A Compact Representation for Bayesian Neural Networks By Removing Permutation Symmetry

Tim Z. Xiao[1,2] Weiyang Liu[3,4] Robert Bamler[1]

[1]University of Tübingen [2]IMPRS-IS [3]University of Cambridge
[4]Max Planck Institute for Intelligent Systems, Tübingen

1. Background: Permutation Symmetry in Neural Networks

- For a neural network, e.g.,

$$x \mapsto \sigma_2(\omega_2\,\sigma_1(\omega_1\,x + b_1) + b_2)$$

  we can apply permutation P s.t.

$$\omega_1' := P\,\omega_1,\; b_1' := P\,b_1,\; \omega_2' := \omega_2\,P^{-1},$$

  which does not change the function.

- Interpolation $\mathbf{W}_\lambda := (\lambda - 1)\mathbf{W}_0 + \lambda \mathbf{W}_1$ between two trained networks with weights $\mathbf{W}_0, \mathbf{W}_1$, s.t. $\mathbf{W} = \{\omega_1, \omega_2, b_1, b_2\}$, introduces a loss barrier (top plot).

- Rebasin (Ainsworth, 2023) removes the loss barriers (bottom plot)

Paper PDF

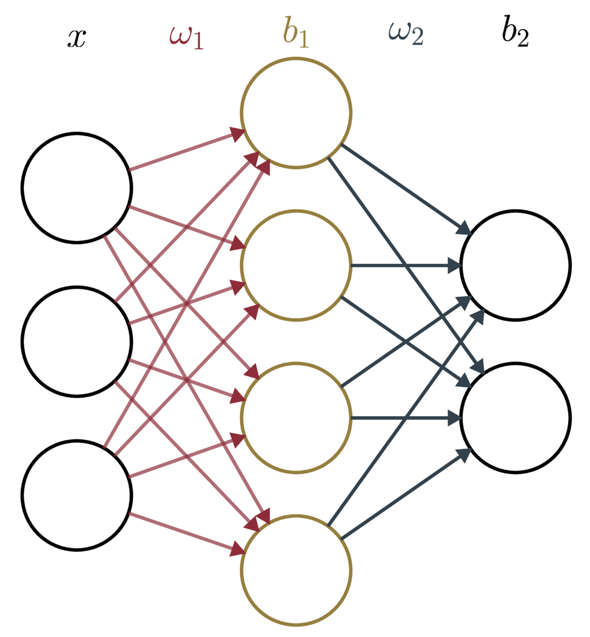


Figure 1: Permutation invariance for neurons in the same layer.

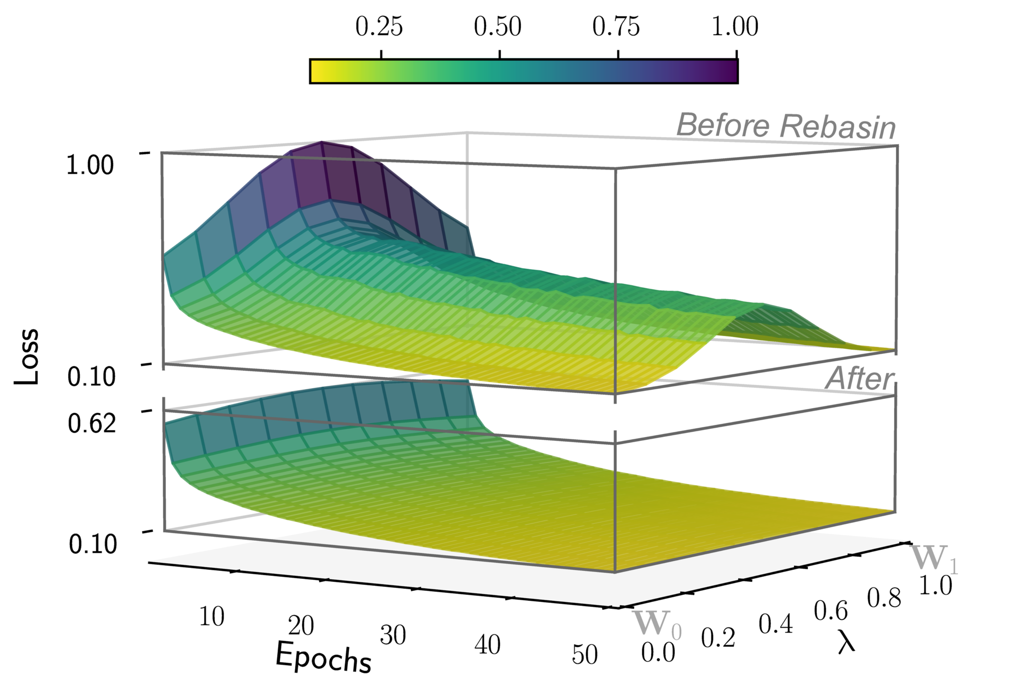


Figure 2: Training dynamics for models with $\mathbf{W}_0$ and $\mathbf{W}_1$, and their interpolations $\mathbf{W}_\lambda$.

2. Quantifying Permutations in Weight Space by Number of Transpositions

- **Number of Transpositions (NoTs)** – Measuring the magnitude of permutation with the minimal number of pairwise swaps (i.e., transpositions). We can then meaningfully quantify weight-space distances by a pair $\left(||\mathbf{W}_0 - P\mathbf{W}_1||_2^2, \mathrm{NoT}(P)\right)$.

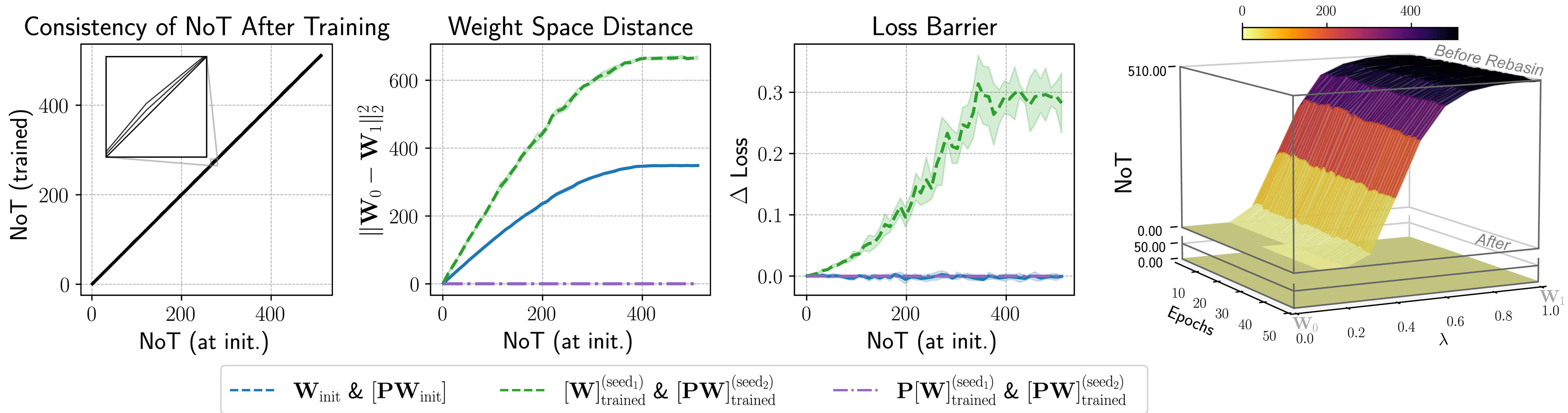


Figure 3: Left three: effect of permuting initial weights by different Number of Transpositions (NoT) on NoT after training, weight-space distance, and loss barrier (shaded regions: $\pm 1\sigma$ over 5 runs). Right: NoT changes monotonically along the interpolation $\mathbf{W}_\lambda$ between two models $\mathbf{W}_0$ and $\mathbf{W}_1$.

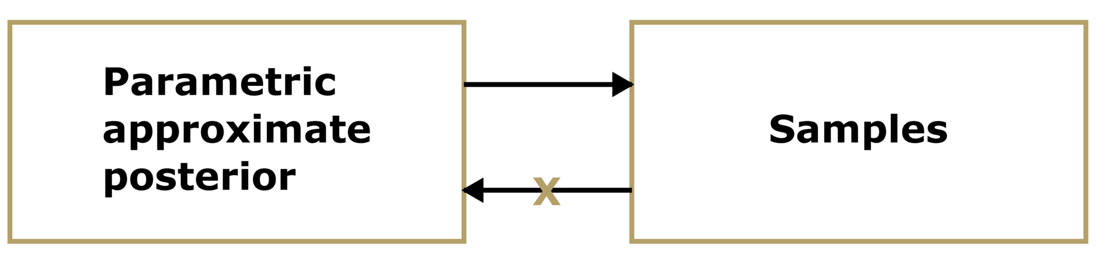
3. A Unifying Compact Representation for Bayesian Neural Networks

Problem:

- In BNNs, instead of $\arg\max_{\mathbf{W}}\; p(\mathcal{D}\,|\,\mathbf{W})$, we want $p(\mathbf{W}\,|\,\mathcal{D}) = \frac{p(\mathbf{W})p(\mathcal{D}\,|\,\mathbf{W})}{p(\mathcal{D})}$.
- The predictive distribution $p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*, \mathcal{D}) = \int p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*, \mathbf{W})\; p(\mathbf{W}\,|\,\mathcal{D})\,\mathrm{d}\mathbf{W}$.
- Two categories of representations for $p(\mathbf{W}\,|\,\mathcal{D})$:
  1. **Parametric methods**, e.g., variational inference (VI) and Laplace approximation.
  2. **Sampling methods**, e.g., Hamiltonian Monte Carlo (HMC), deep ensembles.
- **There is no unifying representation!**

Our Proposed Solution:

- **Conjecture:** *the quasi-convexity conjecture from prior works (Ainsworth, 2023) suggests that the posterior is close to unimodal once we remove the permutation degrees of freedom.*
- **Unify the representations:**
  1. Rebase into one basin
  2. Fit a simple unimodel distribution for $p(\mathbf{W}\,|\,\mathcal{D})$, e.g., Gaussian with the rebased sample mean and variance.



(a) $p(\mathbf{W}\,|\,\mathcal{D})$ has many modes $\rightarrow$ Fitting a parametric model to the samples is difficult.

(b) Rebasin makes it easier.

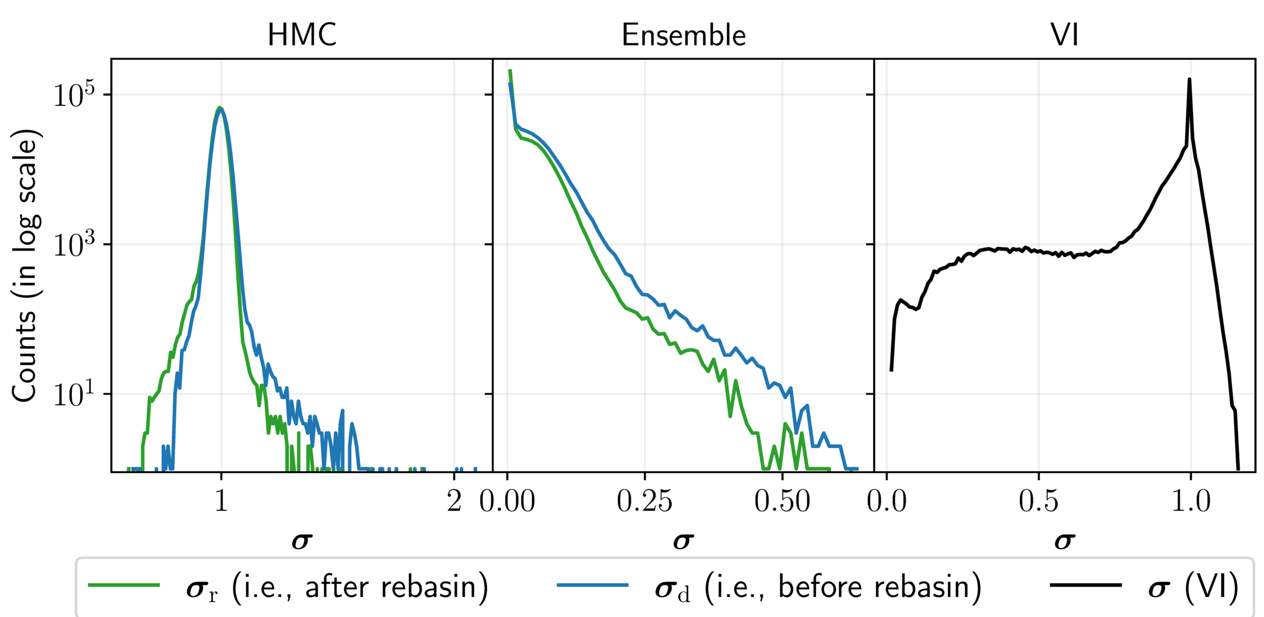
Evaluations:

Table 1: Performance of different BNNs (q_{d}: before rebasin; q_{r}: after rebasin) on their agreement (Equation (1)) and total variation (TV; Equation (2)) to HMC samples, and on their test set accuracy.

	HMC			Ensemble			VI
	Sample	$q_{\mathrm{d}}(\mathbf{W})$	$q_{\mathrm{r}}(\mathbf{W})$	Sample	$q_{\mathrm{d}}(\mathbf{W})$	$q_{\mathrm{r}}(\mathbf{W})$	$q(\mathbf{W})$
(↑) Agreement with HMC samples	1.	0.1212	0.8249	0.9931	0.5239	0.9868	0.9885
(↓) TV to HMC samples	0.	0.8641	0.6570	0.0229	0.7210	0.0495	0.0235
Test Accuracy (%) of Samples	98.43	11.11	82.34	98.66	52.25	97.72	98.11
Test Accuracy (%) of $\boldsymbol{\mu}_{\mathrm{d}}$ and $\boldsymbol{\mu}_{\mathrm{r}}$	N/A	28.06	92.25	N/A	86.40	97.97	98.04

$$\mathrm{Agree.}(p, p_{\mathrm{HMC}}) = \frac{1}{|\mathcal{D}_{\mathrm{test}}|}\sum_{\boldsymbol{x}^*\in\mathcal{D}_{\mathrm{test}}} I\Big[\arg\max_{\boldsymbol{y}^*} p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*, \mathcal{D}) = \arg\max_{\boldsymbol{y}^*} p_{\mathrm{HMC}}(\boldsymbol{y}^*\,|\,\boldsymbol{x}_i^*, \mathcal{D})\Big]; \quad (1)$$

$$\mathrm{TV}(p, p_{\mathrm{HMC}}) = \frac{1}{|\mathcal{D}_{\mathrm{test}}|}\sum_{\boldsymbol{x}^*\in\mathcal{D}_{\mathrm{test}}}\frac{1}{2}\sum_{\boldsymbol{y}^*}\big|p(\boldsymbol{y}^*\,|\,\boldsymbol{x}_i^*, \mathcal{D}) - p_{\mathrm{HMC}}(\boldsymbol{y}^*\,|\,\boldsymbol{x}_i^*, \mathcal{D})\big|. \quad (2)$$



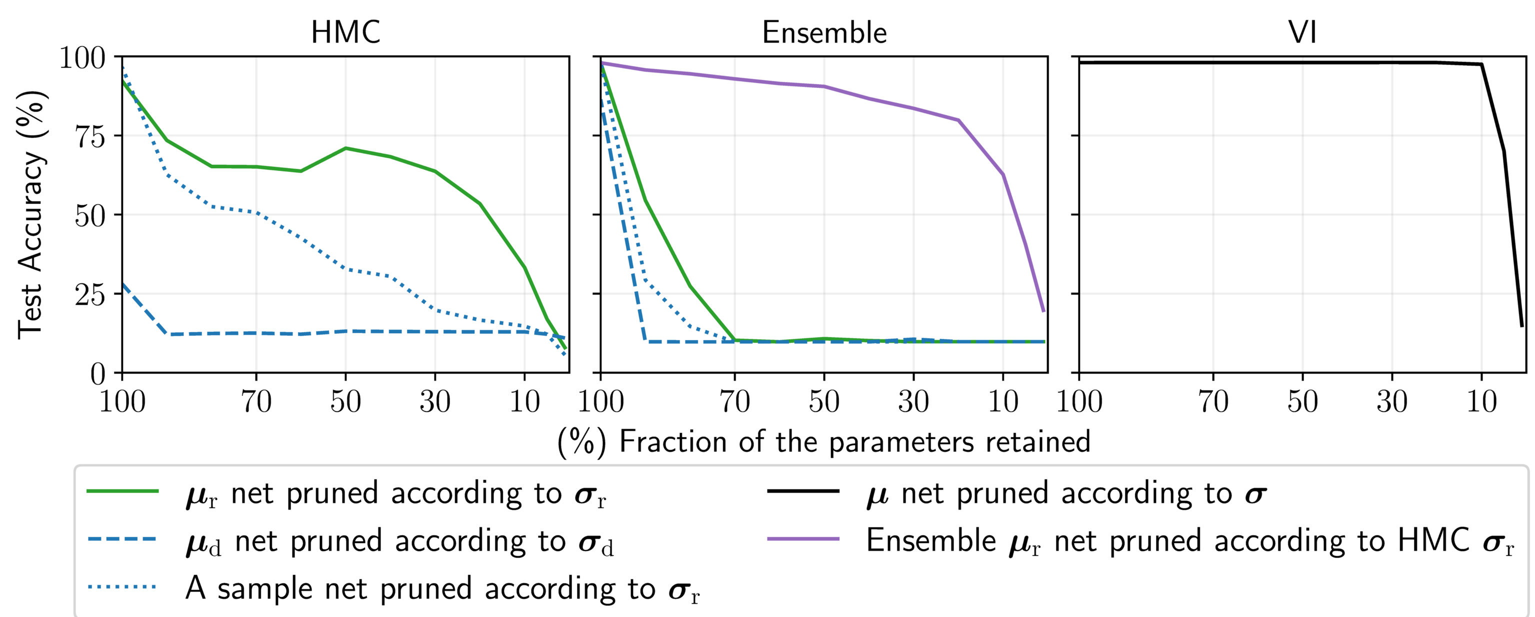


Figure 5: Left: histograms of the standard deviation σ of weights before (σ_{d}) and after (σ_{r}) rebasin. Right: test accuracy vs. various levels of weight pruning (retaining only weights with lowest σ).

Presentation at NeurIPS 2023 Workshop on Unifying Representations in Neural Models (UniReps)