Equivariance and Augmentation for Bayesian Neural Networks

Miaowen Dong[1] Axel Flinth[*2] Jan E. Gerken[*1]

[1]Chalmers University of Technology
[2]Umeå University
[*]Equal contribution

GeUmetric Deep Learning Workshop 2026
August 19, 2026

CHALMERS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Two roads to equivariance

Many tasks have known symmetries: medical imaging, molecules, physics.

Equivariant architectures

- Symmetry built in layer-by-layer
- Exact
- Specialized architectures

Data augmentation

- Train on transformed data
- Any architecture
- Only approximate

Known result

Averaging the predictions of independently initialized networks trained on augmented data (a deep ensemble) becomes exactly equivariant in the infinite-ensemble limit (Gerken and Kessel, 2024; Nordenfors and Flinth, 2025).

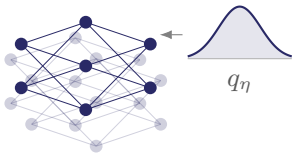
A cheaper "equivariance in expectation"

Idea: one Bayesian neural network instead of an ensemble, and the **posterior predictive** instead of the average over members.

M independent networks

$$\frac{1}{M} \sum_{m=1}^{M} f(x; \theta_m)$$

one network, weights $\sim q_\eta$



$$\frac{1}{T} \sum_{t=1}^{T} f(x; \theta^{(t)})$$

Background: BNNs via variational inference

Network $f(\cdot\,;\theta)$, prior $p_0(\theta)$, dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N_0}$.

- The posterior $p(\theta \mid \mathcal{D})$ is intractable: replace it by the closest q_η in a tractable family $\mathcal{Q} = \{q_\eta\}$, e.g. mean-field Gaussian $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$
- Fit η by minimizing the β-weighted ELBO loss

$$\mathcal{L}_\beta(\eta) = \underbrace{\tfrac{1}{N}\mathbb{E}_{q_\eta}[-\log p(\mathcal{D}_{\mathrm{aug}} \mid \theta)]}_{\text{explain the data}} + \beta \underbrace{D_{\mathrm{KL}}(q_\eta \,\|\, p_0)}_{\text{stay near the prior}}$$

- Predict with the posterior predictive, via Monte Carlo

$$F_\eta(x) = \mathbb{E}_{\theta\sim q_\eta}[f(x;\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(x;\theta^{(t)})$$

Background: symmetry acting on data and parameters

- Augmentation: finite group G, $\mathcal{D}_{\text{aug}} = \{(\rho_{\mathcal{X}}(g)x_i, \rho_{\mathcal{Y}}(g)y_i) \mid g \in G\}$
- G acts compatibly on inputs, outputs, and parameters:

$$f(\rho_{\mathcal{X}}(g)x; \theta) = \rho_{\mathcal{Y}}(g) f\big(x; \rho_{\Theta}(g)^{-1}\theta\big)$$

- Push-forward of a weight distribution: $\mathcal{T}_g \# q(B) = q(\rho(g)^{-1}B)$

Equivariance of the predictive map

F_η is G_ε-equivariant if $\Delta_F^{\text{eq}}(\eta) = \mathbb{E}_{x,g}\big[\|F_\eta(gx) - gF_\eta(x)\|^2\big] \leq \varepsilon$.

Central question: does augmented training drive $\Delta_F^{\text{eq}}(\eta) \to 0$, and how fast?

Step 1: When does G stay inside the variational family?

Exponential families: $q_\eta(\theta) = h(\theta)\exp(\eta^\top T(\theta) - A(\eta))$, with base measure h and sufficient statistic T.

Theorem (Closedness under push-forward)

$\mathcal{Q}$ is closed under $\mathcal{T}_g\#$ iff for every g there are M_g, d_g, b_g, c_g with

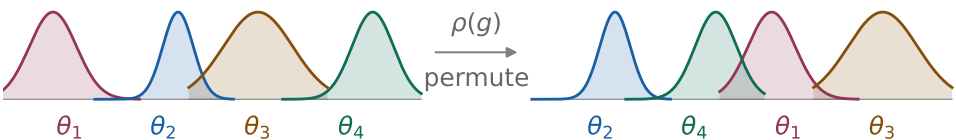
$$T(\rho(g)^{-1}\theta) = M_g T(\theta) + d_g,$$
$$h(\rho(g)^{-1}\theta)\,\big|\det D(\rho(g)^{-1})(\theta)\big| = h(\theta)\,\exp(b_g^\top T(\theta) + c_g).$$

Then $\mathcal{T}_g\#q_\eta = q_{\phi_g(\eta)}$ with the **affine action** $\phi_g(\eta) = M_g^\top \eta + b_g$ on natural parameters, the same b_g as above.

Example: the everyday BNN setup qualifies

Mean-field Gaussian $q_\eta = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$, group acting by permuting the weights.

$q_\eta = \mathcal{N}(\mu,\ \mathrm{diag}(\sigma^2))$ $\qquad$ $\mathcal{T}_g \# q_\eta = \mathcal{N}(R_g\mu,\ \mathrm{diag}(R_g\sigma^2))$

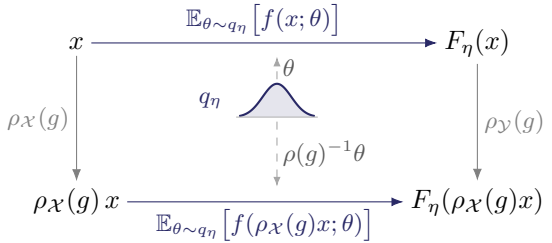


same family, entries reordered: ϕ_g is a permutation of η

Step 2: Symmetric parameters $\Rightarrow$ equivariant predictions

Proposition

$\eta \in H_G := \{\eta \mid \phi_g(\eta) = \eta, \forall g \in G\} \;\Rightarrow\; F_\eta$ is exactly G-equivariant.



$$\mathbb{E}_{\theta\sim q_\eta}\big[f(\rho_{\mathcal{X}}(g)x;\theta)\big] \overset{(1)}{=} \rho_{\mathcal{Y}}(g)\,\mathbb{E}_{\theta\sim q_\eta}\big[f(x;\rho(g)^{-1}\theta)\big] \overset{(2)}{=} \rho_{\mathcal{Y}}(g)\,\mathbb{E}_{\theta\sim q_\eta}\big[f(x;\theta)\big]$$

(1) compatible actions move g into the *weights*; (2) $\eta \in H_G$: same law, hence the same average.

Where augmentation enters: an invariant likelihood

The **likelihood** says how well weights θ explain the data:

$$L_{\text{aug}}(\theta) \;=\; p(\mathcal{D}_{\text{aug}} \mid \theta) \;=\; \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p(y \mid f(x;\theta))$$

- $\mathcal{D}_{\text{aug}}$ is closed under G, and the observation noise is G-invariant
- $\Rightarrow\; L_{\text{aug}}(\rho(g)^{-1}\theta) = L_{\text{aug}}(\theta)$: **augmentation makes the likelihood invariant**, whatever the architecture

This is the *only* place where augmentation enters the theory.

Main theorem: augmented training preserves H_G

Assumptions: invariant likelihood (from augmentation), invariant prior, volume-preserving (e.g. orthogonal) action, closed family.

Theorem (Equivariance of VI)

(i) $\mathcal{L}_\beta(\phi_g(\eta)) = \mathcal{L}_\beta(\eta)$;

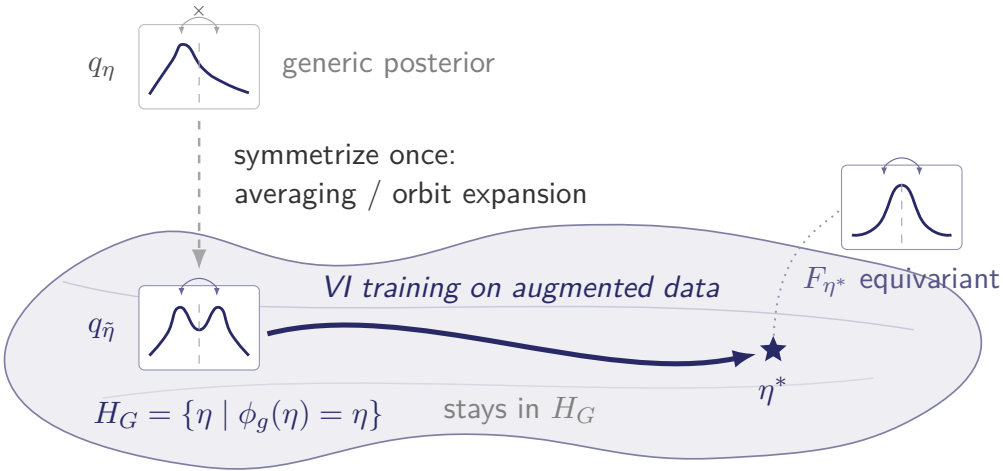
(ii) $\nabla_\eta \mathcal{L}_\beta(\phi_g(\eta)) = M_g^{-1} \nabla_\eta \mathcal{L}_\beta(\eta)$, so for orthogonal M_g the update $U(\eta) = \eta - \alpha \nabla_\eta \mathcal{L}_\beta(\eta)$ commutes with ϕ_g: $U(\phi_g(\eta)) = \phi_g(U(\eta))$;

(iii) $\eta^{(0)} \in H_G \Rightarrow \eta^{(t)} \in H_G$ for all $t \geq 0$;

(iv) $H^* = \arg\min_\eta \mathcal{L}_\beta(\eta)$ is closed under ϕ_g; if the optimum is unique, $\phi_g(\eta^*) = \eta^*$.

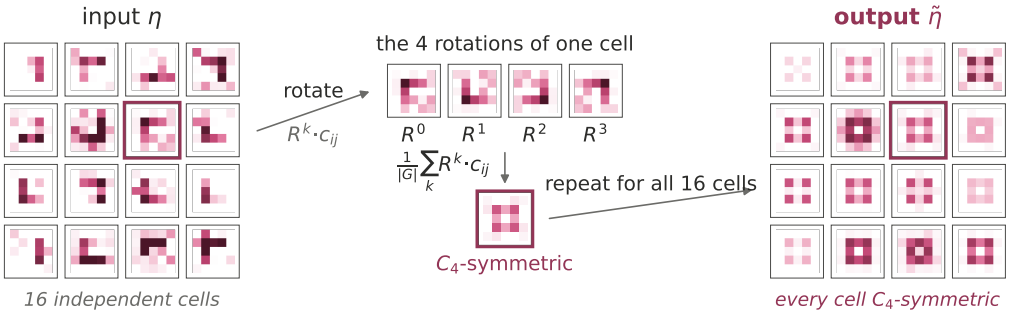
Takeaway: get η into H_G *once*, SGD on augmented data keeps the BNN exactly equivariant forever.

Symmetrize once, and training never leaves H_G



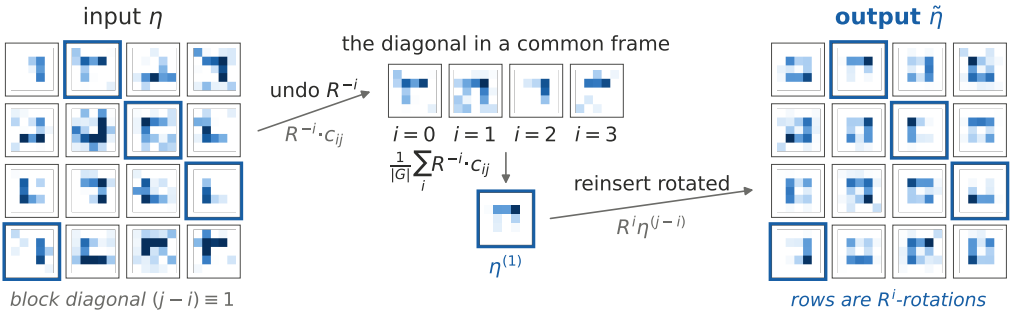
Into H_G (1/3): geometric averaging

Orbit-average the natural parameters, separately for each filter: $\tilde{\eta} = \frac{1}{|G|} \sum_g \phi_g(\eta)$. For orthogonal actions this is exactly the orthogonal projection onto H_G.



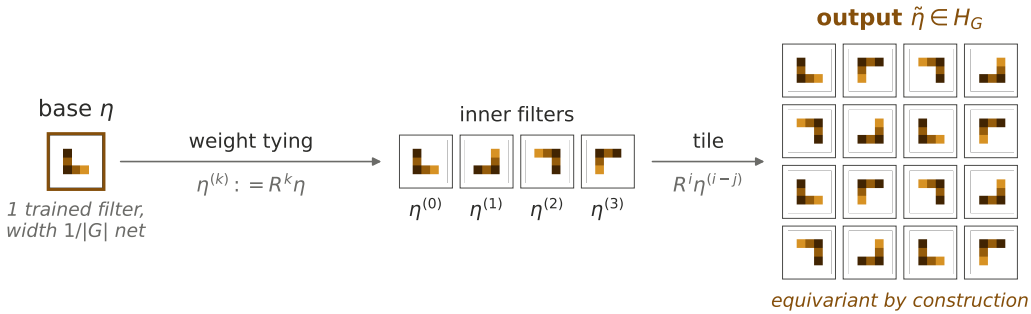
Into H_G (2/3): projection

Combine GCNN action (Cohen and Welling, 2016): rotate the filters and roll them along the channel dimension, so the orbit average couples different cells.



Into H_G (3/3): orbit expansion

GCNN-inspired (Cohen and Welling, 2016), but in the opposite direction: train a width-$1/|G|$ base net, then expand it to full width.



What if the prior is wrong, or sampling is finite?

Non-invariant prior

$$R_{\text{aug}}(\eta^*_\beta(p_0)) \leq R^*_{\text{aug}} + \beta\, C(p_0), \qquad \beta = 1/N$$

prior misalignment is suppressed by augmented dataset size N.

Finite Monte Carlo sampling

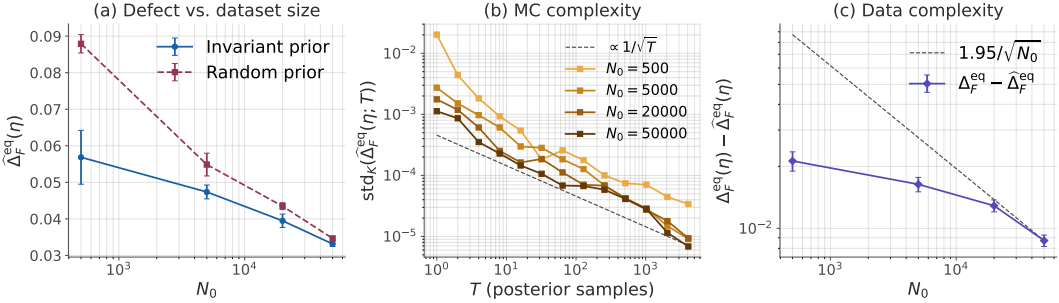
$$\widehat{\Delta}^{\text{eq}}_F(\eta) \leq \Delta^{\text{eq}}_F(\eta) + C_{\text{data}}\sqrt{\frac{\log(2/\delta)}{N_0}} + C_{\text{weight}}\sqrt{\frac{\log(2/\delta)}{T}} \quad \text{w.p. } 1-\delta$$

measured defect decays as $1/\sqrt{T}$, $1/\sqrt{N_0}$, where T is the sampling size and N_0 is the dataset size before augmentation.

Experiments I: the theory holds numerically

C_4 rotations on FashionMNIST, conv BNN, mean-field Gaussian VI.

- Invariant and random priors converge as N_0 grows
- measured defect exceeds the true one by $\mathcal{O}(1/\sqrt{T})$ (fitted slopes ≈ -0.5)
- Training gap follows the predicted $1/\sqrt{N_0}$



Experiments II: orbit expansion wins

FashionMNIST, $N_0 = 5\,000$, C_4 augmentation, SGD, 5 seeds:

Method	Acc. (%) ↑	OSP ↑	Sym. KL ↓
Baseline (augmentation only)	79.0 ± 0.6	0.927	0.0217
Geometric averaging	78.6 ± 0.4	0.950	0.0094
Projection	78.4 ± 0.8	0.931	0.0174
Orbit expansion	$\mathbf{80.4 \pm 0.4}$	$\mathbf{0.963}$	$\mathbf{0.0057}$

- All methods more equivariant; orbit expansion also *more accurate*
- Residual non-equivariance = the predicted Monte Carlo defect
- AdamW breaks update commutativity ⇒ benefit shrinks, as the theory predicts

Conclusion

Summary

- Augmented VI training preserves the equivariant subspace H_G
- Prior effects vanish as $1/N$; sampling defect as $1/\sqrt{T} + 1/\sqrt{N_0}$
- **Orbit expansion**: better accuracy *and* equivariance, one training run

Limitations & outlook

- Exponential families: mixture / flow posteriors need a new closedness argument
- Finite groups: continuous G within reach via Haar sampling

Paper: `arxiv.org/abs/2606.26273`

References

Gerken, Jan E. and Pan Kessel (July 2024). "Emergent Equivariance in Deep Ensembles". In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, pp. 15438–15465. arXiv: 2403.03103 [cs].

Nordenfors, Oskar and Axel Flinth (2025). *Ensembles provably learn equivariance through data augmentation*. arXiv: 2410.01452 [cs.LG].

Cohen, Taco and Max Welling (June 2016). "Group Equivariant Convolutional Networks". In: *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, pp. 2990–2999. arXiv: 1602.07576. (Visited on 09/23/2023).