results show that samplers like MCLMC (Robnik and Seljak, 2024) retain efficiency in high-dimensional regimes under certain assumptions, supporting the practical feasibility of sampling in large BNNs. Another limitation is the conditioning on a fixed assignment ς. While this assumption is not overly restrictive for single-chain approaches—since a single chain is unlikely to traverse between different assignments—a complete posterior characterization requires marginalization over all possible assignments. For ensembles of chains, such marginalization provides a more appropriate theoretical foundation.

References

Azulay, S., Moroshko, E., Nacson, M. S., Woodworth, B. E., Srebro, N., Globerson, A., and Soudry, D. (2021). On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Bona-Pellissier, J., Bachoc, F., and Malgouyres, F. (2023). Parameter Identifiability of a Deep Feedforward ReLU Neural Network. *Machine Learning*, 112(11):4431–4493.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: Composable transformations of Python+NumPy programs.

Cabezas, A., Corenflos, A., Lao, J., and Louf, R. (2024). BlackJAX: Composable Bayesian inference in JAX.

Chen, T., Fox, E., and Guestrin, C. (2014). Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691. PMLR.

Coker, B., Bruinsma, W. P., Burt, D. R., Pan, W., and Doshi-Velez, F. (2022). Wide mean-field bayesian neural networks ignore the data. In *International Conference on Artificial Intelligence and Statistics*, pages 5276–5333. PMLR.

Corti, A., Pacelli, R., Rotondo, P., and Gherardi, M. (2025). Microscopic and collective signatures of feature learning in neural networks.

Cui, H., Krzakala, F., and Zdeborová, L. (2023). Bayes-optimal learning of deep random networks of extensive-width. In *International Conference on Machine Learning*, pages 6468–6521. PMLR.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. (2021). Laplace Redux – Effortless Bayesian Deep Learning. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Dold, D., Kobialka, J., Palm, N., Sommer, E., Rügamer, D., and Dürr, O. (2025). Paths and Ambient Spaces in Neural Loss Landscapes. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR.

Dold, D., Rügamer, D., Sick, B., and Dürr, O. (2024). Semi-Structured Subspace Inference. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. (2018). Essentially No Barriers in Neural Network Energy Landscape. In *Proceedings of the 35th International Conference on Machine Learning*.

Du, S. S., Hu, W., and Lee, J. D. (2018). Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31.

Dua, D. and Graff, C. (2017). UCI Machine Learning Repository.

Duffield, S., Donatella, K., Chiu, J., Klett, P., and Simpson, D. (2025). Scalable bayesian learning with posteriors. In *The Thirteenth International Conference on Learning Representations*.

Fanaee-T, H. (2013). Bike Sharing Dataset. UCI Machine Learning Repository.

Farquhar, S., Smith, L., and Gal, Y. (2020). Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Fortuin, V. (2022). Priors in Bayesian Deep Learning: A Review. *International Statistical Review*, 90(3):563–591.

Freeman, C. D. and Bruna, J. (2017). Topology and Geometry of Half-Rectified Network Optimization. In *Proceedings of the 5th International Conference on Learning Representations*.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, page 10.

Gelberg, Y., van der Ouderaa, T. F., van der Wilk, M., and Gal, Y. (2024). Variational Inference Failures Under Model Symmetries: Permutation Invariant Posteriors for Bayesian Neural Networks. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding Predictive Information Criteria for Bayesian Models. *Statistics and Computing*, 24(6):997–1016.

Ghorbani, B., Krishnan, S., and Xiao, Y. (2019). An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR.

Hecht-Nielsen, R. (1990). On the Algebraic Structure of Feedforward Network Weight Spaces. In *Advanced Neural Computers*, pages 129–135. Elsevier.

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. (2020). Subspace Inference for Bayesian Deep Learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 1169–1179.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. (2021). What Are Bayesian Neural Network Posteriors Really Like? In *Proceedings of the 38th International Conference on Machine Learning, PMLR 139,*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Kim, S., Mishkin, A., and Pilanci, M. (2025). Exploring the loss landscape of regularized neural networks via convex duality. In *The Thirteenth International Conference on Learning Representations*.

Kolb, C., Frost, L., Bischl, B., and Rügamer, D. (2025a). Differentiable Sparsity via D-Gating: Simple and Versatile Structured Penalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kolb, C., Müller, C. L., Bischl, B., and Rügamer, D. (2026). Smoothing the Edges: Smooth Optimization for Sparse Regularization using Hadamard Overparametrization. *Machine Learning*. To appear.

Kolb, C., Weber, T., Bischl, B., and Rügamer, D. (2025b). Deep Weight Factorization: Sparse Learning Through the Lens of Artificial Symmetries. In *The Thirteenth International Conference on Learning Representations*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kunin, D., Raventós, A., Dominé, C., Chen, F., Klindt, D., Saxe, A., and Ganguli, S. (2024). Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. *Advances in Neural Information Processing Systems*, 37:81157–81203.

Kurle, R., Januschowski, T., Gasthaus, J., and Wang, Y. B. (2021). On Symmetries in Variational Bayesian Neural Nets.

Laurent, O., Aldea, E., and Franchi, G. (2024). A Symmetry-Aware Exploration of Bayesian Neural Network Posteriors. In *The Twelfth International Conference on Learning Representations*.

Liu, C., Zhu, L., and Belkin, M. (2020). Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 7.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. (2019). Invariant and Equivariant Graph Networks. In *International Conference on Learning Representations*.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76.

Mishkin, A., Sahiner, A., and Pilanci, M. (2022). Fast Convex Optimization for Two-Layer ReLU Networks: Equivalent Model Classes and Cone Decompositions. In *International Conference on Machine Learning*, pages 15770–15816. PMLR.

Nalisnick, E. T. (2018). *On Priors for Bayesian Neural Networks*. PhD thesis, University of California, Irvine.

Navon, A., Shamsian, A., Achituve, I., Fetaya, E., Chechik, G., and Maron, H. (2023). Equivariant architectures for learning in deep weight spaces. *arXiv preprint arXiv:2301.12780*.

Nguyen, Q. (2019). On connected sublevel sets in deep learning. In *International conference on machine learning*, pages 4790–4799. PMLR.

Papamarkou, T. (2023). Approximate blocked Gibbs sampling for Bayesian neural networks. *Statistics and Computing*, 33.

Papamarkou, T., Hinkle, J., Young, M. T., and Womble, D. (2022). Challenges in Markov Chain Monte Carlo for Bayesian Neural Networks. *Statistical Science*, 37(3).

Papamarkou, T., Skoularidou, M., Palla, K., Aitchison, L., Arbel, J., Dunson, D., Filippone, M., Fortuin, V., Hennig, P., Hernández-Lobato, J. M., Hubin, A., Immer, A., Karaletsos, T., Khan, M. E., Kristiadi, A., Li, Y., Mandt, S., Nemeth, C., Osborne, M. A., Rudner, T. G. J., Rügamer, D., Teh, Y. W., Welling, M., Wilson, A. G., and Zhang, R. (2024). Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR.

Parhi, R. and Nowak, R. D. (2023). Deep learning meets sparse regularization: A signal processing perspective. *IEEE Signal Processing Magazine*, 40(6):63–74.

Paulin, D., Whalley, P. A., Chada, N. K., and Leimkuhler, B. J. (2025). Sampling from bayesian neural network posteriors with symmetric minibatch splitting langevin dynamics. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Pavliotis, G. A. (2014). Stochastic processes and applications. *Texts in applied mathematics*, 60.

Petzka, H., Trimmel, M., and Sminchisescu, C. (2020). Notes on the Symmetries of 2-Layer ReLU-Networks. In *Proceedings of the northern lights deep learning workshop*, volume 1, pages 6–6.

Phuong, M. and Lampert, C. (2020). Functional vs. Parametric Equivalence of ReLU Networks. In *8th International Conference on Learning Representations*.

Pourzanjani, A. A., Jiang, R. M., and Petzold, L. R. (2017). Improving the Identifiability of Neural Networks for Bayesian Inference. In *Second Workshop on Bayesian Deep Learning*.

Robnik, J. and Seljak, U. (2024). Fluctuation without dissipation: Microcanonical langevin monte carlo. In *Symposium on Advances in Approximate Bayesian Inference*, pages 111–126. PMLR.

Rolnick, D. and Kording, K. (2020). Reverse-Engineering Deep ReLU Networks. In *International conference on machine learning*. PMLR.

Roy, H., Miani, M., Ek, C. H., Hennig, P., Pförtner, M., Tatzel, L., and Hauberg, S. (2024). Reparameterization invariance in approximate bayesian inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sen, D., Papamarkou, T., and Dunson, D. (2024). Bayesian Neural Networks and Dimensionality Reduction. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*. Chapman and Hall/CRC.

Sigillito, V., Wing, S., Hutton, L., and Baker, K. (1989). Ionosphere. UCI Machine Learning Repository.

Simsek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W., and Brea, J. (2021). Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances. In *Proceedings of the 38 Th International Conference on Machine Learning*.

Sommer, E., Diao, K., Robnik, J., Seljak, U., and Rügamer, D. (2026). Can microcanonical langevin dynamics leverage mini-batch gradient noise?

Sommer, E., Robnik, J., Nozadze, G., Seljak, U., and Rügamer, D. (2025). Microcanonical Langevin Ensembles: Advancing the Sampling of Bayesian Neural Networks. In *The Thirteenth International Conference on Learning Representations*.

Sommer, E., Wimmer, L., Papamarkou, T., Bothmann, L., Bischl, B., and Rügamer, D. (2024). Connecting the Dots: Is Mode-Connectedness the Key to Feasible Sample-Based Inference in Bayesian Neural Networks? In *Proceedings of the 41st International Conference on Machine Learning*. PMLR.

Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. (2016). Bayesian optimization with robust bayesian neural networks. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Trippe, B. and Turner, R. (2017). Overpruning in variational bayesian neural networks.

Tsanas, A. and Xifara, A. (2012). Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools. *Energy and Buildings*, 49.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019). Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pages 6458–6467. PMLR.

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the Bayes posterior in deep neural networks really? In *Proceedings of the 37th International Conference on Machine Learning*. PMLR.

Wiese, J. G., Wimmer, L., Papamarkou, T., Bischl, B., Günnemann, S., and Rügamer, D. (2023). Towards efficient MCMC sampling in Bayesian neural networks by exploiting symmetry. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 459–474. Springer.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms.

Yeh, I.-C. (1998). Modeling of Strength of High-Performance Concrete Using Artificial Neural Networks. *Cement and Concrete research*, 28(12).

Zhang, R., Li, C., Zhang, J., Chen, C., and Wilson, A. G. (2020). Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning. In *Proceedings of the Eighth International Conference on Learning Representations*.

Ziyin, L. (2024). Symmetry induces structure and constraint of learning. In *International Conference on Machine Learning*, pages 62847–62866. PMLR.

Ziyin, L., Wang, M., Li, H., and Wu, L. (2024). Parameter Symmetry and Noise Equilibrium of Stochastic Gradient Descent. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ziyin, L., Xu, Y., and Chuang, I. (2025a). Neural thermodynamics i: Entropic forces in deep and universal representation learning. *arXiv preprint arXiv:2505.12387*.

Ziyin, L., Xu, Y., and Chuang, I. (2025b). Remove Symmetries to Control Model Expressivity. In *The Thir-*

teenth International Conference on Learning Representations.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, provided in Sections 3 and 4 and Appendices C and D.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, link provided in Appendix D.1.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, in Appendices B and C.
   (b) Complete proofs of all theoretical results. Yes, in Appendix C.
   (c) Clear explanations of any assumptions. Yes, in Sections 3 and 4 and Appendices B and C.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, in Appendix D.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, in Appendix D.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, in Appendix D and all respective experiment descriptions.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, in Appendix D.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. Yes, we use code and reference it in Appendix D.
   (b) The license information of the assets, if applicable. Not Applicable. All code used is open-source under common licenses that allow usage and extension.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

On the Interplay of Priors and Overparametrization in Bayesian Neural Network Posteriors: Supplementary Materials

A ADDITIONAL RELATED WORK

Previous works related to sampling-based inference frame symmetries as equioutput parameter states, where different parameter sets $\mathbf{w}$ in the weight space $\mathcal{W}$ of the neural network lead to the same functional mapping (Hecht-Nielsen, 1990; Wiese et al., 2023), i.e., $\exists \mathbf{w}, \tilde{\mathbf{w}} \in \mathcal{W}, \mathbf{w} \neq \tilde{\mathbf{w}} : f_{\mathbf{w}}(\mathbf{x}) = f_{\tilde{\mathbf{w}}}(\mathbf{x}) \, \forall \mathbf{x} \in \mathcal{X}$, that is, the same unnormalized log-posterior value if $\mathbf{w}, \tilde{\mathbf{w}}$ have the same prior probability which is shown for the special case of permutation symmetries in Wiese et al. (2023). Showing functional equivalence of networks when parameters admit an equivalence relationship, i.e., $\mathbf{w} \sim \tilde{\mathbf{w}} \Rightarrow f_{\mathbf{w}}(\mathbf{x}) = f_{\tilde{\mathbf{w}}}(\mathbf{x})$ is more straightforward (Bona-Pellissier et al., 2023; Petzka et al., 2020; Phuong and Lampert, 2020; Pourzanjani et al., 2017), while deriving parameter equivalence from equivalent outputs, i.e., $f_{\mathbf{w}}(\mathbf{x}) = f_{\tilde{\mathbf{w}}}(\mathbf{x}) \Rightarrow \mathbf{w} \sim \tilde{\mathbf{w}}$ requires stronger and often impractical assumptions (Bona-Pellissier et al., 2023; Phuong and Lampert, 2020; Rolnick and Kording, 2020) which is commonly summarized as the non-identifiability of parameters in neural networks. Ziyin (2024) abstracts symmetries into a common mirror reflection symmetry that structures the loss landscape (Ziyin et al., 2024). Each symmetry carves out a low-capacity subspace defined by a linear constraint (e.g., $O^\top \mathbf{w} = 0$) that acts as an "absorbing state" for optimization dynamics. As detailed in Ziyin (2024); Ziyin et al. (2024, 2025a,b), symmetries in combination with sufficient L_2 regularization strength $\tau > \tau_0$ for the penalty $\tau ||\mathbf{w}||_2^2$ affect SGD optimization by partitioning the parameter space into symmetry-aligned as well as symmetry-orthogonal subspaces. In multiplicative overparametrization, L_2 regularization then forces weights to be balanced across groups of shared neurons (Kolb et al., 2025a,b, 2026).

Overparametrization Induces Connectedness The optimization landscape of neural networks is profoundly affected by overparametrization. In under-parameterized models, the loss surface can be fraught with poor local minima, but as the number of parameters increases, this landscape often simplifies, facilitating optimization. Seminal work has shown that for sufficiently wide networks without regularization, the sublevel sets of the loss function become connected, effectively eliminating "bad local valleys" and ensuring that all global minima reside within a single, large basin (Freeman and Bruna, 2017; Simsek et al., 2021; Nguyen, 2019; Kim et al., 2025). This connectivity is largely driven by the inherent symmetries of neural networks. For instance, the permutation symmetry, which creates numerous equivalent discrete minima in a minimally-sized network, can generate a single connected manifold of global minima with the addition of just one extra neuron per layer (Simsek et al., 2021). This transition to a more benign landscape, which may not be locally convex but often satisfies favorable conditions like the Polyak-Lojasiewicz condition (Liu et al., 2020), is a key reason why gradient-based optimization succeeds in large-scale models. Adding regularization fundamentally alters the loss landscape compared to the unregularized case. For instance, L_2 regularization (i.e., weight decay) breaks the positive homogeneity of ReLU units, which can reduce the number of optimal solutions from infinite to finite and create a more structured set of optimizers (Kim et al., 2025). For models trained with mixed L_1 and L_2 regularization, asymptotic connectivity of sublevel sets has also been shown with the loss barrier to connect two minima shrinking as overparametrization increases (Freeman and Bruna, 2017).

Next to the trivial case of the infinite-width limit of overparametrization (i.e., lazy regime, Jacot et al., 2018), the "proportional regime" where the ratio of the number of observations and the number of hidden neurons remains fixed as both increase, $\frac{n}{M} = \alpha$ as $n, M \to \infty$, has been considered in the literature (Cui et al., 2023; Corti et al., 2025). As opposed to the lazy regime, here, training elicits complex correlations between the parameters, fundamentally changing the model's internal structure to adapt to the data while resembling a Gaussian process in the function space (Corti et al., 2025). This also relates to Trippe and Turner (2017), analyzing a potential pathology where more expressive variational approximations of the BNN posterior can paradoxically lead to worse

performance as the optimization process induces conditional independence between parameters and data.

BNN Priors Prior choice is a widely debated issue in BDL, with opinions ranging from the claim that priors in weight space are meaningless due to high dimensionality and identifiability concerns, to the assertion that meaningful priors do exist, though not in the form of the commonly used isotropic Gaussian priors with constant variance (Fortuin, 2022; Vladimirova et al., 2019).

B OMITTED DEFINITIONS AND ASSUMPTIONS

We start by stating our assumptions and defining overparametrization as a network with excessive neurons that overparametrizes a minimum norm interpolant. This is the basis for our results in Section 4.1.

Assumption 1 (Overparametrization)**.** Let $f(\mathbf{x}) = \sum_{m=1}^{M} w_{2,m}\phi(\mathbf{w}_{1,m}^\top \mathbf{x})$ and assume there is a *p-M^*-1* interpolant $f^*(\mathbf{x}) = \sum_{m=1}^{M^*} w_{2,m}^*\phi(\mathbf{w}_{1,m}^{*\top}\mathbf{x})$ that attains the minimal norm cost among all exact interpolants of width $\leq M^*$. Assume the feature vectors $\Phi(\mathbf{w}_{1,m}^*) \in \mathbb{R}^n$ with $\Phi(\mathbf{v}) = \{\phi(\langle \mathbf{v}, \mathbf{x}_i\rangle)\}_{i=1}^n$ are nonzero, linearly independent (i.e., the network is not overparametrized) and identifiable up to permutations.

Furthermore, in our experiments, we use the following layerwise overparametrization notion.

Definition 3 (Layerwise overparametrization)**.** Let layer l have M_l hidden units and define the layer's feature matrix on the training inputs $\mathbf{X} = (\mathbf{x}_i)_{i=1}^n$ by

$$\mathbf{\Xi}_l(\mathbf{w}) \;:=\; \big(\xi_{l,m}(\mathbf{x}_i;\mathbf{w})\big)_{i\in[n],\, m\in[M_l]} \in \mathbb{R}^{n\times M_l},$$

where $\xi_{l,m}$ is the contribution of hidden unit m to the next layer. We call layer l *overparametrized* if $\operatorname{rank}\big(\mathbf{\Xi}_l(\mathbf{w}^{\text{ref}})\big) =: M_l^* < M_l$ at a reference parameter $\mathbf{w}^{\text{ref}}$. The rank deficiency is $r_l := M_l - M_l^* > 0$.

Overparametrization means that the column span of $\mathbf{\Xi}_l$ has dimension M_l^* while M_l units parameterize it. Any two parameter vectors with the same $\mathbf{\Xi}_l$ produce the same next-layer input up to the linear map W_{l+1}.

For transparency, we collect all assumptions underlying the results in Section 4. Corollary 1 requires Assumption 2 and Assumption 3. Corollary 2 further requires Assumption 4. The experimental results do not rely on any of these assumptions and serve as independent validation.

Assumption 2 (Manifold regularity)**.** For a fixed assignment ς, the minimum-norm manifold $\mathcal{M}_\varsigma$ is a closed embedded C^2 manifold of $\mathbb{R}^d$ with positive reach, i.e., every point in $\mathcal{M}_\varsigma^\varepsilon$ has a unique nearest point on $\mathcal{M}_\varsigma$.

Assumption 3 (Volume factorization)**.** In the ε-tube around $\mathcal{M}_\varsigma$, the contribution of directions orthogonal to the reallocation coordinates $\boldsymbol{\rho}^{(\varpi)}$ to the tube volume is independent of $\boldsymbol{\rho}^{(\varpi)}$.

Assumption 4 (Existence of the weak limit)**.** The tube-conditioned posterior $\mathbb{P}_n^\varepsilon$ (Definition 2) converges weakly as $\varepsilon \downarrow 0$ to a well-defined limit $\mathbb{P}_n$ on $\mathcal{M}_\varsigma$.

Assumption 4 is required for the moment bounds in Corollary 2. Under Assumption 2, this is expected to hold, but we do not verify it for specific ReLU network architectures.

C OMITTED THEORETICAL RESULTS, PROOFS AND DERIVATIONS

The statements made in this section, referring to statements and proofs of Section 4.1, assume the following overparametrization.

Definition 4. We define the surjective map $\varsigma : [M] \to [M^*]$, referred to as the assignment or reallocation map, such that for every $\varpi \in [M^*]$, $\sum_{m\in G_\varpi} \rho_m = 1$ with coefficients $(\rho_m)_{m\in[G_\varpi]} \in \Delta^{k_\varpi - 1}$ on a simplex with dimension defined by $k_\varpi := |G_\varpi|$ and $G_\varpi := \{m \in [M] : \varsigma(m) = \varpi\}$. Further, let $\Delta := \prod_{\varpi\in[M^*]} \Delta^{k_\varpi - 1}$.

Lemma 1. *Under fixed assignment map ς, Assumption 1 and according to Definition 1,*

$$\mathcal{V}_\varpi := \{w_{2,m}, \mathbf{w}_{1,m}, m \in [M] : \\ \mathbf{w}_{1,m} = \sqrt{\rho_m}\mathbf{w}_{1,\varsigma(m)}^*, w_{2,m} = \operatorname{sign}(w_{2,\varsigma(m)}^*)\sqrt{\rho_m}|w_{2,\varsigma(m)}^*|, (\rho_m)_{m\in[M]} \in \Delta\} \;\in\; \mathcal{M}.$$

Proof. We must show (1) $f(\mathbf{x}, \mathbf{w}) = f^*(\mathbf{x}, \mathbf{w}^*)$ for all $\mathbf{x} \in \mathcal{X}$ and all $\mathbf{w} \in \mathcal{V}_\varpi$, and (2) $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w}^*)$, where $\mathcal{R}(\mathbf{w}) := \sum_{m=1}^{M} \|\mathbf{w}_{1,m}\|_2^2 + w_{2,m}^2$. By the definition of $\mathcal{V}_\varpi$, for each hidden index $m \in [M]$, we set

$$\mathbf{w}_{1,m} = \sqrt{\rho_m}\, \mathbf{w}^*_{1,\varsigma(m)}, \qquad w_{2,m} = \operatorname{sign}\big(w^*_{2,\varsigma(m)}\big)\, \sqrt{\rho_m}\, \big|w^*_{2,\varsigma(m)}\big|,$$

with coefficients $\rho_m \geq 0$ satisfying $\sum_{m \in G_\varpi} \rho_m = 1$ for every $\varpi \in [M^*]$. Using the 1-homogeneity of ReLU, $\phi(c\,t) = c\,\phi(t)$ for all $c \geq 0$, we have the following.

1) Function equality.

$$\begin{aligned}
f(\mathbf{x}, \mathbf{w}) &= \sum_{m=1}^{M} w_{2,m}\, \phi(\mathbf{w}_{1,m}^\top \mathbf{x}) \\
&= \sum_{m=1}^{M} \operatorname{sign}\big(w^*_{2,\varsigma(m)}\big) \sqrt{\rho_m}\, |w^*_{2,\varsigma(m)}|\, \phi\big(\sqrt{\rho_m}\, \mathbf{w}^{*\top}_{1,\varsigma(m)} \mathbf{x}\big) \\
&= \sum_{m=1}^{M} \operatorname{sign}\big(w^*_{2,\varsigma(m)}\big)\, \rho_m\, |w^*_{2,\varsigma(m)}|\, \phi\big(\mathbf{w}^{*\top}_{1,\varsigma(m)} \mathbf{x}\big) \\
&= \sum_{\varpi=1}^{M^*} \operatorname{sign}(w^*_{2,\varpi})\, |w^*_{2,\varpi}|\, \phi\big(\mathbf{w}^{*\top}_{1,\varpi} \mathbf{x}\big) \big(\textstyle\sum_{m \in G_\varpi} \rho_m\big) \\
&= \sum_{\varpi=1}^{M^*} w^*_{2,\varpi}\, \phi\big(\mathbf{w}^{*\top}_{1,\varpi} \mathbf{x}\big) \;=\; f^*(\mathbf{x}, \mathbf{w}^*),
\end{aligned}$$

since $\operatorname{sign}(a)\,|a| = a$ and $\sum_{m \in G_\varpi} \rho_m = 1$.

2) Norm equality.

$$\begin{aligned}
\mathcal{R}(\mathbf{w}) &= \sum_{m=1}^{M} \|\mathbf{w}_{1,m}\|_2^2 + w_{2,m}^2 \\
&= \sum_{m=1}^{M} \rho_m\, \|\mathbf{w}^*_{1,\varsigma(m)}\|_2^2 \;+\; \rho_m\, \big|w^*_{2,\varsigma(m)}\big|^2 \\
&= \sum_{\varpi=1}^{M^*} \Big(\|\mathbf{w}^*_{1,\varpi}\|_2^2 + \big|w^*_{2,\varpi}\big|^2\Big) \big(\textstyle\sum_{m \in G_\varpi} \rho_m\big) \\
&= \sum_{\varpi=1}^{M^*} \Big(\|\mathbf{w}^*_{1,\varpi}\|_2^2 + \big|w^*_{2,\varpi}\big|^2\Big) \;=\; \mathcal{R}(\mathbf{w}^*).
\end{aligned}$$

Thus $\mathbf{w} \in \mathcal{M}$ by Definition 1, which proves the claim. □

C.1 Derivation of Theorem 1

We first restate Theorem 1 again using the more formal setup of Lemma 1:

Theorem 3. *Assume the setup of Lemma 1. Fix any assignment $\varsigma : [M] \to [M^*]$. Let $\boldsymbol{\rho}^{(\varpi)} = (\rho_m)_{m \in G_\varpi}$. On $\mathcal{M}_\varsigma$,*

$$\boldsymbol{\rho}^{(\varpi)} \sim \text{Dirichlet}\left(\tfrac{1}{2}, \ldots, \tfrac{1}{2}\right) \quad \textit{for each } \varpi = 1, \ldots, M^*,$$

and the random vectors $\boldsymbol{\rho}^{(1)}, \ldots, \boldsymbol{\rho}^{(M^*)}$ are independent.

Proof. By Lemma 1, for fixed ς, the minimum–norm manifold is parametrized blockwise with blocks G_ϖ. Within one of these blocks G_ϖ, all solutions $\boldsymbol{\rho}^{(\varpi)} \in \Delta^{k_\varpi - 1}$. We first prove the distribution result and then derive its origin. First, parameterize each block $G_\varpi, \varpi \in [M^*]$, by nonnegative amplitudes $\boldsymbol{\alpha}^{(\varpi)} = (\alpha_m)_{m \in G_\varpi} \in \mathbb{S}_+^{k_\varpi - 1} := \left\{(\alpha_m)_{m \in G_\varpi} \in \mathbb{R}_{\geq 0}^{k_\varpi} : \sum_{m \in G_\varpi} \alpha_m^2 = 1\right\}$ and define $\boldsymbol{\rho}^{(\varpi)} = (\rho_m)_{m \in G_\varpi}$, $\rho_m := \alpha_m^2$. The Dirichlet

distribution follows from the change of measure. Let $\mathbf{v}^{(\varpi)} = (v_m)_{m \in G_\varpi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{k_\varpi})$. Then $\boldsymbol{\alpha}^{(\varpi)} \overset{d}{=} \frac{|\mathbf{v}^{(\varpi)}|}{\|\mathbf{v}^{(\varpi)}\|_2}$ and $\rho_m = \alpha_m^2 = \frac{v_m^2}{\sum_{j \in G_\varpi} v_j^2}$. Since v_m^2 are independent χ_1^2 variables, $\boldsymbol{\rho}^{(\varpi)}$ is Dirichlet$(\frac{1}{2}, \dots, \frac{1}{2})$. Independence follows from independence of elements in $\mathbf{v}^{(\varpi)}$.

It remains to justify that, conditional on the fixed assignment ς and on $\mathbf{w} \in \mathcal{M}_\varsigma$, the induced law of the amplitude vector $\boldsymbol{\alpha}^{(\varpi)} \in \mathbb{S}_+^{k_\varpi - 1}$ is the uniform surface measure (on the positive orthant), and that different blocks factorize. By Lemma 1, the likelihood is constant on $\mathcal{M}_\varsigma$ as is the (radial) prior. Hence the posterior restricted to $\mathcal{M}_\varsigma$ is proportional to the intrinsic volume measure on $\mathcal{M}_\varsigma$. On $\mathcal{M}_\varsigma$, $\boldsymbol{\omega}_m = \alpha_m \boldsymbol{\omega}_\varpi^*$ with $\boldsymbol{\alpha}^{(\varpi)} \in \mathbb{S}_+^{k_\varpi - 1}$. The map $\boldsymbol{\alpha}^{(\varpi)} \mapsto (\boldsymbol{\omega}_m)_{m \in G_\varpi}$ is linear and its Jacobian has constant magnitude $(\|\boldsymbol{\omega}_\varpi^*\|_2)^{k_\varpi - 1}$, independent of $\boldsymbol{\alpha}^{(\varpi)}$. Therefore, the induced density of $\boldsymbol{\alpha}^{(\varpi)}$ is constant on $\mathbb{S}_+^{k_\varpi - 1}$, i.e., $\boldsymbol{\alpha}^{(\varpi)}$ is uniform on $\mathbb{S}_+^{k_\varpi - 1}$. Moreover, the parametrization is a product over blocks, so the intrinsic volume measure (hence the posterior) factorizes across blocks and independence of $\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(M^*)}$ follows. □

C.2 Proof of Corollary 1

Proof. For a fixed block G_ϖ, write each duplicated neuron as

$$\boldsymbol{\omega}_m = r_m \boldsymbol{u}_\varpi^*, \qquad r_m := \|\boldsymbol{\omega}_m\| = \sqrt{\rho_m}\, \|\boldsymbol{\omega}_\varpi^*\|,$$

where $\boldsymbol{u}_\varpi^* := \boldsymbol{\omega}_\varpi^* / \|\boldsymbol{\omega}_\varpi^*\|$ is a unit vector. In $\mathbb{R}^{p+1}$, the set of vectors with fixed norm r_m forms a p-dimensional sphere $\mathbb{S}^p(r_m)$. Hence, in a small ε-neighborhood of the ray $\{t\, \boldsymbol{u}_\varpi^* : t \geq 0\}$, the orthogonal directions correspond locally to perturbations within this p-dimensional sphere. The p-dimensional surface measure of $\mathbb{S}^p(r_m)$ scales as

$$\mathrm{Vol}\big(\mathbb{S}^p(r_m)\big) \propto r_m^p.$$

Therefore, the cross-sectional volume in the p orthogonal directions associated with the m-th duplicated neuron scales as $r_m^p = (\sqrt{\rho_m}\, \|\boldsymbol{\omega}_\varpi^*\|)^p \propto \rho_m^{p/2}$. All remaining factors are independent of $\boldsymbol{\rho}$ and cancel after normalization. Therefore, the total tube-volume weight attached to a configuration $\boldsymbol{\rho}$ is

$$\mathrm{Vol}(\boldsymbol{\rho}) \;\propto\; \prod_{m=1}^{k} \rho_m^{p/2}.$$

To obtain the density of $\boldsymbol{\rho}$, we multiply this volume by the density $\prod_{m=1}^k \rho_m^{-1/2}$ corresponding to Dirichlet$(\frac{1}{2})$. This yields

$$p_{\mathrm{tube}}(\boldsymbol{\rho}) \;\propto\; \prod_{m=1}^{k} \rho_m^{-1/2}\, \rho_m^{p/2} = \prod_{m=1}^{k} \rho_m^{\frac{p+1}{2} - 1}.$$

This is exactly the density of a symmetric Dirichlet distribution with concentration parameters $(\frac{p+1}{2}, \dots, \frac{p+1}{2})$. Independence across blocks follows from the product structure of the parameterization for fixed ς. □

C.3 Proof of Corollary 2

We again state Corollary 2 more precisely.

Corollary 4. *Under the setup of Theorem 3, fixing an assignment ς and under the tube-law $\mathbb{P}_n$ induced on the corresponding manifold $\mathcal{M}_\varsigma$, let G_ϖ be a block of size k_ϖ and let w_m and $w_{m'}$ denote two scalar weights from the same block with $w_m \neq w_{m'}$. Then*

$$\mathbb{E}_{\mathbb{P}_n}(w_m) = O(k_\varpi^{-1/2}), \qquad \mathrm{Cov}_{\mathbb{P}_n}(w_m, w_{m'}) = O(k_\varpi^{-2}).$$

Assuming the group size k_ϖ to increase proportionally with M, i.e., $\exists c > 0 : k_\varpi > cM$, one has $\mathbb{E}_{\mathbb{P}_n}(w) = O(M^{-1/2})$ and $\mathrm{Cov}_{\mathbb{P}_n}(w_m, w_{m'}) = O(M^{-2})$. If w_m and $w_{m'}$ belong to different blocks, $\mathrm{Cov}_{\mathbb{P}_n}(w_m, w_{m'}) = 0$.

Proof. In the following, we assume all expectations and covariances to be taken w.r.t. the distribution $\mathbb{P}_n$ and drop the corresponding index. Further, to make a statement w.r.t. M while working with group size k_ϖ, we assume that both grow proportionally, i.e., $\exists c > 0 : k_\varpi > cM$.

Fix a block G_ϖ of size $k := k_\varpi$ and write the duplicate amplitudes as $a_m \geq 0$ with $\sum_{m \in G_\varpi} a_m^2 = 1$ and $\rho_m := a_m^2$. Following Corollary 1 for fixed assignment, $\boldsymbol{\rho} := (\rho_m)_{m \in G_\varpi} \sim \text{Dirichlet}(\frac{p+1}{2}, \ldots, \frac{p+1}{2})$. This implies that the marginal distribution of each reallocation is $\rho_m \sim \text{Beta}(\alpha, (k-1)\alpha)$ for $\alpha = (p+1)/2$.

Mean: Since for every $w_m, m \in G_\varpi$, it holds $w_m = \sqrt{\rho_m} w^*_\varpi$, it follows directly

$$\mathbb{E}[w_m] = \mathbb{E}[\sqrt{\rho_m} w^*_\varpi] = \mathbb{E}[\sqrt{\rho_m}]\ w^*_\varpi = \mathbb{E}[\sqrt{\rho_m}]\ w^*_\varpi = \frac{\Gamma(\alpha + \frac{1}{2})\Gamma(k\alpha)}{\Gamma(\alpha)\Gamma(k\alpha + \frac{1}{2})} w^*_\varpi = \Theta(k^{-1/2}) = \mathcal{O}(M^{-1/2}).$$

assuming $\exists c > 0 : k > cM$.

Covariance within a block: Applying the same logic for two weights $w_m, w_{m'} \in G_\varpi$ with $w_m \neq w_{m'}$, we have

$$\begin{aligned}\text{Cov}(w_m, w_{m'}) &= \mathbb{E}[w_m w_{m'}] - \mathbb{E}[w_m]\mathbb{E}[w_{m'}] = (w^*_\varpi)^2 \text{Cov}(w_m, w_{m'}) \\ &= (w^*_\varpi)^2 \left(\frac{\Gamma(\alpha + \frac{1}{2})^2, \Gamma(k\alpha)}{\Gamma(\alpha)^2, \Gamma(k\alpha + 1)} - \left(\frac{\Gamma(\alpha + \frac{1}{2})\Gamma(k\alpha)}{\Gamma(\alpha)\Gamma(k\alpha + \frac{1}{2})} \right)^2 \right) = \mathcal{O}(k^{-2}) = \mathcal{O}(M^{-2}).\end{aligned}$$

Covariance between blocks: Last, since $\text{Cov}(\rho_m, \rho_{m'}) = 0$ for $\varsigma(m) \neq \varsigma(m')$, it follows $\text{Cov}_{\mathbb{P}_n}(w_m, w_{m'}) = 0$.

□

C.4 Proof of Theorem 2

Proof. We first repeat that we have

$$\mathbb{E}_\pi[\langle \mathbf{W}_l, \nabla_{\mathbf{W}_l} \mathcal{L}(\mathbf{w}) \rangle_F] \ = \ \mathbb{E}_\pi[\langle \mathbf{W}_{l+1}, \nabla_{\mathbf{W}_{l+1}} \mathcal{L}(\mathbf{w}) \rangle_F], \tag{4}$$

which is based on the pointwise positive homogeneity of $f(\mathbf{x}) = \mathbf{W}_L \phi(\mathbf{W}_{L-1} \phi(\cdots(\mathbf{W}_1 \mathbf{x})))$ (Du et al., 2018). This provides the inner identity of (4), which we extend under standard regularity conditions to the expectation in the stationary limit of a consistent sampler (Pavliotis, 2014).

Stein's identity for π with the test function $f(\mathbf{W}) = \mathbf{W}_l$ gives

$$\mathbb{E}_\pi\big[\langle \mathbf{W}_l, \nabla_{\mathbf{W}_l} \mathcal{L}_{\boldsymbol{\tau}}(\mathbf{w}) \rangle_F\big] \ = \ d_l.$$

Using $\nabla_{\mathbf{W}_l} \mathcal{L}_{\boldsymbol{\tau}}(\mathbf{w}) = \nabla_{\mathbf{W}_l} \mathcal{L}(\mathbf{w}) + \tau_l^{-2} \mathbf{W}_l$,

$$\mathbb{E}_\pi\big[\langle \mathbf{W}_l, \nabla_{\mathbf{W}_l} \mathcal{L}(\mathbf{w}) \rangle_F\big] \ + \ \frac{1}{\tau_l^2} \mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] \ = \ d_l.$$

Apply the same identity with $l{+}1$ and subtract. By the homogeneity identity,

$$\frac{1}{\tau_l^2} \mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] - \frac{1}{\tau_{l+1}^2} \mathbb{E}_\pi\big[\|\mathbf{W}_{l+1}\|_F^2\big] \ = \ d_l - d_{l+1}.$$

□

Corollary 5. *Under the assumptions of the Layer-balance theorem, define*

$$B_l \ := \ \frac{1}{\tau_l^2} \mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] \ - \ d_l, \qquad h = 1, \ldots, L.$$

Then B_l is constant across layers, i.e.,

$$B_1 \ = \ B_2 \ = \ \cdots \ = \ B_L.$$

In particular, if there exists an index h_0 such that $\mathbb{E}_\pi[\|\mathbf{W}_{h_0}\|_F^2] = \tau_{h_0}^2 d_{h_0}$, then

$$\mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] \ = \ \tau_l^2 d_l \quad \textit{for all } h = 1, \ldots, L.$$

C.5 Proof of Corollary 5

Proof. By the Layer-balance theorem, for each adjacent pair $(l, l+1)$,

$$\frac{1}{\tau_l^2}\,\mathbb{E}_\pi\big[\|\mathbf{W}_l\|_F^2\big] - \frac{1}{\tau_{l+1}^2}\,\mathbb{E}_\pi\big[\|\mathbf{W}_{l+1}\|_F^2\big] \;=\; d_l - d_{l+1}.$$

Rearranging gives $\big(\tau_l^{-2}\mathbb{E}_\pi\|\mathbf{W}_l\|_F^2 - d_l\big) = \big(\tau_{l+1}^{-2}\mathbb{E}_\pi\|\mathbf{W}_{l+1}\|_F^2 - d_{l+1}\big)$, hence $B_l = B_{l+1}$ for all l. Transitivity yields $B_1 = \cdots = B_L$. If for some l_0 we have $\mathbb{E}_\pi\|\mathbf{W}_{l_0}\|_F^2 = \tau_{l_0}^2 d_{l_0}$, then $B_{l_0} = 0$ and thus $B_l = 0$ for all l, which implies $\mathbb{E}_\pi\|\mathbf{W}_l\|_F^2 = \tau_l^2 d_l$ for all l. □

Proposition 1. *Assume a homogeneous network $f(\mathbf{w})$ that admits a rescaling symmetry between layer weights of layer l and $l+1$ and a layer-specific penalty $\mathcal{R}(\mathbf{w}) = \ldots + \lambda_l||\mathbf{W}_l||_F^2 + \lambda_{l+1}||\mathbf{W}_{l+1}||_F^2 + \ldots$. Along any rescaling that preserves the represented function, the penalty is minimized precisely when*

$$\frac{\|\mathbf{W}_l\|_F}{\|\mathbf{W}_{l+1}\|_F} \;=\; \sqrt{\frac{\lambda_{l+1}}{\lambda_l}}.$$

Proof. By positive 1-homogeneity, for any $a > 0$ the transformation

$$\mathbf{W}_l' = a\,\mathbf{W}_l, \qquad \mathbf{W}_{l+1}' = a^{-1}\mathbf{W}_{l+1},$$

leaves $f(\mathbf{w})$ unchanged. Hence $\mathcal{L}(\mathbf{w})$ is invariant along this 1-dimensional rescaling orbit, and minimizing $\mathcal{L}_\lambda$ on the orbit reduces to minimizing

$$\mathcal{R}_{l,l+1}(a) \;=\; \lambda_l\|\mathbf{W}_l'\|_F^2 + \lambda_{l+1}\|\mathbf{W}_{l+1}'\|_F^2 \;=\; \lambda_l a^2 x^2 + \lambda_{l+1}a^{-2}y^2,$$

where $x = \|\mathbf{W}_l\|_F$, $y = \|\mathbf{W}_{l+1}\|_F$. The function $\mathcal{R}_{l,l+1}(a)$ is strictly convex in a^2 and coercive, so it has a unique minimizer. Differentiating gives

$$\mathcal{R}_{l,l+1}'(a) = 2\lambda_l a x^2 - 2\lambda_{l+1}a^{-3}y^2 = 0 \iff \lambda_l a^4 x^2 = \lambda_{l+1}y^2.$$

At the minimizer,

$$\frac{\|\mathbf{W}_l'\|_F}{\|\mathbf{W}_{l+1}'\|_F} = \frac{ax}{a^{-1}y} = a^2\frac{x}{y} = \sqrt{\frac{\lambda_{l+1}}{\lambda_l}}.$$

□

Remark 2. The following results are based on the balancedness property induced by the Gaussian prior. For simplicity, we assume $\tau_l \equiv \tau\,\forall l \in [L]$, but note that the following results, Theorem 1 and Corollary 2 could be analogously derived using layer-specific priors.

Corollary 6. *Under the assumptions of Theorem 2, let $\mathbf{a}_{l,j}$ denote the incoming weight vector of hidden unit j in layer l and $\mathbf{v}_{l,j}$ its outgoing weight vector. Following the same logic as in Theorem 2 by transferring Du et al. (2018) to our sampling setup, we have*

$$\mathbb{E}_\pi\big[\mathbf{a}_{l,j}^\top \nabla_{\mathbf{a}_{l,j}}\mathcal{L}(\mathbf{w})\big] \;=\; \mathbb{E}_\pi\big[\mathbf{v}_{l,j}^\top \nabla_{\mathbf{v}_{l,j}}\mathcal{L}(\mathbf{w})\big].$$

Then for every hidden unit j in layer l,

$$\frac{1}{\tau_l^2}\,\mathbb{E}_\pi\big[\|\mathbf{a}_{l,j}\|_2^2\big] \;-\; \frac{1}{\tau_{l+1}^2}\,\mathbb{E}_\pi\big[\|\mathbf{v}_{l,j}\|_2^2\big] \;=\; d_{l,j}^{\text{in}} \;-\; d_{l,j}^{\text{out}},$$

where $d_{l,j}^{\text{in}}$ and $d_{l,j}^{\text{out}}$ are the input and output dimensions of neuron (l, j), respectively.

Remark 3. Summing the identities of Corollary 6 over all neurons j in a given layer h recovers the layer-level balance of Theorem 2.

C.6 Proof of Corollary 6

Proof. Apply Stein's identity (see, e.g., Liu and Wang, 2016) to the test function $f(\mathbf{W}) = \mathbf{a}_{l,j}$ to obtain

$$\mathbb{E}_\pi\big[\mathbf{a}_{l,j}^\top \nabla_{\mathbf{a}_{l,j}} \mathcal{L}(\mathbf{w})\big] \;+\; \frac{1}{\tau_l^2}\,\mathbb{E}_\pi\big[\|\mathbf{a}_{l,j}\|_2^2\big] \;=\; d_{l,j}^{\text{in}}.$$

Similarly, for $f(\mathbf{W}) = \mathbf{v}_{l,j}$,

$$\mathbb{E}_\pi\big[\mathbf{v}_{l,j}^\top \nabla_{\mathbf{v}_{l,j}} \mathcal{L}(\mathbf{w})\big] \;+\; \frac{1}{\tau_{l+1}^2}\,\mathbb{E}_\pi\big[\|\mathbf{v}_{l,j}\|_2^2\big] \;=\; d_{l,j}^{\text{out}}.$$

Subtracting these two equalities and invoking the neuron-wise backprop homogeneity identity cancels the likelihood terms, which yields the claim. □

C.7 Assumptions & Details for Section 4.2: Overparametrization and Prior Conformity

The approximation $\mathbf{H}^\star \approx \mathbf{J}^\top \boldsymbol{\Upsilon} \mathbf{J} + 2\lambda \mathbf{I}$ is the Gauss-Newton approximation of the true Hessian of the negative log-posterior. This approximation is valid under the conditions that $\mathbf{w}^{\text{ref}}$ represents a good model fit, i.e., the gradients of the loss w.r.t. the outputs are close to zero, as well as the network output function being nearly linear w.r.t. the weights $\mathbf{w}$ around $\mathbf{w}^{\text{ref}}$ (Martens, 2020).

The statement $\ker(\mathbf{J}) \cap \mathcal{W} \neq \{\mathbf{0}\}$ follows directly from the definition of an overparametrized layer in Appendix B.

We now show that $(\mathbf{H}^\star)^{-1}\big|_{\ker(\mathbf{J})} = (2\lambda)^{-1}\mathbf{I} = \tau^2\mathbf{I}$, so that the matrix $(\mathbf{H}^\star)^{-1}$, when restricted to the subspace $\ker(\mathbf{J})$, acts as a scalar multiplication by $(2\lambda)^{-1}$.

Let $\boldsymbol{v} \in \ker(\mathbf{J})$ be an arbitrary non-zero vector. By definition, this means $\mathbf{J}\boldsymbol{v} = \mathbf{0}$. Then applying the Hessian approximation $\mathbf{H}^\star$ yields:

$$\mathbf{H}^\star \boldsymbol{v} = (\mathbf{J}^\top \boldsymbol{\Upsilon} \mathbf{J} + 2\lambda \mathbf{I})\boldsymbol{v} = \mathbf{J}^\top \boldsymbol{\Upsilon} \mathbf{J}\boldsymbol{v} + 2\lambda(\mathbf{I}\boldsymbol{v}) = 2\lambda \boldsymbol{v}.$$

We obtain $(\mathbf{H}^\star)^{-1}\boldsymbol{v} = (2\lambda)^{-1}\boldsymbol{v}$ and it follows that $(\mathbf{H}^\star)^{-1}\big|_{\ker(\mathbf{J})} = (2\lambda)^{-1}\mathbf{I} = \tau^2\mathbf{I}$.

D EXPERIMENTAL DETAILS AND FURTHER ANALYSES

D.1 Experimental Setup

Software For sampling-based inference results, we implement our analysis in Python and mainly rely on the `jax` (Bradbury et al., 2018) and `BlackJAX` (Cabezas et al., 2024) libraries. For some comparisons, we used the `posteriors` package (Duffield et al., 2025). Our code is available at `https://github.com/EmanuelSommer/bnn_interplay_priors_overparam`.

Computing Environment The experiments were conducted on two NVIDIA RTX A6000 GPUs and an AMD Ryzen™ Threadripper™ PRO 5000WX/3000WX CPU with 64 cores. For most experiments, 10 chains were sampled in parallel on the CPU, enabling efficient parallelization and allowing multiple experiments to run concurrently. For larger-scale experiments involving thousands of chains, 50 chains were sampled in parallel to maximize resource utilization. For larger CNNs, we used parallel training on GPUs.

Datasets Table 1 summarizes the benchmark datasets utilized in our experiments. For all tabular benchmarks, if not specified otherwise, we use a 70% train, 10% validation, and 20% test split as well as a fully connected model architecture with 3 hidden layers, 16 neurons per layer. For all image classification benchmarks, we use the suggested train/test split and CNNs of varying size.

Performance Evaluation To quantify the quality of the posterior predictive approximation and thus the UQ capabilities of the models we use the log posterior predictive density (LPPD) (Gelman et al., 2014; Wiese et al.,

Table 1: Benchmark datasets overview.

Dataset	Size	Features	Source
Airfoil	1503	5	Dua and Graff (2017)
Bikesharing	17379	13	Fanaee-T (2013)
Concrete	1030	8	Yeh (1998)
Energy	768	8	Tsanas and Xifara (2012)
Ionosphere	351	34	Sigillito et al. (1989)
F(ashion)-MNIST	60000	28x28	Xiao et al. (2017)
CIFAR-10	60000	28x28	Krizhevsky et al. (2009)

2023; Sommer et al., 2025) over a test set $\mathcal{D}_{\text{test}}$, defined as

$$\text{LPPD} = \frac{1}{n_{\text{test}}} \sum_{(\boldsymbol{y}^*, \boldsymbol{x}^*) \in \mathcal{D}_{\text{test}}} \log \left(\frac{1}{K \cdot S} \sum_{k=1}^{K} \sum_{s=1}^{S} p\left(\boldsymbol{y}^* | \boldsymbol{\theta}^{(k,s)}(\boldsymbol{x}^*)\right) \right). \tag{5}$$

Here, K denotes the number of chains, S the number of samples per chain, and $\boldsymbol{\theta}^{(k,s)}$ the parameters from the s-th sample of the k-th chain. Intuitively, the LPPD quantifies how well the predictive distribution aligns with the observed labels, with higher values indicating higher density coverage, i.e., improved UQ performance.

In addition, we employ the root mean squared error (RMSE) for regression and the accuracy for classification tasks to check for the accuracy of point predictions.

D.2 Experimental Details for Figure 5

Below, let data be $(x_i, y_i)_{i=1}^n$, noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with known σ^2, and

$$f(x) = a\,\text{ReLU}(bx) + c\,\text{ReLU}(dx), \qquad a, b, c, d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Define $\phi_{1,i}(b) = \text{ReLU}(bx_i)$, $\phi_{2,i}(d) = \text{ReLU}(dx_i)$, the design

$$\Phi(b,d) = \begin{bmatrix} \phi_{1,1}(b) & \phi_{2,1}(d) \\ \vdots & \vdots \\ \phi_{1,n}(b) & \phi_{2,n}(d) \end{bmatrix} \in \mathbb{R}^{n \times 2}, \quad \theta = \begin{bmatrix} a \\ c \end{bmatrix}, \quad y = (y_1, \ldots, y_n)^\top.$$

The likelihood and prior are given by

$$p(y \mid \theta, b, d) \;=\; \mathcal{N}\big(y;\, \Phi(b,d)\theta,\, \sigma^2 I_n\big), \qquad p(\theta, b, d) \;=\; \mathcal{N}(\theta; 0, I_2)\,\mathcal{N}(b; 0, 1)\,\mathcal{N}(d; 0, 1).$$

Then, the kernel of the joint posterior is

$$p(\theta, b, d \mid y) \;\propto\; \exp\Big(-\tfrac{1}{2\sigma^2}\|y - \Phi(b,d)\theta\|^2 - \tfrac{1}{2}\|\theta\|^2 - \tfrac{1}{2}b^2 - \tfrac{1}{2}d^2 \Big).$$

Because the model is linear in (a, c) given (b, d), we get conjugacy for $\theta \mid b, d, y$ and a conditional posterior for a, c given b, d given by

$$\theta \mid b, d, y \;\sim\; \mathcal{N}\big(\mu_{ac}(b,d),\, \Sigma_{ac}(b,d)\big),$$

with

$$\Sigma_{ac}(b,d) \;=\; \Big(I_2 + \tfrac{1}{\sigma^2}\Phi(b,d)^\top \Phi(b,d)\Big)^{-1}, \qquad \mu_{ac}(b,d) \;=\; \Sigma_{ac}(b,d)\,\tfrac{1}{\sigma^2}\Phi(b,d)^\top y.$$

Therefore, the marginal likelihood for b, d and their posterior (by integrating out (a, c)) is

$$p(y \mid b, d) \;=\; \mathcal{N}\big(y;\, 0,\, \sigma^2 I_n + \Phi(b,d)\Phi(b,d)^\top\big),$$

hence

$$p(b, d \mid y) \;\propto\; \mathcal{N}\big(y;\, 0,\, \sigma^2 I_n + \Phi(b,d)\Phi(b,d)^\top\big)\,\mathcal{N}(b; 0, 1)\,\mathcal{N}(d; 0, 1).$$

This pair has no closed form. It can be explored by MAP optimization of $\log p(b, d \mid y)$ or by MCMC. Further, the full posterior factorization is given by

$$p(a, b, c, d \mid y) \;=\; p(a, c \mid b, d, y)\; p(b, d \mid y),$$

with $p(a, c \mid b, d, y)$ Gaussian as above and $p(b, d \mid y)$ given up to a normalizing constant.

D.3 The Role of the Bias

Extending the exposition about the role of the bias in Section 4.2, we provide an illustrative marginal bias distribution across layers of a ReLU network in Figure 11.

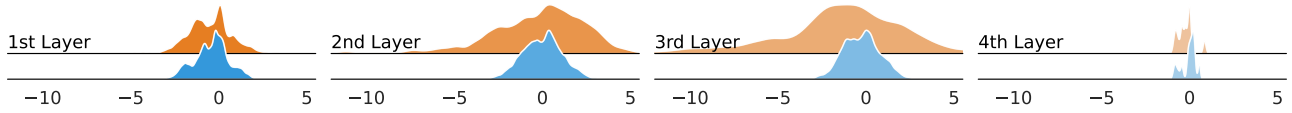


Figure 11: Empirical marginal distribution of the bias of a four-hidden-layer MLP with a normal prior on the bias as well as a uniform prior on the bias.

In frequentist training of ReLU neural networks, it is common practice not to regularize the bias terms, as they primarily shift the activation thresholds and are not prone to uncontrolled growth like weights. In contrast, BNNs explicitly place priors over all parameters, including biases, which induces a form of regularization. Figure 11 illustrates this difference: when a uniform prior is imposed, the distributions appear wider, reflecting greater posterior uncertainty, whereas introducing isotropic Gaussian priors yields narrower marginals. Importantly, both cases exhibit similar shapes in their high-density regions near the mode, indicating that regularization primarily reduces tail spread without altering the local geometry of the posterior around the most probable values.

D.4 Experimental Details of Figure 9

For each posterior sample, we identify clusters of neurons with similar activation patterns. We compute the cosine similarity matrix $\tilde{\boldsymbol{\Xi}}^\top \tilde{\boldsymbol{\Xi}}$ of the standardized feature columns (excluding dead or constant units) and form clusters as connected components of the thresholded similarity graph, where the threshold is calibrated via a Bonferroni correction under the null hypothesis of independent activation columns. For each cluster with indices j, we compute

$$S = \min_{\substack{\boldsymbol{v}^\top \mathbf{1}=0 \\ \|\boldsymbol{v}\|_2=1}} \|\boldsymbol{\Xi}_{:,j}\boldsymbol{v}\|_2.$$

Here $\boldsymbol{v}$ represents a zero-sum reweighting of outgoing weights within the cluster. We use a zero-sum reweighting to redistribute weight among neurons within a given cluster. Small S indicates that such reweightings barely affect the network output, which is a likelihood-flat direction. We use unstandardized features $\boldsymbol{\Xi}$ (not $\tilde{\boldsymbol{\Xi}}$) for this computation to ensure S reflects the magnitude of the actual output change under reweighting, not just the collinearity of activations.

D.5 Exploring the Limits of BDEs

We extend the analysis of Sommer et al. (2024), who only consider 12 to 10k chains of 1k samples each. We also use a more than twice as large fully-connected neural network (4 hidden layers of 16 neurons each) to perform distributional regression. For this, we use the recently proposed MILE approach (Sommer et al., 2025) and configure it exactly as suggested by the authors. Due to the immense computational load of sampling this amount of chains and also evaluating the posterior samples (compressing the samples roughly amounts to 100GB for a single experiment), we carefully select four benchmark datasets, namely, `airfoil` and `bikesharing` for distributional regression, `ionosphere` for tabular classification, and `Fashion-MNIST` for image classification. For the latter, we do not use MILE but rather employ scale-adapted SGHMC (Springenberg et al., 2016) for computational efficiency. In our analysis, we focus on two major aspects. First, we analyze how the performance of the model develops when adding chains to the Bayesian Model Average (BMA). Second, in the spirit of Sommer et al. (2024) we take a closer look at bivariate margins of the empirical posterior derived from SAI.

Performance Evolution The cumulative performance, which we—focusing on UQ quality—measure with the LPPD, of adding chains to the BMA obviously depends on the order in which chains are added. Thus, we consider 5 different orderings and report means and standard deviations of the cumulative LPPD over chains in Figure 12 and 13. The results suggest that with even a rather small number of chains, the performance saturates quite fast, but slowly increases further until exhibiting a very strong performance for 10k chains. Parallelizing 10-20 chains on modern hardware is very easy and comes with no considerable cost overhead over single-chain

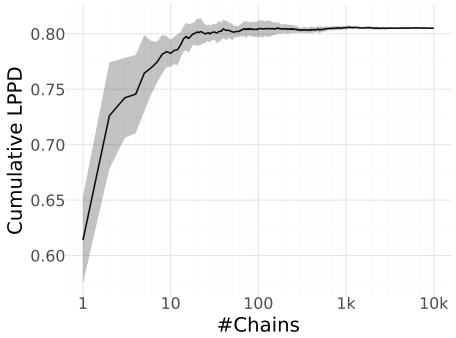
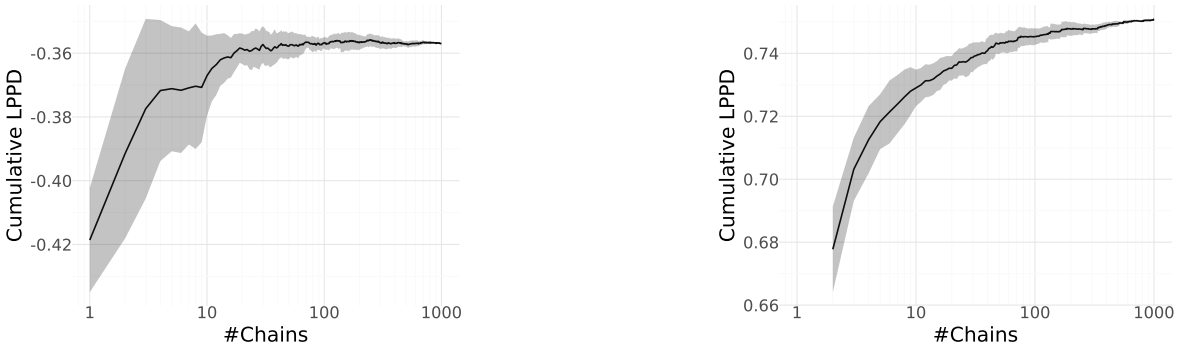


Figure 12: Cumulative LPPD over the number of chains (standard deviation across 5 random chain orderings) on the `airfoil` dataset.

sampling. This also has very positive implications on memory requirements and inference time, rendering the approach practically feasible.



(a) `CNNv1` on Fashion-MNIST with SG-MCMC.

(b) FCN on `bikesharing` with MILE.

Figure 13: Cumulative LPPD over the number of chains (standard deviation across 5 random chain orderings) for the same `CNNv1` setting on Fashion-MNIST as in Figure 14 (left) and the same fully-connected neural network (FCN) as the one in Figure 12 but fitted on the larger `bikesharing` dataset. As a reference, the DE with 1k members only achieves an LPPD of -0.3627 for `CNNv1` and 0.4935 for the FCN. Also, the predictive accuracy of the BDE (1k) exceeds that of the DE (1k) in both settings (left: accuracy $0.8715 > 0.8706$, right: RMSE $0.2258 < 0.2508$).

Underparametrized Model In order to contrast the bivariate marginal posterior densities of the above-considered overparametrized models with previously analyzed underparametrized models, we also considered the small f_1 architecture of Wiese et al. (2023) and just as in their work, fitted it on the `airfoil` dataset, but now with 8M posterior samples and the MILE sampler. Figure 17 displays the obtained marginal densities and shows clearly that in the underparametrized setting, very distinct symmetry patterns emerge in the margins of the sampled posterior. One of these marginal plots is also featured in Figure 1.

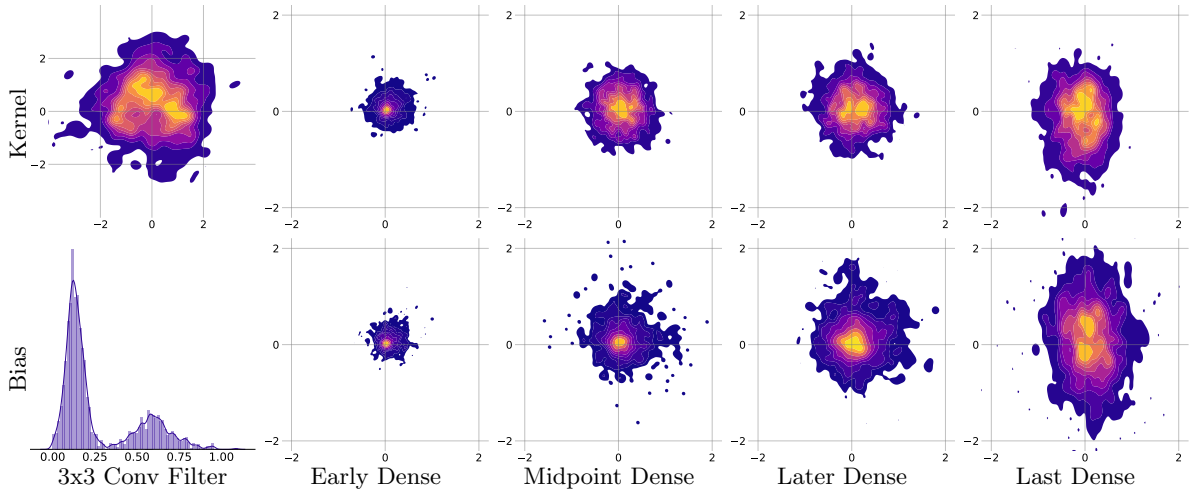


Figure 14: (Bivariate) marginal posterior densities of a small 4-hidden layer convolutional BNN fitted on the Fashion-MNIST dataset (`CNNv1` of Sommer et al., 2025). The grid visualizes the empirical densities of 1M posterior samples obtained from 1k independent chains. The rows and columns (Conv, three hidden, and output weights) display representative densities of randomly chosen weights of the network. We employ scale-adapted SGHMC (Springenberg et al., 2016) for sampling, use an ensemble of 10 with 10k warmup steps, 10k sampling steps, thinning of 10, step size 0.001, momentum decay 0.05, batch size 256, and standard normal isotropic priors.

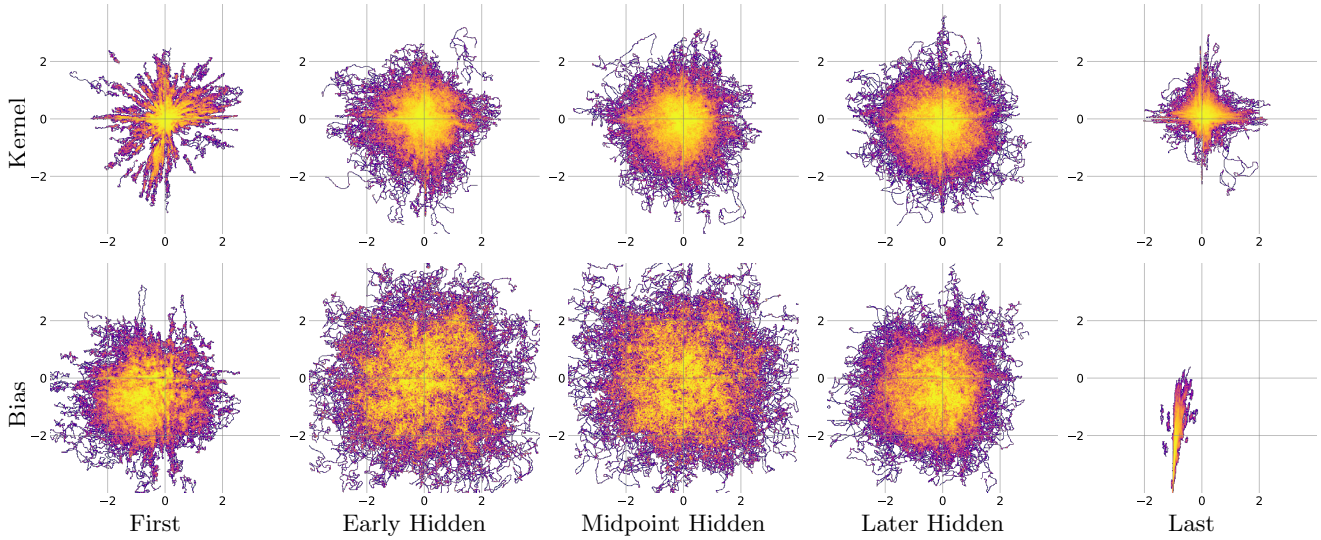


Figure 15: (Bivariate) marginal posterior densities of a 4-hidden layer fully-connected BNN fitted on the `bikesharing` dataset (regression task, and same architecture as in Figure 6). The grid visualizes the empirical densities of 1M posterior samples obtained from 1k independent chains. The rows and columns (input, three hidden, and output weights) display representative densities of randomly chosen weights of the network.

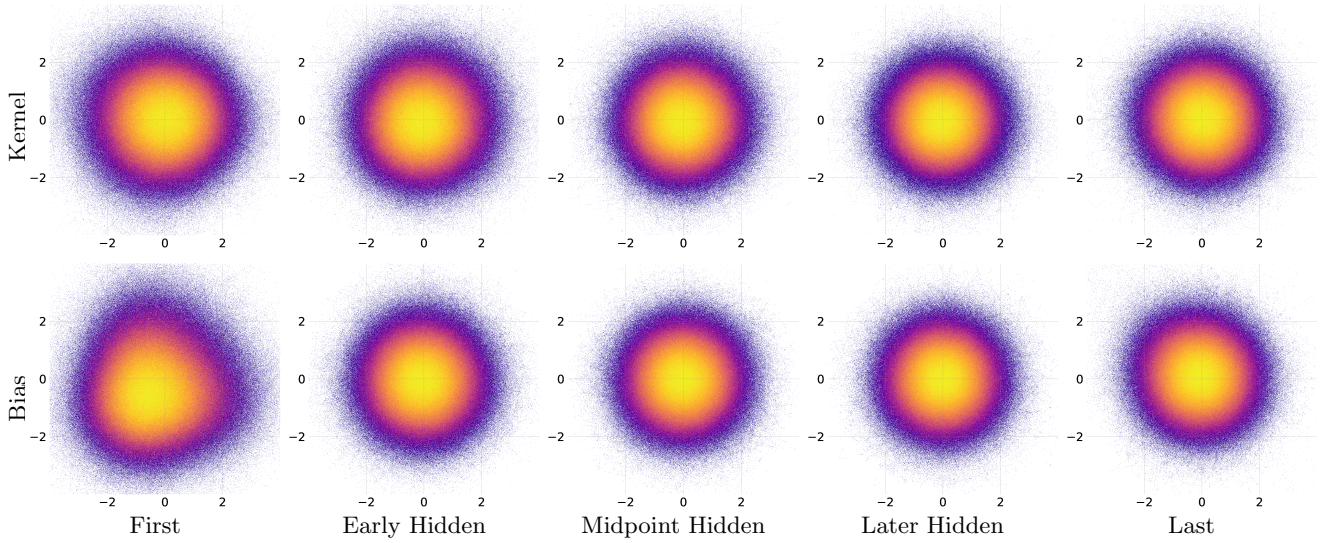


Figure 16: (Bivariate) marginal posterior densities of a 4-hidden layer fully-connected BNN fitted on the `ionosphere` dataset (binary classification task, and same architecture as in Figure 6). The grid visualizes the empirical densities of 8M posterior samples obtained from 8k independent chains. The rows and columns (input, three hidden, and output weights) display representative densities of randomly chosen weights of the network.

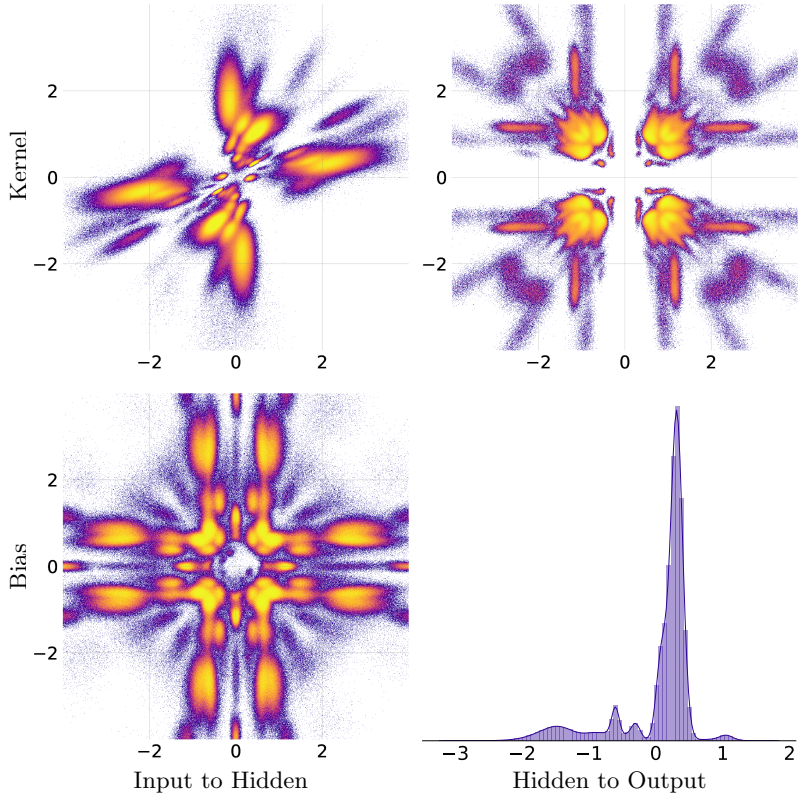


Figure 17: (Bivariate) marginal posterior densities of the underparametrized one-hidden layer fully-connected BNN fitted on the `airfoil` dataset (following the small f_1 architecture of Wiese et al. (2023)). The grid visualizes the empirical densities of 8M posterior samples obtained from 8k independent chains. The rows and columns (input and output weights) display representative densities of randomly chosen weights of the network.

E BENCHMARKS

In this section, we provide benchmarks of sampling-based inference, both tabular regression with MLPs as well as image classification with CNN architectures. Overall, we find that sampling-based inference outperforms approximate Bayesian inference methods, delivering superior predictive performance and uncertainty quantification. We first describe the experimental setup in detail before showing the respective results.

UCI Benchmark For the UCI benchmark presented in Table 2, we fit classical mean regression to the different tasks corresponding to the datasets described in Table 1. In the process, we always use a fully-connected feed-forward neural network with 3 hidden layers of size 16 each, resulting in about 700 total model parameters. If sampling from the posterior is done, we use 1000 samples per ensemble member (chain). We describe the configuration of the employed methods one by one:

- For the **Laplace approximation** (LA), we utilize a JAX-based implementation to first train MAP solutions using the Adam optimizer with decoupled weight decay (Loshchilov and Hutter, 2019) for 10000 epochs with a learning rate of 0.005 to then carry out last-layer LA with a generalized Gauss Newton Hessian approximation and closed-form predictive approximation as detailed in Daxberger et al. (2021). The variance of the predictive distribution is calculated according to Daxberger et al. (2021) with a small additional noise variance term.
- For **mean-field variational inference** (MFVI), we utilize a Gaussian posterior approximation with independence assumption. We optimize the evidence lower bound (ELBO) for 5000 epochs with the Adam optimizer and a learning rate of 0.005. The variance of the predictive distribution is calculated as the variance over the predictions made with 100 samples from the fitted approximate posterior with a small additional observation noise term.
- As the recently proposed **Microcanonical Langevin Ensemble** (MILE) approach provides both an optimized **Deep Ensemble** (DE) and a **Bayesian Deep Ensemble** (BDE), we follow the suggested setup of Sommer et al. (2025), i.e., the DE is optimized with the Adam optimizer with decoupled weight decay with (memberwise) early stopping and the sampling then uses the proposed auto-tuning strategy of MILE comprising 50k steps before then providing 1k samples (after the thinning of 10k samples).

Each method is evaluated using three distinct train-test splits to assess the robustness of its performance.

Table 2: Mean RMSE (↓) and LPPD (↑) results (± standard deviation across 3 train-test splits) for a 3 hidden-layer fully-connected neural network on regression tasks. Numbers in brackets indicate the number of ensemble members/chains.

	Dataset	Laplace	MFVI	DE (10)	MCMC
LPPD	Airfoil	-1.056 ± 0.003	-0.975 ± 0.004	-0.293 ± 0.096	$\mathbf{0.016 \pm 0.293}$
	Bikesharing	-1.046 ± 0.001	-0.990 ± 0.005	-0.223 ± 0.181	$\mathbf{-0.060 \pm 0.096}$
	Concrete	-1.131 ± 0.036	-0.998 ± 0.007	-0.510 ± 0.189	$\mathbf{0.042 \pm 0.056}$
	Energy	-1.046 ± 0.004	-0.945 ± 0.002	1.561 ± 0.101	$\mathbf{1.947 \pm 0.047}$
RMSE	Airfoil	0.237 ± 0.013	0.276 ± 0.009	0.269 ± 0.016	$\mathbf{0.184 \pm 0.016}$
	Bikesharing	0.252 ± 0.006	0.318 ± 0.018	$\mathbf{0.253 \pm 0.015}$	0.262 ± 0.018
	Concrete	0.482 ± 0.100	0.350 ± 0.025	0.297 ± 0.032	$\mathbf{0.270 \pm 0.034}$
	Energy	0.065 ± 0.008	0.126 ± 0.007	0.050 ± 0.001	$\mathbf{0.041 \pm 0.003}$

Image Classification on CIFAR-10 Using SGHMC We extend the tabular UCI benchmarks of Table 2 to an image classification task on CIFAR-10 using a small ResNet with 76106 parameters. We employ scale-adapted SGHMC (Springenberg et al., 2016) for sampling, use an ensemble of 10 with 5k warmup steps, 25k sampling steps, thinning of 250, step size 0.001, momentum decay 0.05, batch size 256, and standard normal isotropic priors. Notably, compared to DE optimization, **sampling required only 25% additional compute**. Table 3 displays both predictive and UQ performance as well as wallclock time in comparison with optimization-based approximate inference methods implemented via the novel `posteriors` package (Duffield et al., 2025). Again, the sampling approach provides the best predictive UQ and precision at a negligible cost in wall-clock time. Due to the non-trivial optimization of some VI methods (for example, caused by noisy gradients), sampling can be faster and more robust to fit even in these larger-scale settings.

Table 3: Image classification task on CIFAR-10 using a small ResNet with 76106 parameters. We employ scale-adapted SGHMC for sampling, use an ensemble of 10 with 5k warmup steps, 25k sampling steps, thinning of 250, step size 0.001, momentum decay 0.05, batch size 256, and standard normal isotropic priors. Mean accuracy (↑) and LPPD (↑) results (± standard deviation) for CIFAR10 classification are reported. Numbers in brackets indicate ensemble members/chains.

	Laplace	**MFVI**	**DE (10)**	**BDE (10)**
Accuracy (↑)	0.7851 ± 0.0067	0.7133 ± 0.0102	0.8222 ± 0.0013	$\mathbf{0.8273 \pm 0.0020}$
LPPD (↑)	-1.6731 ± 0.0571	-0.8410 ± 0.0193	-0.5341 ± 0.0050	$\mathbf{-0.5149 \pm 0.0030}$
Wallclock Time in Minutes (↓)	12.92 ± 0.41	15.78 ± 0.27	$\mathbf{11.59 \pm 0.34}$	14.41 ± 0.02

Computational Cost Comparison Between DE and BDE To assess the trade-off between computational cost and predictive performance, we compare Deep Ensembles and Bayesian Deep Ensembles on the `airfoil` and `ionosphere` datasets. While BDE for the same number of members is significantly more expensive—taking 12.22 and 8.25 times more wall-clock time per member on `airfoil` and `ionosphere`, respectively—it consistently outperforms DE in terms of LPPD. The results of the experiments are given in Table 4. For instance, BDE(1) on `ionosphere` achieves an LPPD of -0.1544 ± 0.0119 compared to DE(8000)'s -0.2923, despite the latter using nearly 1000 times more compute. For the `airfoil` dataset to ensure a fair comparison in terms of runtime, we compare BDE(10) to DE(125) over 3 replications and again observe superior sampling performance for the same computational effort.

Table 4: Computational cost and predictive performance comparison between DE and BDE on the `airfoil` and `ionosphere` datasets.

Model	**Ensemble Members**	**Dataset**	**LPPD** (↑)	**RMSE/Accuracy**
DE	125	`airfoil`	0.0285 ± 0.0266	0.2870 ± 0.0240
BDE	10	`airfoil`	$\mathbf{0.7660 \pm 0.0075}$	$\mathbf{0.1415 \pm 0.0079}$
DE	8000	`ionosphere`	-0.2923	0.9154
BDE	1	`ionosphere`	$\mathbf{-0.1544}$	**0.9436**