EQUIVARIANCE AND AUGMENTATION FOR BAYESIAN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Symmetries are important for many deep learning tasks, ranging from applications in the sciences to medical imaging. However, there is an ongoing debate about whether to impose symmetry constraints on the neural network architecture (yielding equivariant neural networks) or learn them from augmented training data. Although equivariant networks are well-studied theoretically, much less is known about data augmentation, since analyzing augmentation requires control over the training dynamics. Inspired by recent results showing that augmented infinite deep ensembles are exactly equivariant, we study data augmentation for Bayesian neural networks (BNNs) trained with variational inference. We focus on variational distributions in the exponential family and derive conditions under which exact equivariance is reached. We furthermore obtain bounds on the equivariance error and introduce three novel symmetrization techniques which boost the effect of data augmentation in this setting. We conduct extensive numerical experiments which show that one of our symmetrization methods (orbit expansion) outperforms the baseline in both equivariance and overall performance. Our code is available at this anonymous repository.

1 INTRODUCTION

In recent years, symmetric learning tasks have become an important field of study. After an initial focus on specialized equivariant networks which impose the symmetry constraints layer-by-layer (Bronstein et al., 2017), attention has recently shifted towards learning symmetries from augmented training data (Chen et al., 2020). This approach has the advantage of, given an efficient symmetry transformation mechanism, being straightforward to implement and can be used together with highly-optimized and well-performing architectures (Wang et al., 2024). However, as the symmetry is merely learned and not imposed, it is only achieved approximately. Therefore, novel techniques are required which improve the symmetry gain derived from augmented training.

The explicit layer-wise equivariant neural networks are readily accessible to theoretical analysis, while data augmentation is considerably harder to study since it involves the training dynamics (see

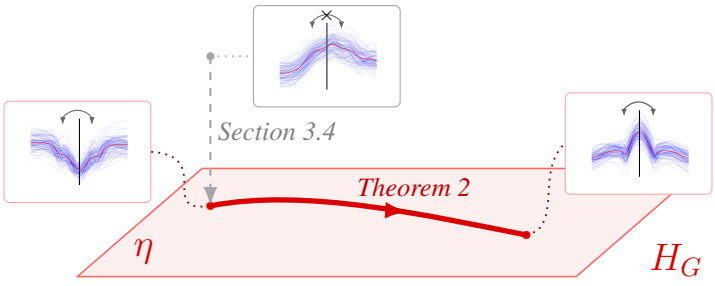


Figure 1: Natural parameters η for a variational distribution in the exponential family that lie in H_G correspond to symmetric BNNs, here exemplified with a reflection symmetry. Our main Theorem 2 implies that H_G is invariant for augmented training. Through the symmetrization strategies described in Section 3.4, we can increase the equivariance of the final model.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the related works section). It is however possible to show that in *expectation over initializations*, data augmentation leads to exact equivariance (Gerken & Kessel, 2024; Nordenfors & Flinth, 2025a). A practical but costly method for approximating such expected values is training a deep ensemble.

The purpose of this work is to investigate a cheaper way to realize such "equivariance in expectation": training Bayesian neural networks (BNNs) with variational inference on augmented data. In this setting, sampling from the posterior predictive distribution replaces the inference step on the ensemble and yields Bayesian uncertainty estimates. In contrast to the one training run per ensemble member for deep ensembles, only one training run is necessary to obtain the variational posterior. Furthermore, since BNNs show stable out of distribution behavior, they are particularly well suited for small datasets, the regime in which data augmentation can be expected to have the largest effect.

Our main contributions are the following:

- We study BNNs trained on augmented data with a variational distribution from the exponential family. Under some mild assumptions, we show that when the prior is invariant and training starts from an invariant variational distribution, it stays invariant throughout training. This generalizes a similar result from Nordenfors et al. (2025) for non-Bayesian network training.
- We derive bounds for the deviation of the variational distribution from equivariance if the prior is not equivariant. We furthermore prove bounds for the equivariance error in the predictions due to finite sampling. These theoretical results are validated numerically.
- We introduce three symmetrization methods (geometric averaging, projection and orbit expansion) that can be applied during training to improve the equivariance properties of the BNN. We test these techniques in extensive numerical experiments for image classification. *Orbit expansion* outperforms the baseline in both model performance and equivariance.

2 Related Work

Equivariant neural networks. The issue of symmetry, i.e. *invariance* and *equivariance*, of deep neural nets has grown into an entire subfield called *geometric deep learning* (Bronstein et al., 2017). The most prominent construction of equivariant networks is the layerwise one. This strategy originates from GCNNs, Group Convolutional Neural Networks (Cohen & Welling, 2016), but has by now been generalized to virtually any symmetry induced by any group (Kondor & Trivedi, 2018; Keriven & Peyré, 2019; Gerken et al., 2023). There are also other strategies, such as learning from invariants (Villar et al., 2021), frame averaging (Puny et al., 2022), fundamental domain projection (Aslan et al., 2023) and group averaging. There are also works in which symmetries are enforced approximately, for instance through so called weight annealing (Wang et al., 2022).

Augmentation and training dynamics. The question of the effect of data augmentation on the training dynamics of neural networks have been treated in several simplified contexts, such as feature averaged models (Lyle et al., 2020; Elesedy & Zaidi, 2021) as well as linear neural networks (Lawrence et al., 2022; Chen & Zhu, 2024; Duan & Montúfar, 2026). In these cases, it is often possible to prove equivalence of augmentation and equivariance. Fully non-linear networks were treated in Nordenfors et al. (2025); Nordenfors & Flinth (2025b), whose results we generalize to Bayesian networks here.

Empirical studies on the difference between augmentation and restrictions are plentiful. More systematic treatments are Gandikota et al. (2021); Gerken et al. (2022); Brehmer et al. (2025).

Bayesian neural networks. Bayesian approaches to deep learning have been studied for many years (MacKay, 1992) since they offer uncertainty estimates for neural networks that are otherwise black-box models (see the PhD thesis by Gal (2016) for an overview). However, making BNNs practically applicable requires integrating variational inference into the deep learning training methodology (Graves, 2011; Blundell et al., 2015; Kingma & Welling, 2022). For a review of BNNs with a focus on practical applications, see Jospin et al. (2022).

Not many prior works consider symmetries in relation to BNNs. van der Ouderaa & van der Wilk (2022) propose a probabilistic group-averaging of BNNs in order to realize a soft symmetry constraint, optimized on the data. Closest to our work is Mourdoukoutas et al. (2021) which uses a

specific prior that combines different weight-sharing schemes and therefore symmetry constraints. During training, the network learns which symmetry fits the data best. In contrast, we consider training on augmented data with a generic prior that does not impose weight-sharing.

3 THEORY

We develop a theoretical framework for understanding how data augmentation induces equivariance in variational Bayesian inference. We proceed in three steps: first, we characterize when exponential families are closed under group actions (Section 3.2); second, we show that data-augmented training makes the ELBO invariant, and how that affects the training (Section 3.3); third, we propose symmetrization mechanisms and analyze their properties (Section 3.4).

3.1 PRELIMINARIES

In this section, we introduce the mathematical tools used throughout this paper. We begin with exponential families, which provide the structural backbone for our theoretical analysis, before reviewing variational inference and the group theoretic concepts needed to formalize symmetry.

Exponential families. An exponential family of probability distributions is defined through a base measure $h(\theta)$, a sufficient statistic $T(\theta) \in \mathbb{R}^k$ and a log-partition function $A(\eta) := \log \int_\Theta h(\theta) \exp(\eta^\top T(\theta)) d\theta$. A distribution belongs to the exponential family if its density has the form

$$q_\eta(\theta) = h(\theta) \exp(\eta^\top T(\theta) - A(\eta)) \,, \tag{1}$$

where $\eta \in H \subseteq \mathbb{R}^k$ is the *natural parameter*. Examples of such families include normal, exponential and log-normal distributions. An exponential family $\mathcal{Q} := \{q_\eta(\theta) \mid \eta \in H\}$, is *regular* if H is open, and *minimal* if the components of $T(\theta)$ are affinely independent.

Group actions and push-forward distributions. We assume that a group G acts on the input space $\mathcal{X}$, the output space $\mathcal{Y}$, and the parameter space Θ via representations $\rho_\mathcal{X}$, $\rho_\mathcal{Y}$, and ρ_Θ respectively. Throughout, we assume that ρ_Θ is compatible with the data representations $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ in the sense that

$$f(\rho_\mathcal{X}(g)x; \theta) = \rho_\mathcal{Y}(g) f(x; \rho_\Theta(g)^{-1}\theta). \tag{2}$$

Following the layered decomposition $\Theta = \bigoplus_\ell \Theta_\ell$ of a neural network's parameters, we assume that ρ_Θ acts per layer, i.e., $\rho_\Theta = \bigoplus_\ell \rho_{\Theta_\ell}$. There is a canonical construction of such a ρ_Θ: one introduces representations ρ_{Θ_ℓ} on the hidden layers and then imposes equivariance on each linear layer with respect to the two representations acting on the input and output of that layer. As long as the non-linearities are equivariant with respect to the same representations (which is always the case for pointwise non-linearities and permutation representations) one obtains equation 2. For more details, see Nordenfors et al. (2025). The choice of ρ_Θ, or equivalently the choice of the hidden layer representations, is a modeling decision that we revisit in Section 3.4. When there is no risk of confusion, we write gx and gy as shorthands for $\rho_\mathcal{X}(g)x$ and $\rho_\mathcal{Y}(g)y$, but retain the explicit notation $\rho(g)\theta$ for the parameter space, where the representation structure is central to our analysis. The push-forward of a distribution q under $g \in G$ is defined as

$$\mathcal{T}_g \# q(B) := q(\rho(g)^{-1} B) \tag{3}$$

for any measurable set B, with density $\mathcal{T}_g \# q(\theta) = q(\rho(g)^{-1}\theta) \left|\det D(\rho(g)^{-1})(\theta)\right|$. A distribution q is called G-invariant if $\mathcal{T}_g \# q = q$ for all $g \in G$.

Bayesian neural networks (BNN) and variational inference (VI). A neural network is a parametrized function $f(\cdot; \theta) : \mathcal{X} \to \mathcal{Y}$. They define conditional label distributions $p(y \,|\, x, \theta) = p(y \,|\, f(x; \theta))$. In the Bayesian treatment, we put a prior $p_0(\theta)$ on the weights and calculate the posterior $p(\theta \mid \mathcal{D})$ given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N_0}$ via Bayes' rule. However, since the true posterior distribution is typically intractable due to the high-dimensional marginal likelihood integral, one makes a variational inference approximation $q_\eta \in \mathcal{Q}$, and then chooses the parameters η to maximize the *evidence lower bound* (Bishop, 2006)

$$\mathrm{ELBO}(\eta) := \mathbb{E}_{q_\eta}[\log p(\mathcal{D} \mid \theta)] - D_{\mathrm{KL}}(q_\eta \| p_0) \,. \tag{4}$$

The posterior predictive is approximated by Monte Carlo sampling from q_η:

$$F_\eta(x) := \mathbb{E}_{\theta\sim q_\eta}[f(x;\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(x;\theta^{(t)}), \quad \theta^{(t)} \sim q_\eta . \tag{5}$$

Data augmentation. Given a finite group G and a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N_0}$ drawn i.i.d. from a data distribution $P_\mathcal{X}$, data augmentation constructs an augmented dataset $\mathcal{D}_{\text{aug}} = \{(gx_i, gy_i) \mid g \in G, (x_i, y_i) \in \mathcal{D}\}$, expanding the dataset by a factor of $|G|$. One could also consider a continuous compact group by sampling finitely many Haar-distributed group elements. This comes with some subtleties that we discuss in Appendix A. In the following, we instead restrict to the case of finite G. Following the mini-batch scaling convention of Graves (2011) and Blundell et al. (2015), instead of using the ELBO equation 4 directly, we work with a β-weighted objective:

$$\mathcal{L}_\beta(\eta) := \underbrace{\tfrac{1}{N}\,\mathbb{E}_{q_\eta}[-\log p(\mathcal{D}_{\text{aug}} \mid \theta)]}_{=:\, R_{\text{aug}}(\eta)} + \beta D_{\text{KL}}(q_\eta \| p_0), \tag{6}$$

where $N = |G| \cdot N_0$ is the augmented dataset size. $\mathcal{L}_\beta$ with $\beta = 1/N$ is the rescaled negative ELBO.

If we assume the following invariance of the forward model:

$$p(gy \mid gf(x;\theta)) = p(y \mid f(x;\theta)), \tag{7}$$

the compatibility of ρ_Θ with the data representation implies that the augmented likelihood inherits the symmetry of the data. Note that condition equation 7 is weak – it is e.g. satisfied for $y \sim f(x;\theta) + e$ for a G-invariant error e – and furthermore independent of the architecture f.

Proposition 1 (Invariant augmented likelihood)**.** *Let $L_{\text{aug}}(\theta) := p(\mathcal{D}_{\text{aug}} \mid \theta)$ be the likelihood of the augmented dataset. Under the compatibility assumption equation 2 as well as the invariance assumption equation 7, the likelihood is G-invariant:*

$$L_{\text{aug}}(\rho(g)^{-1}\theta) = L_{\text{aug}}(\theta),\ \forall g \in G,\ \theta \in \Theta . \tag{8}$$

The proof is given in Appendix B.

Equivariance. A single network $f(\cdot\,;\theta)$ is G-equivariant if $f(gx;\theta) = gf(x;\theta)$ for all $x \in \mathcal{X}$, $g \in G$. More generally, the predictive map F_η defined in equation 5 is G_ε-*equivariant* if the equivariance defect $\Delta_F^{\text{eq}}(\eta)$ is bounded by ε,

$$\Delta_F^{\text{eq}}(\eta) := \mathbb{E}_{x\sim P_\mathcal{X},\, g\sim\nu_G}\left[\|F_\eta(gx) - gF_\eta(x)\|^2\right] \le \varepsilon. \tag{9}$$

Here, ν_G denotes the normalized Haar measure on G. When $\varepsilon = 0$, this reduces to exact G-equivariance of F_η $P_\mathcal{X}$-a.s. One central question is whether minimizing $\mathcal{L}_\beta$ on $\mathcal{D}_{\text{aug}}$ drives $\Delta_F^{\text{eq}}(\eta)$ to zero, and at what rate.

3.2 Closed exponential family under group actions

Neural networks can learn equivariance efficiently from augmented data only when the parameter space is closed under group transformations (Nordenfors et al., 2025). Similarly, in the Bayesian setting, we restrict the variational family to those closed under group transformations. As the following theorem shows, this imposes constraints on the exponential family. We provide the proof in Appendix C.

Theorem 1 (Closure under push-forward)**.** *A regular minimal exponential family $\mathcal{Q}$ is closed under the push-forward operator $\mathcal{T}_g$ for all $g \in G$ if and only if for each g there exist $M_g \in \mathbb{R}^{k\times k}$, $d_g \in \mathbb{R}^k$, $b_g \in \mathbb{R}^k$ and $c_g \in \mathbb{R}$ such that for all $\theta \in \Theta$,*

$$T(\rho(g)^{-1}\theta) = M_g T(\theta) + d_g, \tag{10}$$

$$h(\rho(g)^{-1}\theta)\left|\det D(\rho(g)^{-1})(\theta)\right| = h(\theta)\exp(b_g^\top T(\theta) + c_g). \tag{11}$$

When these conditions hold, $\mathcal{T}_g \# q_\eta = q_{\eta_g}$ where $\eta_g = M_g^\top \eta + b_g$, and the transformation $\phi_g : \eta \mapsto M_g^\top \eta + b_g$ forms an affine action of G on the space of natural parameters H.

A family $\mathcal{Q}$ is hence closed if and only if there exists an affine representation of the group acting on the ambient space $\mathbb{R}^k$ of the space of natural parameters, that makes $T : \Theta \to \mathbb{R}^k$ equivariant, and with respect to which the base measure transforms according to equation 11.

Example 1 (Mean-field Gaussian under permutation actions)**.** *Consider the mean-field Gaussian family $\mathcal{Q} = \{\mathcal{N}(\mu, \mathrm{diag}(\sigma^2)) : \mu \in \mathbb{R}^d, \sigma^2 \in \mathbb{R}^d_{>0}\}$ on $\Theta = \mathbb{R}^d$. This family has sufficient statistic $T(\theta) = (\theta, \theta \odot \theta) \in \mathbb{R}^d \oplus \mathbb{R}^d$, constant base measure $h(\theta) = (2\pi)^{-d/2}$, and natural parameter $\eta = (\eta_1, \eta_2) \in \mathbb{R}^d \oplus \mathbb{R}^d$ given by $(\eta_1)_i = \mu_i/\sigma_i^2$ and $(\eta_2)_i = -1/(2\sigma_i^2)$.*

Let G act on Θ by a permutation representation $\rho(g) = R_g$. Then, both conditions of Theorem 1 are met: permutation matrices commute with the Hadamard product, so $T(R_g^{-1}\theta) = (R_g^{-1} \oplus R_g^{-1})T(\theta)$ giving $M_g = R_g^{-1} \oplus R_g^{-1}$ and $d_g = 0$. Since $|\det R_g^{-1}| = 1$ and h is constant, equation 11 is fulfilled by $b_g = c_g = 0$. Using $(R_g^{-1})^\top = R_g$, the induced natural-parameter transformation is

$$\phi_g(\eta) = M_g^\top \eta = (R_g \oplus R_g)\eta, \tag{12}$$

which simply amounts to permuting the entries of η_1 and η_2. Equivalently,

$$\mathcal{T}_g \# \mathcal{N}(\mu, \mathrm{diag}(\sigma^2)) = \mathcal{N}(R_g\mu, \mathrm{diag}(R_g\sigma^2)), \tag{13}$$

so the family is indeed closed.

3.3 Data augmentation induces equivariance

The representation ϕ_g fixes the set of symmetric parameters $H_G := \{\eta \in H \mid \phi_g(\eta) = \eta, \forall\, g \in G\}$. Just as in the single network case, parameters in H_G correspond to equivariant predictive maps.

Proposition 2. *For $\eta \in H_G$, F_η is G-equivariant.*

Proof. If $\eta \in H_G$, $\mathcal{T}_g \# q_\eta = q_{\phi_g(\eta)} = q_\eta$. Hence, if $\theta \sim q_\eta$, $\rho(g)\theta \sim q_\eta$ for every $g \in G$. This together with the compatibility assumption implies that $f(gx; \theta) = gf(x; \rho(g^{-1})\theta) \overset{\mathrm{d}}{=} gf(x; \theta)$, where the last equality is in distribution. Taking the expectation over θ yields $F_\eta(gx) = gF_\eta(x)$, which is the claim. □

Our main result states that, under the conditions presented below, H_G is invariant under gradient descent on augmented data. Initializing η there guarantees equivariant F_η throughout training.

Assumption 1 (Invariant prior)**.** *The prior $p_0 = q_{\eta_0}(\theta)$ for some η_0 is G-invariant, i.e., $\mathcal{T}_g \# p_0 = p_0$ for all $g \in G$, or equivalently $\phi_g(\eta_0) = \eta_0$.*

Assumption 2 (Volume-preserving action)**.** $|\det D(\rho(g)^{-1})(\theta)| = 1$ *for all $g \in G$ and $\theta \in \Theta$.*

Both these assumptions are natural: an invariant prior matches the hypothesis of a G-symmetric task, and volume-preservation is automatic for the, in practice very common, orthogonal actions.

Theorem 2 (Equivariance of VI)**.** *Assumptions 1 and 2 imply the following for variational inference using invariant likelihoods and closed exponential families:*

(i) ***Loss invariance.*** *The loss $\mathcal{L}_\beta$ is invariant under the natural parameter transformation:*

$$\mathcal{L}_\beta(\eta) = \mathcal{L}_\beta(\phi_g(\eta)), \quad \forall g \in G,\, \eta \in H. \tag{14}$$

(ii) ***Gradient equivariance and update commutativity.*** *The gradient of the loss $\mathcal{L}_\beta$ satisfies*

$$\nabla_\eta \mathcal{L}_\beta(\phi_g(\eta)) = M_g^{-1} \nabla_\eta \mathcal{L}_\beta(\eta), \quad \forall g \in G,\, \eta \in H. \tag{15}$$

Consequently, for M_g orthogonal, gradient descent $\mathcal{U}(\eta) = \eta - \alpha \nabla_\eta \mathcal{L}_\beta(\eta)$ commutes with ϕ_g:

$$\mathcal{U}(\phi_g(\eta)) = \phi_g(\mathcal{U}(\eta)). \tag{16}$$

(iii) ***Invariant subspace is preserved.*** *For M_g orthogonal, the equivariant subspace H_G is preserved by gradient updates: if $\eta^{(0)} \in H_G$, then $\eta^{(t)} \in H_G$ for all $t \geq 0$.*

(iv) ***Optimal solutions inherit symmetry.*** *The set of optimal natural parameters $H^* = \arg\min_\eta \mathcal{L}_\beta(\eta)$ is closed under ϕ_g for all $g \in G$. If the optimum is unique, then $\phi_g(\eta^*) = \eta^*$.*

The proof is given in Appendix D. This formalizes the intuition behind data augmentation: by making the training objective symmetric, augmentation ensures that H_G is invariant – and in the case of a unique minimum, even that the minimum is in H_G.

Note that Theorem 2 does not rely on augmented data directly, only on an invariant likelihood (which is implied by data augmentation according to Proposition 1). Therefore, the theorem is still applicable when the symmetry is present in the data but unknown. However in such a case, it is hard to guarantee that the prior is invariant. The next theorem, which we prove in Appendix E, shows that as the dataset size increases, the effect of a non-invariant prior vanishes.

Theorem 3 (Prior-independent convergence)**.** *Assume Assumption 2, and that the likelihood is invariant. Let* $R^*_{\mathrm{aug}} := \inf_\eta R_{\mathrm{aug}}(\eta)$ *and* $\eta^*_\beta(p_0) \in \arg\min_\eta \mathcal{L}_\beta(\eta; p_0)$*. We then have*

$$R_{\mathrm{aug}}(\eta^*_\beta(p_0)) \leq R^*_{\mathrm{aug}} + \beta \cdot C(p_0), \tag{17}$$

for any prior p_0*, where* $C(p_0) := \inf_{\eta^* \in H^*_R} D_{\mathrm{KL}}(q_{\eta^*} \| p_0)$ *and* $H^*_R := \arg\min_\eta R_{\mathrm{aug}}(\eta)$ *is the set of risk-optimal natural parameters. As* $\beta \to 0$*, all minimizers achieve* R^*_{aug} *regardless of the prior.*

Note that the quantity $C(p_0)$ measures how well the prior aligns with the risk-optimal orbit. Since $\beta = 1/N$, with sufficient data, it becomes irrelevant. In contrast, R^*_{aug} is asymptotically constant; It will almost surely converge to the true likelihood $\mathbb{E}_{(x,y)\sim P_{\mathcal{X}}} \mathbb{E}_{g\sim\nu_G} \mathbb{E}_{\theta\sim q_\eta}[-\log p((gx, gy) \mid \theta)]$.

Next, we address the fact that the true predictive maps F_η in practice must be approximated with finite sample means. This will indeed cause a Monte Carlo equivariance defect $\widehat{\Delta}^{\mathrm{eq}}_F(\eta)$ (defined formally in equation 77), but that defect vanishes as the number of samples grows, at the expected rate, as the following theorem shows (proof in Appendix F). We verify it and Theorem 3 in Section 4.1.

Theorem 4 (Complexity for G_ε-equivariance)**.** *Suppose* $f(x; \theta)$ *is uniformly bounded in* $x \in \mathcal{X}$ *and* q_η*-a.s. in* θ*, and that the* G*-action on* $\mathcal{Y}$ *is orthogonal. Then there exist constants* C_{data} *and* C_{weight} *such that, for any fixed* η*, with probability at least* $1 - \delta$ *over the* N_0 *i.i.d. inputs* $x_i \sim P_{\mathcal{X}}$ *and the* T *i.i.d. posterior samples* $\theta^{(t)} \sim q_\eta$*, the Monte Carlo equivariance defect satisfies*

$$\widehat{\Delta}^{\mathrm{eq}}_F(\eta) \leq \Delta^{\mathrm{eq}}_F(\eta) + C_{\mathrm{data}} \sqrt{\frac{\log(2/\delta)}{N_0}} + C_{\mathrm{weight}} \sqrt{\frac{\log(2/\delta)}{T}}\,. \tag{18}$$

In other words, for any $\varepsilon_0, \varepsilon_{\mathrm{data}}, \varepsilon_{\mathrm{weight}} > 0$*, if* $\Delta^{\mathrm{eq}}_F(\eta) < \varepsilon_0$*, it suffices to choose*

$$N_0 \geq \frac{C^2_{\mathrm{data}}}{\varepsilon^2_{\mathrm{data}}} \log\left(\frac{2}{\delta}\right), \quad T \geq \frac{C^2_{\mathrm{weight}}}{\varepsilon^2_{\mathrm{weight}}} \log\left(\frac{2}{\delta}\right) \tag{19}$$

in order for $\widehat{\Delta}^{\mathrm{eq}}_F(\eta) < \varepsilon_0 + \varepsilon_{\mathrm{data}} + \varepsilon_{\mathrm{weight}}$ *with probability at least* $1 - \delta$*.*

3.4 Symmetrization of the variational posterior

According to Theorem 2(iii), the space of parameters corresponding to equivariant variational distributions H_G is preserved during training. Hence, if we ensure that the parameters lie in H_G at some point during training, we will arrive at an equivariant variational posterior at the end of it. Therefore, we propose two complementary mechanisms that place η into H_G, one by projecting an existing posterior, one by constructing the posterior from a smaller base, see Figure 1.

Orbit averaging. Given a posterior distribution q_η at any point during training, we construct a G-invariant posterior by averaging its natural parameter over the orbit:

$$\tilde{\eta} = \frac{1}{|G|} \sum_{g\in G} \phi_g(\eta). \tag{20}$$

We can interpret this operation in two ways: First, $\tilde{\eta}$ is the natural parameter of the *geometric mean distribution* $q_{\tilde{\eta}} \propto \left[\prod_{g\in G} \left(\mathcal{T}_g \# q_\eta\right)(\theta)\right]^{\frac{1}{|G|}}$ (which lies in $\mathcal{Q}$ by Lemma 1 in Appendix G). Secondly, for orthogonal ϕ_g, $\tilde{\eta}$ is the orthogonal projection of η onto the invariant subspace H_G under the Euclidean inner product – this in particular shows that $\tilde{\eta} \in H_G$.

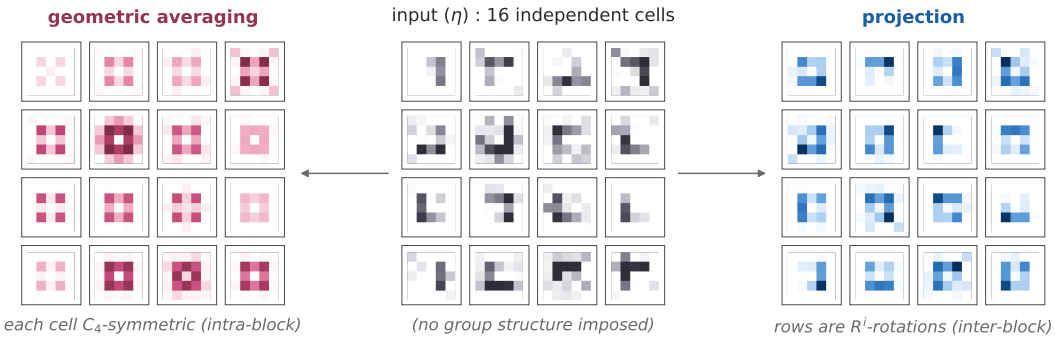


Figure 2: Two specializations of the orbit averaging equation 20 on the second convolutional layer of a C_4-equivariant network. Both start from the same 4×4 block of variational parameters η.

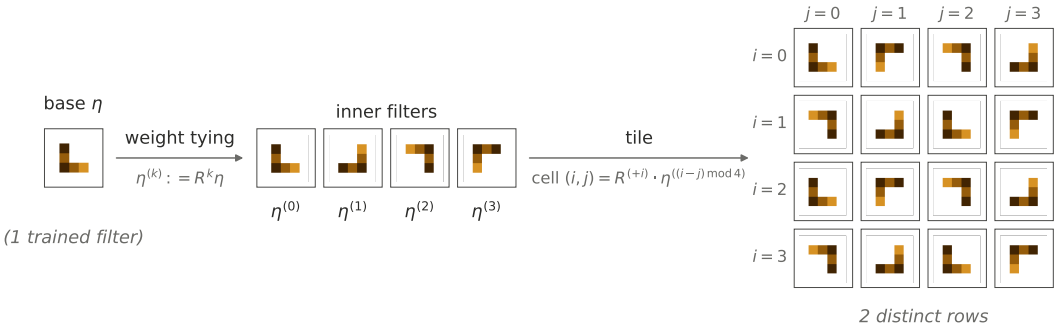


Figure 3: Filter arrangement for orbit expansion under C_4, starting from a single base filter η.

The mechanism specializes through the choice of ρ_Θ. Let us exemplify this in the case of C_4 acting on convolutional layers: First, we can define a representation acting by rotating each individual convolutional channel and secondly a representation that additionally permutes channel blocks, inspired by group convolutional neural networks (GCNNs) from Cohen & Welling (2016). The corresponding average operations on the filters are depicted in Figure 2. We will refer to the former as **geometric averaging**, and the latter as **projection**. We derive both in detail in Appendix H.

Orbit expansion. Instead of symmetrizing a general parameter of the intended size, we can also first train a small parameter and then, in a manner inspired by the weight-sharing structure of GCNNs, extend it to an equivariant one of bigger size. Concretely, a base network with $1/|G|$ of the target channel width is trained in Stage 1, yielding a posterior $q_{\eta_{\text{small}}}$. At the start of Stage 2, the network is expanded to full width by inserting rotated versions of each filter in $|G| \times |G|$-blocks (see Figure 3). We refer to this strategy as **orbit expansion**. The resulting posterior $q_{\eta_{\text{full}}}$ has a G-equivariant block structure at the parameter level. In Appendix I, we compare three different expansion strategies.

4 Experiments

In our experiments, we consider the cyclic group C_4 of rotations of multiples of $90°$ acting on FashionMNIST (Xiao et al., 2017). This setup allows us to augment exactly and enables extensive Monte Carlo sampling. Throughout our experiments, we use Gaussian variational distributions which satisfy the constraints of Theorem 1. In Appendix J, we provide additional ablations for other variational distributions. Code is provided at this anonymous repository.

4.1 Validation of Theorems

We start by verifying the predictions of Theorems 3 and 4. We use a Bayesian neural network with two convolutional layers and one linear layer and vary the dataset size $N_0 \in$

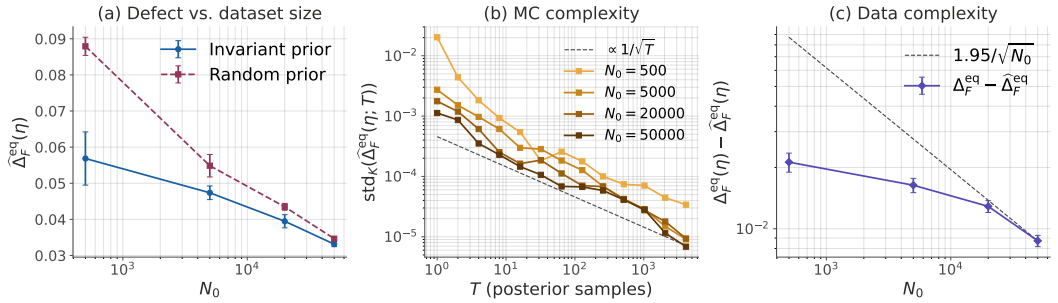


Figure 4: Empirical validation of Theorems 3 and 4. **(a)** The empirical equivariance defect decreases with N_0 under both invariant and random Gaussian priors, and the two curves converge. **(b)** K-fold standard deviation of the Monte Carlo estimate $\widehat{\Delta}_F^{\mathrm{eq}}(\eta; T)$ across $K = 10$ independent runs at each T. The decay matches the $\mathcal{O}(1/\sqrt{T})$ rate predicted by Theorem 4 (dotted reference line, slope -0.5). **(c)** The training gap satisfies the $C_{\mathrm{data}}/\sqrt{N_0}$ envelope of Theorem 4 with $C_{\mathrm{data}} \approx 1.95$. Error bars are ± 1 s.d. across 5 seeds.

$\{500, 5000, 20000, 50000\}$. For Theorem 3, we compare an invariant isotropic Gaussian prior $p_0 = \mathcal{N}(0, I)$ against a non-invariant prior with means drawn from a normal distribution. Posterior predictives are estimated with T Monte Carlo samples. The results are presented in Figure 4.

Priors. The empirical defect $\widehat{\Delta}_F^{\mathrm{eq}}$ decreases monotonically with N_0 under both priors, from 0.057 ± 0.007 to 0.033 ± 0.001 (invariant) and 0.088 ± 0.003 to 0.035 ± 0.001 (random); the two curves converge as N_0 grows, consistent with Theorem 3.

Monte Carlo complexity. Theorem 4 predicts that the gap between the T-sample Monte Carlo estimate $\widehat{\Delta}_F^{\mathrm{eq}}(\eta; T)$ and the true predictive map decays as $\mathcal{O}(1/\sqrt{T})$. We assess this directly by drawing $K = 10$ independent T-sample estimates at each T and tracking their standard deviation across the K runs. The results are shown in Figure 4(b). The theorem is indeed verified: a log-log fit of the K-fold std against T yields slopes of -0.45, -0.66, -0.52, -0.51 for $N_0 \in \{500, 5000, 20000, 50000\}$ respectively, all close to the predicted -0.5.

Data complexity. In Figure 4(c), we use $T_{\max} = 1024$ as a proxy for $T \to \infty$ and plot the training gap $\Delta_F^{\mathrm{eq}}(\eta) - \widehat{\Delta}_F^{\mathrm{eq}}(\eta)$ together with $\mathcal{O}(1/\sqrt{N_0})$ envelope from Theorem 4, with $C_{\mathrm{data}} \approx 1.95$. The measured gap decreases with N_0 in close agreement with the predicted $1/\sqrt{N_0}$ behavior.

4.2 Symmetrization on image classification

Next, we empirically compare the symmetrization mechanisms from Section 3.4.

Setup. We train a convolutional Bayesian neural network with two convolutional layers with 32 and 64 channels, followed by a classification layer using a diagonal Gaussian variational family and a standard isotropic Gaussian prior. The training set consists of 5 000 randomly chosen FashionMNIST images, each augmented with the full C_4 orbit (producing 20 000 training samples).

The parameter space of each convolutional layer decomposes along the channel dimension. We implement the orbit-averaging mechanism of Section 3.4 under two natural choices of ρ_Θ introduced there: **geometric averaging** (per-filter C_4-symmetrization, channel arrangement unconstrained) and **projection** (averaging across filters and channel-block permutation). Both satisfy the closure conditions of Theorem 1, so the results of Section 3 apply uniformly. **Orbit expansion** is implemented with the GCNN-expansion from Figure 3. All methods share the same architecture, prior, optimizer, learning rate, epoch count, and number of MC samples; they differ only in when and how the posterior is symmetrized. We use SGD as the optimizer throughout the main text: its update rule is linear in the gradient and therefore commutes exactly with ϕ_g (Theorem 2(ii)), so the equivariant subspace H_G is preserved under training.

Each symmetrization mechanism is applied at epoch $n \in \{0, 20, 100\}$ to study the effect of trigger timing (Section 3.4). After the intervention, all methods continue training until epoch 500. As a baseline, we train a randomly initialized network on augmented data.

Metrics. We report: (1) classification accuracy (%), (2) orbit same prediction (OSP): the fraction of predictions that agree with the mode on the orbit, averaged over the test set, and (3) symmetric KL divergence between predictive distributions on original and rotated inputs.

Results. Our results are summarized in Table 1. First note that none of our strategies yields perfect equivariance, as expected from the Monte Carlo residual quantified by Theorem 4, present even when $\eta \in H_G$. However, all strategies yield more equivariant models, and the orbit expansion model leads to a better overall performance.

We note that the performance across all metrics and strategies generally decreases when the intervention is applied later. This is natural, since a later trigger time exactly corresponds to a shorter training time post reinitialization. OSP and Sym.KL of the orbit expansion are the only exceptions to this trend, but these improvements start from values already better than those of all other methods.

Table 1: Effect of symmetrization mechanism and timing on FashionMNIST ($N_0 = 5\,000$, C_4 augmentation, $10\,000$ test samples), all with SGD. Results are mean $\pm$ std over 5 seeds. **Bold**: best in each column. Shaded: best equivariance–accuracy trade-off. Full results across optimizers and trigger epochs are in Table 4.

Method	Epoch	Acc. (%) ↑	OSP ↑	Sym. KL ↓
Baseline	—	$79.0_{\pm 0.6}$	$0.927_{\pm .003}$	$0.0217_{\pm .0028}$
Geo. avg.	0	$78.6_{\pm 0.4}$	$0.950_{\pm .003}$	$0.0094_{\pm .0009}$
	20	$78.6_{\pm 0.8}$	$0.952_{\pm .005}$	$0.0094_{\pm .0020}$
	100	$78.0_{\pm 1.1}$	$0.946_{\pm .005}$	$0.0112_{\pm .0020}$
Proj.	0	$78.4_{\pm 0.8}$	$0.931_{\pm .004}$	$0.0174_{\pm .0026}$
	20	$78.5_{\pm 0.2}$	$0.931_{\pm .004}$	$0.0180_{\pm .0026}$
	100	$77.2_{\pm 0.9}$	$0.929_{\pm .012}$	$0.0182_{\pm .0046}$
Orbit exp.	0	$80.4_{\pm 0.4}$	$0.963_{\pm .003}$	$0.0057_{\pm .0004}$
	20	$\mathbf{80.4}_{\pm 0.6}$	$0.966_{\pm .004}$	$0.0048_{\pm .0009}$
	100	$80.0_{\pm 0.5}$	$\mathbf{0.968}_{\pm .003}$	$\mathbf{0.0042}_{\pm .0007}$

The orbit expansion yields the best results. We hypothesize that the reason for this is that the orbit expansion results in parameters with additional symmetries compared to the orbit averaging techniques, due to the restricted number of base filters (cf. Figures 3 and 5). These additional symmetries seem to induce more stability. To further explore such effects however remains future work.

In Appendix K we repeat the comparison with AdamW, whose adaptive rescaling violates the update commutation equation 16; there an earlier trigger *hurts* equivariance, so H_G is less stable. This is further evidence for the correctness of our theory.

5 Conclusion and Limitations

In this paper, we theoretically analyzed the effect of data augmentation for variational inference of Bayesian neural networks. We showed that (under some weak conditions) when the underlying exponential family of variational distributions is closed under a symmetry and an invariant prior is used, a set H_G of symmetric parameters corresponding to equivariant predictive maps becomes invariant under gradient training on augmented data. Based on this theorem, we proposed and tested several strategies to initialize the net in H_G. The orbit insertion strategy proved especially efficient, boosting both the performance and the equivariance compared to a baseline.

Throughout our work, we have made a number of assumptions that limit the scope of our results. In particular, we focused on exponential families. Although this is a broad class of probability distributions, some notable distributions are not part of it, such as Gaussian mixtures. Furthermore, we have restricted our analysis (and correspondingly our experiments) to finite groups to enable a straightforward definition of the (finite) augmented dataset. An extension to continuous groups would require an augmentation approach based on sampling from the symmetry group. Although this is outside the scope of this work, we do not see a conceptual obstacle to extending our results in this way (see also the discussion in Appendix A). A final limitation is the compatibility assumption equation 2. Given the flexibility of choice of intermediate representations, this restriction is however mild.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, assist in the writing of proofs, implement methods, and assist with translation. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to create and modify scientific figures, suggest experimental parameters, search for information, and improve readability. We have reviewed all AI-assisted work. We checked the LLM-generated claims and proofs to ensure that they were consistent with our original ideas. The LLM-generated code was verified and tested for correctness by one of the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Benjamin Aslan, Daniel Platt, and David Sheard. Group invariant machine learning by fundamental domain projections. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, pp. 181–218. PMLR, 2023.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006. ISBN 978-0-387-31073-2.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1613–1622. PMLR, June 2015.

Johann Brehmer, Sönke Behrends, Pim De Haan, and Taco Cohen. Does equivariance matter at scale? *Transactions on Machine Learning Research*, April 2025. ISSN 2835-8856.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *The Journal of Machine Learning Research*, 21(1):9885–9955, 2020.

Ziyu Chen and Wei Zhu. On the implicit bias of linear equivariant steerable networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2990–2999. PMLR, June 2016.

J. L. Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3):455–486, 1940. ISSN 00029947, 10886850.

Hao Duan and Guido Montúfar. Understanding learning invariance in deep linear networks. In Michael Bleher, Freya Jensen, Levin Maier, Diaaeldin Taha, and Anna Wienhard (eds.), *Proceedings of the Geometry, Topology, and Machine Learning Workshop*, volume 325 of *Proceedings of Machine Learning Research*, pp. 1–45. PMLR, 10–14 Nov 2026.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2959–2969. PMLR, 2021.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Kanchana Vaishnavi Gandikota, Jonas Geiping, Zorah Lähner, Adam Czapliński, and Michael Moeller. Training or architecture? How to incorporate invariance in neural networks. *arXiv:2106.10044*, 2021.

Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Equivariance versus augmentation for spherical images. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 7404–7421. PMLR, 2022.

Jan E. Gerken and Pan Kessel. Emergent Equivariance in Deep Ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 15438–15465. PMLR, July 2024.

Jan E. Gerken, Jimmy Aronsson, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Geometric deep learning and equivariant neural networks. *Artificial Intelligence Review*, June 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10502-7.

Alex Graves. Practical Variational Inference for Neural Networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, May 2022. ISSN 1556-6048. doi: 10.1109/MCI.2022.3155327.

Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114v11 [stat.ML]*, 2022.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pp. 2747–2755. PMLR, 2018.

Hannah Lawrence, Kristian Georgiev, Andrew Dienes, and Bobak T Kiani. Implicit bias of linear equivariant networks. In *International Conference on Machine Learning*, pp. 12096–12125. PMLR, 2022.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv:2005.00178 [cs.LG]*, 2020.

David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, May 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448.

Colin McDiarmid. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

Nikolaos Mourdoukoutas, Marco Federici, Georges Pantalos, Mark van der Wilk, and Vincent Fortuin. A Bayesian Approach to Invariant Deep Neural Networks. *arXiv:2107.09301 [cs, stat]*, July 2021.

Oskar Nordenfors and Axel Flinth. Ensembles provably learn equivariance through data augmentation. *arXiv:2410.01452v2 [cs.LG]*, 2025a.

Oskar Nordenfors and Axel Flinth. Data augmentation and regularization for learning group equivariance. In *2025 International Conference on Sampling Theory and Applications (SampTA)*, pp. 1–5, 2025b.

Oskar Nordenfors, Fredrik Ohlsson, and Axel Flinth. Optimization dynamics of equivariant and augmented neural networks. *Transactions on Machine Learning Research*, 2025.

Omri Puny, Matan Atzmon, Edward J Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022.

Tycho F.A. van der Ouderaa and Mark van der Wilk. Learning invariant weights in neural networks. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1992–2001. PMLR, 01–05 Aug 2022.

Soledad Villar, David W Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. *Advances in Neural Information Processing Systems*, 34:28848–28863, 2021.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pp. 23078–23091. PMLR, 2022.

Yuyang Wang, Ahmed A. A. Elhag, Navdeep Jaitly, Joshua M. Susskind, and Miguel Ángel Bautista. Swallowing the Bitter Pill: Simplified Scalable Conformer Generation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 50400–50418. PMLR, July 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747v2 [cs.LG]*, 2017.

A FROM DISCRETE TO CONTINUOUS COMPACT GROUPS

It is often the case in the geometric deep learning literature that virtually all results concerning finite groups can be generalized to a compact, continuous group – simply by replacing the average over the group with an integral over the group with respect to the *Haar measure* ν_G

$$\frac{1}{|G|}\sum_{g\in G} \quad \longrightarrow \quad \int_G \mathrm{d}\nu_G(g).$$

The Haar measure is the unique G-invariant probability measure on G – meaning that if $g \sim \nu_G$, also $hg \sim \nu_G$ for any $h \in G$. Note that for finite groups, the average over G already is the Haar average.

In order for us to do this, there is one minor obstruction: the definition of the augmented likelihood. Note that for a finite group, this is well-defined:

$$p(\mathcal{D}_{\text{aug}}\,|\theta) = \prod_{(x_i,y_i)\in\mathcal{D}_{\text{aug}}} p((x_i,y_i)\,|\theta) = \prod_{g\in G}\prod_{(x_i,y_i)\in\mathcal{D}} p((gx_i,gy_i)\,|\theta).$$

For an infinite group, the product above (as well as the augmented dataset) becomes infinite, and hard to interpret.

We can circumvent this by considering a sampled augmentation – we draw S i.i.d. elements $g_j \in G$ (according to ν_G) and form a randomly augmented dataset $\mathcal{D}_{S-\text{aug}} = \{(g_j x_i, g_j y_i)\}_{i=1,j=1}^{N_0,S}$. The likelihood of this data set is

$$p(\mathcal{D}_{S_{\text{aug}}}\,|\theta) = \prod_{(x_i,y_i)\in\mathcal{D}_{S-\text{aug}}} p((x_i,y_i)\,|\theta) = \prod_{j=1}^{S}\prod_{(x_i,y_i)\in\mathcal{D}} p((g_j x_i,g_j y_i)\,|\theta).$$

Now note that we can maximize this by minimizing the negative log-likelihood, a rescaled version of which becomes

$$-\frac{1}{S}\log p(\mathcal{D}_{\text{aug}}\,|\theta) = -\frac{1}{S}\sum_{j=1}^{S}\sum_{(x_i,y_i)\in\mathcal{D}} \log p((gx_i,gy_i)\,|\theta)\,.$$

As we increase the number of samples, this function converges towards

$$-\int_G \sum_{(x_i,y_i)\in\mathcal{D}} \log p((gx_i,gy_i)\,|\theta)\mathrm{d}\nu_G(g)\,.$$

If we define this function as $-\log p(\mathcal{D}_{\text{aug}}|\theta)$ in the infinite case, all of our results go through, with proofs more or less verbatim.

It should be noted that in practice we will always need to work on a sampled augmentation space, so there will be a discretization error with respect to the group. Also note that in this formulation, we draw the sample once and then train on the resulting dataset – we are not drawing new random group elements in each epoch, which often is done. To analyze this latter stochastic version requires other tools than we have used here, and we leave it to future work.

B Proof of Proposition 1

Proposition 1 (Invariant augmented likelihood). *Let $L_{\text{aug}}(\theta) := p(\mathcal{D}_{\text{aug}} \mid \theta)$ be the likelihood of the augmented dataset. Under the compatibility assumption equation 2 as well as the invariance assumption equation 7, the likelihood is G-invariant:*

$$L_{\text{aug}}(\rho(g)^{-1}\theta) = L_{\text{aug}}(\theta),\ \forall g \in G,\ \theta \in \Theta\,. \tag{8}$$

Proof. The compatibility of ρ_Θ equation 2 with the data representations implies that for any $g \in G$ and any input $x \in \mathcal{X}$,

$$f(gx;\theta) = gf(x;\rho(g)^{-1}\theta), \tag{21}$$

implies

$$\begin{aligned} p((gx,gy)\,|\,\theta) &= p(gy\,|\,gx,\theta)p(gx) = p(gy\,|\,f(gx;\theta))p(gx) && (22)\\ &= p(gy\,|\,gf(x;\rho(g)^{-1}\theta))p(gx) = p(y\,|\,f(x;\rho(g)^{-1}\theta))p(gx), && (23)\end{aligned}$$

or equivalently

$$p((x,y)\,|\,\rho(g)\theta) = p(g^{-1}y\,|\,f(g^{-1}x;\theta))p(x), \tag{24}$$

We can now make the calculation

$$\begin{aligned} L_{\text{aug}}(\rho(g)\theta) &= \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p((x,y)\mid\rho(g)\theta) = \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p(g^{-1}y\,|\,f(g^{-1}x;\theta))p(x) && (25)\\ &= \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p(g^{-1}y\,|\,f(g^{-1}x;\theta)) \cdot \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p(x) && (26)\\ &\overset{*}{=} \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p(y\,|\,f(x;\theta)) \cdot \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p(x) && (27)\\ &= \prod_{(x,y)\in\mathcal{D}_{\text{aug}}} p((x,y)\,|\,\theta) = L_{\text{aug}}(\theta), && (28)\end{aligned}$$

where (*) follows from the fact that $\mathcal{D}_{\text{aug}}$ is closed under group transformations: $h\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{aug}}$ for any $h \in G$. □

C Proof of Theorem 1

Theorem 1 (Closure under push-forward). *A regular minimal exponential family $\mathcal{Q}$ is closed under the push-forward operator $\mathcal{T}_g$ for all $g \in G$ if and only if for each g there exist $M_g \in \mathbb{R}^{k\times k}$, $d_g \in \mathbb{R}^k$, $b_g \in \mathbb{R}^k$ and $c_g \in \mathbb{R}$ such that for all $\theta \in \Theta$,*

$$T(\rho(g)^{-1}\theta) = M_g T(\theta) + d_g, \tag{10}$$

$$h(\rho(g)^{-1}\theta)\left|\det D(\rho(g)^{-1})(\theta)\right| = h(\theta)\exp(b_g^\top T(\theta) + c_g). \tag{11}$$

When these conditions hold, $\mathcal{T}_g \# q_\eta = q_{\eta_g}$ where $\eta_g = M_g^\top \eta + b_g$, and the transformation $\phi_g : \eta \mapsto M_g^\top \eta + b_g$ forms an affine action of G on the space of natural parameters H.

Proof. "⇐" This is a simple calculation

$$\begin{aligned} (T_g \# q_\eta)(\theta) &= h(\rho(g)^{-1}(\theta))\exp(\eta^\top T(\rho(g)^{-1}(\theta)) - A(\eta))\left|\det D(\rho(g)^{-1})(\theta)\right| && (29)\\ &= h(\theta)\exp(b_g^\top T(\theta) + c_g)\exp(\eta^\top(M_g T(\theta) + d_g) - A(\eta)) && (30)\\ &= h(\theta)\exp((M_g^\top\eta + b_g)^\top T(\theta) - (A(\eta) - d_g^\top\eta - c_g)) && (31)\\ &=: h(\theta)\exp(\eta_g^\top T(\theta) - A_g(\eta)) && (32)\end{aligned}$$

It remains to verify that $A_g(\eta) = A(\eta) - d_g^\top\eta - c_g$ is indeed the log-partition function A evaluated at η_g.

Using $\eta_g = M_g^\top \eta + b_g$, the affine condition $M_g T(\theta) = T(\rho(g)^{-1}\theta) - d_g$, and the base-measure condition $h(\theta)\exp(b_g^\top T(\theta)) = h(\rho(g)^{-1}\theta)\,|\det D(\rho(g)^{-1})(\theta)|\,e^{-c_g}$,

$$A(\eta_g) = \log \int h(\theta)\exp\big(\eta^\top M_g T(\theta) + b_g^\top T(\theta)\big)\,d\theta \tag{33}$$

$$= \log \int h(\rho(g)^{-1}\theta)\,e^{-c_g}\,\big|\det D(\rho(g)^{-1})(\theta)\big|\exp\big(\eta^\top (T(\rho(g)^{-1}\theta) - d_g)\big)\,d\theta \tag{34}$$

$$= -c_g - \eta^\top d_g + \log \int h(\rho(g)^{-1}\theta)\exp\big(\eta^\top T(\rho(g)^{-1}\theta)\big)\big|\det D(\rho(g)^{-1})(\theta)\big|\,d\theta \tag{35}$$

$$= -c_g - \eta^\top d_g + \log \int h(\theta')\exp\big(\eta^\top T(\theta')\big)\,d\theta' = A(\eta) - d_g^\top \eta - c_g, \tag{36}$$

where the second-to-last step substitutes $\theta' = \rho(g)^{-1}\theta$, for which $|\det D(\rho(g)^{-1})(\theta)|\,d\theta = d\theta'$. Hence $A_g(\eta) = A(\eta_g)$ and $\mathcal{T}_g \# q_\eta = q_{\eta_g} \in \mathcal{Q}$.

"$\Rightarrow$" Suppose that there exists a mapping $\phi_g : H \to H$ such that for any $\eta \in H$ and $\theta \in \Theta$,

$$q_\eta(\rho(g)^{-1}(\theta))\,\big|\det D(\rho(g)^{-1})(\theta)\big| = q_{\phi_g(\eta)}(\theta). \tag{37}$$

First, we take the logarithm on both sides and use the form of q_η:

$$\begin{aligned} &\log h(\rho(g)^{-1}(\theta)) + \eta^\top T(\rho(g)^{-1}(\theta)) - A(\eta) + \log\big|\det D(\rho(g)^{-1})(\theta)\big| \\ &\quad = \phi_g(\eta)^\top T(\theta) - A(\phi_g(\eta)) + \log h(\theta) \qquad (38) \\ \Rightarrow\quad &\eta^\top T(\rho(g)^{-1}(\theta)) + \log(h(\rho(g)^{-1}(\theta))\,\big|\det D(\rho(g)^{-1})(\theta)\big|) + A(\phi_g(\eta)) - A(\eta) \\ &\quad = \phi_g(\eta)^\top T(\theta) + \log h(\theta)\,. \qquad (39) \end{aligned}$$

Take any $\eta_1, \eta_2 \in H$, and subtract the equation equation 39 for η_2 from that for η_1:

$$\begin{aligned} &(\eta_1 - \eta_2)^\top T(\rho(g)^{-1}(\theta)) - (\phi_g(\eta_1) - \phi_g(\eta_2))^\top T(\theta) \\ &\quad = A(\phi_g(\eta_1)) - A(\eta_1) - A(\phi_g(\eta_2)) + A(\eta_2)\,. \end{aligned} \tag{40}$$

Observe that the right-hand side is independent of θ. Since the exponential family is minimal, $T(\theta)$ is linearly independent. To be precise, the components $\{1, T_1(\theta), \ldots, T_k(\theta)\}$ are linearly independent. Thus, both sides must be equal to some constant $C(\eta)$ independent of θ.

Denote $u = \eta_1 - \eta_2$ and $v = \phi_g(\eta_1) - \phi_g(\eta_2)$. We have:

$$u^\top T(\rho(g)^{-1}(\theta)) - v^\top T(\theta) = C(\eta)\,. \tag{41}$$

Since η_1, η_2 are arbitrary, u can take any value in $\mathbb{R}^k$. Thus, we can choose u to be each standard basis vector in $\mathbb{R}^k$. For each $j = 1, 2, \ldots, k$, we then get

$$T_j(\rho(g)^{-1}(\theta)) = \sum_{i=1}^{k} v_i^{(j)} T_i(\theta) + C_j\,. \tag{42}$$

Denoting $M_g = [v^{(1)}, v^{(2)}, \ldots, v^{(k)}]^\top$ and $d_g = (C_1, C_2, \ldots, C_k)^\top$, we conclude:

$$T(\rho(g)^{-1}(\theta)) = M_g T(\theta) + d_g\,. \tag{43}$$

Next, we substitute this back into equation 39:

$$\begin{aligned} &\eta^\top (M_g T(\theta) + d_g) + \log(h(\rho(g)^{-1}(\theta))\,\big|\det D(\rho(g)^{-1})(\theta)\big|) - A(\eta) \\ &\quad = \log h(\theta) + \phi_g(\eta)^\top T(\theta) - A(\phi_g(\eta)) \qquad (44) \\ \Rightarrow\quad &(\eta^\top M_g - \phi_g(\eta)^\top) T(\theta) + \log(h(\rho(g)^{-1}(\theta))\,\big|\det D(\rho(g)^{-1})(\theta)\big|) - \log h(\theta) + \eta^\top d_g \\ &\quad = A(\eta) - A(\phi_g(\eta))\,. \qquad (45) \end{aligned}$$

Again, since the right-hand side is independent of θ and only depends on η, we take any two $\theta_1, \theta_2 \in \Theta$ and subtract equation 45 for θ_2 from that for θ_1 to obtain:

$$(\eta^\top M_g - \phi_g(\eta)^\top)(T(\theta_1) - T(\theta_2)) + \log \frac{h(\rho(g)^{-1}(\theta_1)) \left|\det D(\rho(g)^{-1})(\theta_1)\right|}{h(\rho(g)^{-1}(\theta_2)) \left|\det D(\rho(g)^{-1})(\theta_2)\right|} - \log \frac{h(\theta_1)}{h(\theta_2)} = 0. \tag{46}$$

Since the logarithmic terms do not depend on η, we can write this as

$$(\eta^\top M_g - \phi_g(\eta)^\top)(T(\theta_1) - T(\theta_2)) = B(\theta_1, \theta_2)\,, \tag{47}$$

for some numbers $B(\theta_1, \theta_2)$. Now, again, since the family is minimal, the components $T(\theta)$ are linearly independent. This implies that we can choose a value for θ_1 and k different values for θ_2, such that the matrix $(T(\theta_1) - T(\theta_2^\ell))_{\ell=1}^k \in \mathbb{R}^{k,k}$ becomes invertible. This yields a linear system of equations in the vector $\eta^\top M_g - \phi_g(\eta)^\top$, which has a unique solution which we denote by $-b_g^\top$:

$$-b_g^\top := \eta^\top M_g - \phi_g(\eta)^\top\,. \tag{48}$$

Since the right hand side above is independent of θ, the same must be true for b_g, so that we get

$$\phi_g(\eta) = M_g^\top \eta + b_g\,. \tag{49}$$

Now plug equation 48 back into equation 45 to obtain

$$-b_g^\top T(\theta) + \log(h(\rho(g)^{-1}(\theta)) \left|\det D(\rho(g)^{-1})(\theta)\right|) - \log h(\theta) = A(\eta) - A(\phi_g(\eta)) - \eta^\top d_g\,. \tag{50}$$

Since the right-hand side is independent of θ, the left-hand side must also be a constant independent of θ. Since the left-hand side is independent of η, this constant is also independent of η. Denoting the constant by c_g, we have:

$$\begin{aligned} &\log(h(\rho(g)^{-1}(\theta)) \left|\det D(\rho(g)^{-1})(\theta)\right|) = \log h(\theta) + b_g^\top T(\theta) + c_g \\ \implies \quad &h(\rho(g)^{-1}(\theta)) \left|\det D(\rho(g)^{-1})(\theta)\right| = h(\theta) \exp(b_g^\top T(\theta) + c_g) \end{aligned} \tag{51}$$

Finally, since the bracketed combination in equation 50 equals the constant c_g established above, combining equation 50 with equation 48 gives

$$A(\phi_g(\eta)) = A(\eta) - d_g^\top \eta - c_g. \tag{52}$$

It remains to show that $g \mapsto \phi_g$ is an affine action of G on H. Each $\phi_g(\eta) = M_g^\top \eta + b_g$ is affine, and M_g is invertible: by minimality the rows $\{v^{(j)}\}$ chosen above are linearly independent, so $M_g = [v^{(1)}, \ldots, v^{(k)}]^\top$ is nonsingular.

For the action property, the identity $e \in G$ has $\rho(e) = \mathrm{id}$, so

$$\mathcal{T}_e \# q_\eta(\theta) = q_\eta(\rho(e)^{-1}\theta) \left|\det D(\rho(e)^{-1})(\theta)\right| = q_\eta(\theta), \tag{53}$$

giving $\phi_e = \mathrm{id}$. For $g_1, g_2 \in G$, the chain rule yields

$$\begin{aligned} \mathcal{T}_{g_1 g_2} \# q_\eta(\theta) &= q_\eta\big(\rho(g_2^{-1})\rho(g_1^{-1})\theta\big) \left|\det D(\rho(g_2^{-1}))(\rho(g_1^{-1})\theta)\right| \left|\det D(\rho(g_1^{-1}))(\theta)\right| \\ &= q_{\phi_{g_2}(\eta)}\big(\rho(g_1^{-1})\theta\big) \left|\det D(\rho(g_1^{-1}))(\theta)\right| = q_{\phi_{g_1}(\phi_{g_2}(\eta))}(\theta), \end{aligned} \tag{54}$$

Thus, we have $\phi_{g_1 g_2}(\eta) = \phi_{g_1}(\phi_{g_2}(\eta))$. Since η is arbitrary, we conclude that $\phi_{g_1 g_2} = \phi_{g_1} \circ \phi_{g_2}$. Hence $g \mapsto \phi_g$ is an affine group action of G on H, which completes the proof. □

D Proof of Theorem 2

Theorem 2 (Equivariance of VI)**.** *Assumptions 1 and 2 imply the following for variational inference using invariant likelihoods and closed exponential families:*

(i) ***Loss invariance.*** *The loss $\mathcal{L}_\beta$ is invariant under the natural parameter transformation:*

$$\mathcal{L}_\beta(\eta) = \mathcal{L}_\beta(\phi_g(\eta)), \quad \forall g \in G,\, \eta \in H. \tag{14}$$

(ii) ***Gradient equivariance and update commutativity.*** *The gradient of the loss $\mathcal{L}_\beta$ satisfies*

$$\nabla_\eta \mathcal{L}_\beta(\phi_g(\eta)) = M_g^{-1} \nabla_\eta \mathcal{L}_\beta(\eta), \quad \forall g \in G,\, \eta \in H. \tag{15}$$

Consequently, for M_g orthogonal, gradient descent $\mathcal{U}(\eta) = \eta - \alpha \nabla_\eta \mathcal{L}_\beta(\eta)$ commutes with ϕ_g:

$$\mathcal{U}(\phi_g(\eta)) = \phi_g(\mathcal{U}(\eta)). \tag{16}$$

(iii) ***Invariant subspace is preserved.*** *For M_g orthogonal, the equivariant subspace H_G is preserved by gradient updates: if $\eta^{(0)} \in H_G$, then $\eta^{(t)} \in H_G$ for all $t \geq 0$.*

(iv) ***Optimal solutions inherit symmetry.*** *The set of optimal natural parameters $H^* = \arg\min_\eta \mathcal{L}_\beta(\eta)$ is closed under ϕ_g for all $g \in G$. If the optimum is unique, then $\phi_g(\eta^*) = \eta^*$.*

Proof. We prove the four claims in turn.

(i) Loss invariance. Assumption 1 implies that $q_{\eta_0} = q_{\phi_g(\eta_0)}$. Consequently,

$$\begin{aligned}
D_{\mathrm{KL}}(q_{\phi_g(\eta)}(\theta) \| q_{\eta_0}(\theta)) &= D_{\mathrm{KL}}(q_{\phi_g(\eta)}(\theta) \| q_{\phi_g(\eta_0)}(\theta)) && (55)\\
&= D_{\mathrm{KL}}(q_\eta(\rho(g)^{-1}\theta) \| q_{\eta_0}(\rho(g)^{-1}\theta)) && (56)\\
&= \int q_\eta(\rho(g)^{-1}\theta) \log \frac{q_\eta(\rho(g)^{-1}\theta)}{q_{\eta_0}(\rho(g)^{-1}\theta)} d(\rho(g)^{-1}\theta) && (57)\\
&= \int q_\eta(\theta) \log \frac{q_\eta(\theta)}{q_{\eta_0}(\theta)} |\det D(\rho(g)^{-1})(\theta)| d\theta && (58)\\
&= \int q_\eta(\theta) \log \frac{q_\eta(\theta)}{q_{\eta_0}(\theta)} d\theta = D_{\mathrm{KL}}(q_\eta(\theta) \| q_{\eta_0}(\theta)), && (59)
\end{aligned}$$

where the third step is the transformation formula and the fourth uses the volume-preserving condition $|\det D(\rho(g)^{-1})(\theta)| = 1$ (Assumption 2).

We now use the invariance of the likelihood proven in Proposition 1 to get

$$\begin{aligned}
\mathbb{E}_{q_{\phi_g(\eta)}(\theta)}[\log p(\mathcal{D}_{\mathrm{aug}} \mid \theta)] &= \int q_{\phi_g(\eta)}(\theta) \log p(\mathcal{D}_{\mathrm{aug}} \mid \theta) d\theta && (60)\\
&= \int q_\eta(\rho(g)^{-1}\theta) \log p(\mathcal{D}_{\mathrm{aug}} \mid \theta) d\theta && (61)\\
&= \int q_\eta(\theta) \log p(\mathcal{D}_{\mathrm{aug}} \mid \rho(g)(\theta)) |\det D(\rho(g))(\theta)| d\theta && (62)\\
&= \int q_\eta(\theta) \log p(\mathcal{D}_{\mathrm{aug}} \mid \theta) d\theta = \mathbb{E}_{q_\eta(\theta)}[\log p(\mathcal{D}_{\mathrm{aug}} \mid \theta)]. && (63)
\end{aligned}$$

We again used the transformation formula and the volume-preserving conditions. Combining the two terms, we get

$$\mathcal{L}_\beta(\phi_g(\eta)) = \tfrac{1}{N} \mathbb{E}_{q_\eta(\theta)}[-\log p(\mathcal{D}_{\mathrm{aug}} \mid \theta)] + \beta D_{\mathrm{KL}}(q_\eta(\theta) \,\|\, q_{\eta_0}(\theta)) = \mathcal{L}_\beta(\eta). \tag{64}$$

(ii) Gradient equivariance and update commutativity. Differentiating the identity $\mathcal{L}_\beta(\phi_g(\eta)) = \mathcal{L}_\beta(\eta)$ with respect to η and using the chain rule together with $\phi_g(\eta) = M_g^\top \eta + b_g$ yields

$$M_g \nabla_\eta \mathcal{L}_\beta\big|_{\phi_g(\eta)} = \nabla_\eta \mathcal{L}_\beta(\eta), \tag{65}$$

so that

$$\nabla_\eta \mathcal{L}_\beta(\phi_g(\eta)) = M_g^{-1} \nabla_\eta \mathcal{L}_\beta(\eta). \tag{66}$$

When M_g is orthogonal, $M_g^{-1} = M_g^\top$, we get

$$\begin{aligned}
\mathcal{U}(\phi_g(\eta)) = \phi_g(\eta) - \alpha \nabla_\eta \mathcal{L}_\beta(\phi_g(\eta)) &= M_g^\top \eta + b_g - \alpha M_g^{-1} \nabla_\eta \mathcal{L}_\beta(\eta) && (67)\\
= M_g^\top (\eta - \alpha \nabla_\eta \mathcal{L}_\beta(\eta)) + b_g &= \phi_g(\mathcal{U}(\eta)). && (68)
\end{aligned}$$

(iii) Invariant subspace is preserved. Let $\eta \in H_G$, i.e., $\phi_g(\eta) = \eta$ for all $g \in G$. By the update commutativity equation 16 from (ii),

$$\phi_g(\mathcal{U}(\eta)) = \mathcal{U}(\phi_g(\eta)) = \mathcal{U}(\eta) \quad \text{for all } g \in G, \tag{69}$$

so $\mathcal{U}(\eta) \in H_G$. By induction on t, if $\eta^{(0)} \in H_G$ then $\eta^{(t)} \in H_G$ for all $t \geq 0$.

(iv) Optimal solutions inherit symmetry. Let $\eta^* \in H^* = \arg\min_\eta \mathcal{L}_\beta(\eta)$. By (i),

$$\mathcal{L}_\beta(\phi_g(\eta^*)) = \mathcal{L}_\beta(\eta^*) = \min_\eta \mathcal{L}_\beta(\eta), \tag{70}$$

so $\phi_g(\eta^*) \in H^*$ for all $g \in G$; i.e., H^* is closed under ϕ_g. If the minimizer is unique, then $\phi_g(\eta^*) = \eta^*$ for all $g \in G$, meaning $\eta^* \in H_G$. □

E Proof of Theorem 3

Theorem 3 (Prior-independent convergence)**.** *Assume Assumption 2, and that the likelihood is invariant. Let* $R^*_{\mathrm{aug}} := \inf_\eta R_{\mathrm{aug}}(\eta)$ *and* $\eta^*_\beta(p_0) \in \arg\min_\eta \mathcal{L}_\beta(\eta; p_0)$*. We then have*

$$R_{\mathrm{aug}}(\eta^*_\beta(p_0)) \leq R^*_{\mathrm{aug}} + \beta \cdot C(p_0), \tag{17}$$

for any prior p_0*, where* $C(p_0) := \inf_{\eta^* \in H^*_R} D_{\mathrm{KL}}(q_{\eta^*} \| p_0)$ *and* $H^*_R := \arg\min_\eta R_{\mathrm{aug}}(\eta)$ *is the set of risk-optimal natural parameters. As* $\beta \to 0$*, all minimizers achieve* R^*_{aug} *regardless of the prior.*

Proof. Note that the log-likelihood does not depend on the prior. Hence, using the same calculation as in the proof of Theorem 2 (i), we show that R_{aug} is G-invariant on the natural parameter space:

$$R_{\mathrm{aug}}(\phi_g(\eta)) = R_{\mathrm{aug}}(\eta), \quad \forall g \in G,\ \eta \in H. \tag{71}$$

In particular, the set $H^*_R := \arg\min_\eta R_{\mathrm{aug}}(\eta)$ is closed under ϕ_g for all $g \in G$.

Fix any $\eta^*_R \in H^*_R$. By the optimality of $\eta^*_\beta(p_0)$ for $\mathcal{L}_\beta(\cdot; p_0)$,

$$R_{\mathrm{aug}}(\eta^*_\beta(p_0)) + \beta D_{\mathrm{KL}}(q_{\eta^*_\beta(p_0)} \| p_0) \leq R_{\mathrm{aug}}(\eta^*_R) + \beta D_{\mathrm{KL}}(q_{\eta^*_R} \| p_0). \tag{72}$$

Dropping the non-negative KL term on the left and using $R_{\mathrm{aug}}(\eta^*_R) = R^*_{\mathrm{aug}}$,

$$R_{\mathrm{aug}}(\eta^*_\beta(p_0)) \leq R^*_{\mathrm{aug}} + \beta D_{\mathrm{KL}}(q_{\eta^*_R} \| p_0). \tag{73}$$

Taking the infimum over $\eta^*_R \in H^*_R$ on the right yields the stated bound

$$R_{\mathrm{aug}}(\eta^*_\beta(p_0)) \leq R^*_{\mathrm{aug}} + \beta \cdot C(p_0). \tag{74}$$

Combined with $R_{\mathrm{aug}}(\eta^*_\beta(p_0)) \geq R^*_{\mathrm{aug}}$ (by the definition of R^*_{aug}), this gives

$$R^*_{\mathrm{aug}} \leq R_{\mathrm{aug}}(\eta^*_\beta(p_0)) \leq R^*_{\mathrm{aug}} + \beta \cdot C(p_0). \tag{75}$$

For any fixed p_0 with $C(p_0) < \infty$, sending $\beta \to 0$ (equivalently $N \to \infty$, since $\beta = 1/N$) yields $R_{\mathrm{aug}}(\eta^*_\beta(p_0)) \to R^*_{\mathrm{aug}}$, independent of p_0. □

F Proof of Theorem 4

Before stating the proof, we recall the definitions of the empirical and Monte Carlo equivariance defects.

Definition 1 (Empirical and Monte Carlo equivariance defects)**.** *Given a dataset* $\{x_i\}_{i=1}^{N_0}$*, the empirical equivariance defect is*

$$\widetilde{\Delta}^{\mathrm{eq}}_F(\eta) := \frac{1}{N_0} \sum_{i=1}^{N_0} \frac{1}{|G|} \sum_{g \in G} \|F_\eta(gx_i) - gF_\eta(x_i)\|^2 . \tag{76}$$

Its Monte Carlo approximation using T *posterior samples is*

$$\widehat{\Delta}^{\mathrm{eq}}_F(\eta) := \frac{1}{N_0} \sum_{i=1}^{N_0} \frac{1}{|G|} \sum_{g \in G} \left\| \widehat{F}_\eta(gx_i) - g\widehat{F}_\eta(x_i) \right\|^2, \quad \widehat{F}_\eta(x) := \frac{1}{T} \sum_{t=1}^{T} f(x; \theta^{(t)}), \tag{77}$$

where $\{\theta^{(t)}\}_{t=1}^T$ *are i.i.d. samples from* $q_\eta(\theta)$*.*

We now restate Theorem 4 and prove it thereafter.

Theorem 4 (Complexity for G_ε-equivariance)**.** *Suppose $f(x;\theta)$ is uniformly bounded in $x \in \mathcal{X}$ and q_η-a.s. in θ, and that the G-action on $\mathcal{Y}$ is orthogonal. Then there exist constants C_{data} and C_{weight} such that, for any fixed η, with probability at least $1-\delta$ over the N_0 i.i.d. inputs $x_i \sim P_{\mathcal{X}}$ and the T i.i.d. posterior samples $\theta^{(t)} \sim q_\eta$, the Monte Carlo equivariance defect satisfies*

$$\widehat{\Delta}_F^{\mathrm{eq}}(\eta) \le \Delta_F^{\mathrm{eq}}(\eta) + C_{\mathrm{data}}\sqrt{\frac{\log(2/\delta)}{N_0}} + C_{\mathrm{weight}}\sqrt{\frac{\log(2/\delta)}{T}}\,. \tag{18}$$

In other words, for any $\varepsilon_0, \varepsilon_{\mathrm{data}}, \varepsilon_{\mathrm{weight}} > 0$, if $\Delta_F^{\mathrm{eq}}(\eta) < \varepsilon_0$, it suffices to choose

$$N_0 \ge \frac{C_{\mathrm{data}}^2}{\varepsilon_{\mathrm{data}}^2}\log\left(\frac{2}{\delta}\right), \quad T \ge \frac{C_{\mathrm{weight}}^2}{\varepsilon_{\mathrm{weight}}^2}\log\left(\frac{2}{\delta}\right) \tag{19}$$

in order for $\widehat{\Delta}_F^{\mathrm{eq}}(\eta) < \varepsilon_0 + \varepsilon_{\mathrm{data}} + \varepsilon_{\mathrm{weight}}$ with probability at least $1-\delta$.

Proof. Suppose that there exists $M > 0$ such that $\|f(x;\theta)\| \le M$ for all $x \in \mathcal{X}$ and q_η-a.s. θ. Let us begin by noting that a simple application of Jensen's inequality shows that for each x,

$$\|F_\eta(x)\| = \|\mathbb{E}_{q_\eta}[f(x;\theta)]\| \le \mathbb{E}_{q_\eta}[\|f(x;\theta)\|] \le M, \tag{78}$$

$$\|\widehat{F}_\eta(x)\| = \frac{1}{T}\left\|\sum_{t=1}^{T} f(x;\theta^{(t)})\right\| \le \frac{1}{T}\sum_{t=1}^{T}\left\|f(x;\theta^{(t)})\right\| \le M. \tag{79}$$

Using the triangle inequality and orthogonality of the group action, this yields

$$\|F_\eta(gx) \pm gF_\eta(x)\| \le 2M, \quad \|\widehat{F}_\eta(gx) \pm g\widehat{F}_\eta(x)\| \le 2M \tag{80}$$

Hence, the quantities we need to estimate are essentially sums of i.i.d. bounded terms. This makes standard concentration inequalities viable. We now carry out the details.

Step 1: Dataset-level concentration. Define

$$Z_i := \frac{1}{|G|}\sum_{g\in G}\|F_\eta(gx_i) - gF_\eta(x_i)\|^2. \tag{81}$$

These terms are bounded, $Z_i \in [0, 4M^2]$, their common expected value is $\mathbb{E}[Z_i] = \Delta_F^{\mathrm{eq}}(\eta)$, and their mean is $\frac{1}{N_0}\sum_{i=1}^{N_0} Z_i = \widetilde{\Delta}_F^{\mathrm{eq}}(\eta)$. Applying Hoeffding's inequality, we get for any $t > 0$,

$$\mathbb{P}\Big(\widetilde{\Delta}_F^{\mathrm{eq}}(\eta) - \Delta_F^{\mathrm{eq}}(\eta) \ge t\Big) \le \exp\left(-\frac{2N_0t^2}{(4M^2)^2}\right) = \exp\left(-\frac{N_0t^2}{8M^4}\right). \tag{82}$$

Setting the right-hand side to $\delta/2$ and solving for t, we get that with probability at least $1-\delta/2$,

$$\widetilde{\Delta}_F^{\mathrm{eq}}(\eta) \le \Delta_F^{\mathrm{eq}}(\eta) + C_{\mathrm{data}}\sqrt{\frac{\log(2/\delta)}{N_0}}, \tag{83}$$

where $C_{\mathrm{data}} := 2\sqrt{2}\,M^2$.

Step 2: Monte Carlo concentration. The estimator $\widehat{\Delta}_F^{\mathrm{eq}}(\eta)$ differs from $\widetilde{\Delta}_F^{\mathrm{eq}}(\eta)$ only through replacing the posterior expectation $F_\eta(x)$ by the empirical average $\widehat{F}_\eta(x)$. Note that in contrast to above, $\mathbb{E}(\widehat{\Delta}_F^{\mathrm{eq}})$ is not equal to $\widetilde{\Delta}_F^{\mathrm{eq}}(\eta)$. There instead is a small bias term. We first control that bias, and then bound $\widehat{\Delta}_F^{\mathrm{eq}}(\eta)$ around its mean via McDiarmid's inequality.

Bias. Writing $D := F_\eta(gx) - gF_\eta(x)$ and $\widehat{D} := \widehat{F}_\eta(gx) - g\widehat{F}_\eta(x)$, we have

$$\mathbb{E}\Big[\|\widehat{D}\|^2\Big] - \|D\|^2 = \mathrm{tr}\Big(\mathrm{Cov}(\widehat{D})\Big) \le \frac{4M^2}{T}, \tag{84}$$

since $\widehat{D}$ is an average of T i.i.d. terms each bounded in norm by $2M$. Averaging over G and the data samples yields

$$\mathbb{E}[\widehat{\Delta}_F^{\mathrm{eq}}(\eta)] - \widetilde{\Delta}_F^{\mathrm{eq}}(\eta) \le \frac{C_{\mathrm{bias}}}{T} \le \frac{C_{\mathrm{bias}}}{\sqrt{T}}, \tag{85}$$

with $C_{\mathrm{bias}} = 8M^2$.

Bounded differences. Replacing a single sample $\theta^{(s)}$ by $\theta'^{(s)}$ changes any $\widehat{F}_\eta(x)$ by at most $\frac{1}{T}\|f(x;\theta^{(s)}) - f(x;\theta'^{(s)})\| \le 2M/T$. By the triangle inequality, the same replacement changes every $\widehat{F}_\eta(gx_i) - g\widehat{F}_\eta(x_i)$ less than $4M/T$. Using the identity $\|a\|^2 - \|b\|^2 = \langle a-b, a+b\rangle$ together with bounds $\|a\| = \|\widehat{F}_\eta(gx_i) - g\widehat{F}_\eta(x_i)\| \le 2M$, $\|b\| = \|\widehat{F}'_\eta(gx_i) - g\widehat{F}'_\eta(x_i)\| \le 2M$ from equation 80, we get

$$\left| \|\widehat{F}_\eta(gx_i) - g\widehat{F}_\eta(x_i)\|^2 - \|\widehat{F}'_\eta(gx_i) - g\widehat{F}'_\eta(x_i)\|^2 \right| \le \frac{4M}{T} \cdot 4M = \frac{16M^2}{T}. \tag{86}$$

Averaging over i and g preserves this bound, so $\widehat{\Delta}_F^{\mathrm{eq}}(\eta)$ satisfies the bounded-differences condition with constants $c_s = 16M^2/T$ for each $s = 1, \dots, T$.

McDiarmid's inequality (McDiarmid, 1989; Doob, 1940) states that any function Φ of independent random variables satisfies

$$\mathbb{P}\big(\mathbb{E}[\Phi] - \Phi \ge t\big) \le \exp\left(-\frac{2t^2}{\sum_{s=1}^T c_s^2}\right), \tag{87}$$

where c_s is the bounded-difference constant of Φ in its s-th argument, i.e. substituting the s-th argument of Φ with a different value changes Φ by at most c_s.

Substituting $\Phi = \widehat{\Delta}_F^{\mathrm{eq}}(\eta)$ and $c_s = 16M^2/T$, we have $\sum_{s=1}^T c_s^2 = T \cdot (16M^2/T)^2 = 256M^4/T$. Hence,

$$\mathbb{P}\Big(\mathbb{E}[\widehat{\Delta}_F^{\mathrm{eq}}(\eta)] - \widehat{\Delta}_F^{\mathrm{eq}}(\eta) \ge t\Big) \le \exp\left(-\frac{Tt^2}{128M^4}\right). \tag{88}$$

Setting the right-hand side to $\delta/2$ and solving for t yields, with probability at least $1 - \delta/2$,

$$\mathbb{E}[\widehat{\Delta}_F^{\mathrm{eq}}(\eta)] - \widehat{\Delta}_F^{\mathrm{eq}}(\eta) \le \frac{16M^2}{\sqrt{2T}}\sqrt{\log(2/\delta)} = C_{\mathrm{mc}}\sqrt{\frac{\log(2/\delta)}{T}}, \tag{89}$$

where $C_{\mathrm{mc}} = 16M^2/\sqrt{2} = 8\sqrt{2}\,M^2$.

Step 3: Combining deviation and bias. We now combine the above to get

$$\widehat{\Delta}_F^{\mathrm{eq}}(\eta) = \Delta_F^{\mathrm{eq}}(\eta) + \underbrace{(\widetilde{\Delta}_F^{\mathrm{eq}}(\eta) - \Delta_F^{\mathrm{eq}}(\eta))}_{equation\ 83} + \underbrace{(\mathbb{E}[\widehat{\Delta}_F^{\mathrm{eq}}(\eta)] - \widetilde{\Delta}_F^{\mathrm{eq}}(\eta))}_{equation\ 85} + \underbrace{(\widehat{\Delta}_F^{\mathrm{eq}}(\eta) - \mathbb{E}[\widehat{\Delta}_F^{\mathrm{eq}}(\eta)])}_{equation\ 89} \tag{90}$$

$$\le \Delta_F^{\mathrm{eq}}(\eta) + C_{\mathrm{data}}\sqrt{\frac{\log(2/\delta)}{N_0}} + C_{\mathrm{bias}}\sqrt{\frac{1}{T}} + C_{\mathrm{mc}} \cdot \sqrt{\frac{\log(2/\delta)}{T}}, \tag{91}$$

which yields the first claim with $C_{\mathrm{weight}} = C_{\mathrm{mc}} + C_{\mathrm{bias}} \cdot \sqrt{1/\log(2)}$.

Step 4: Sample requirements. For the decomposition $\varepsilon = \varepsilon_0 + \varepsilon_{\mathrm{data}} + \varepsilon_{\mathrm{weight}} + \varepsilon_{\mathrm{bias}}$, requiring each term in the bound to be at most its allotted budget gives

$$C_{\mathrm{data}}\sqrt{\frac{\log(2/\delta)}{N_0}} \le \varepsilon_{\mathrm{data}}, \quad C_{\mathrm{weight}}\sqrt{\frac{\log(2/\delta)}{T}} \le \varepsilon_{\mathrm{weight}}. \tag{92}$$

Squaring and solving for N_0 and T yields the stated requirements. □

G GEOMETRIC AVERAGING STAYS IN $\mathcal{Q}$

Lemma 1 (Geometric mean invariance and membership)**.** *Suppose $\mathcal{Q}$ is closed under all push-forwards $\mathcal{T}_g$ as in Theorem 1, i.e., for each $g \in G$ and any $\eta \in H$,*

$$(\mathcal{T}_g \# q_\eta)(\theta) = h(\theta) \exp\big(\eta_g^\top T(\theta) - A_g(\eta)\big), \tag{93}$$

for some $\eta_g \in H$. Then the geometric mean

$$\tilde{q}(\theta) \propto \left[\prod_{g\in G} \Big(\mathcal{T}_g \# q_\eta\Big)(\theta)\right]^{\frac{1}{|G|}} \tag{94}$$

is invariant under the group action and can be written as an exponential-family member with the same base measure h and sufficient statistic T with natural parameter

$$\tilde{\eta} = \frac{1}{|G|} \sum_{g\in G} \eta_g. \tag{95}$$

Proof. Using closure under push-forward, for each g,

$$(\mathcal{T}_g \# q_\eta)(\theta) = h(\theta) \exp\big(\eta_g^\top T(\theta) - A(\eta_g)\big). \tag{96}$$

Taking the $\frac{1}{|G|}$-power and multiplying over all $g \in G$ yields

$$\left[\prod_{g\in G} \Big(\mathcal{T}_g \# q_\eta\Big)(\theta)\right]^{\frac{1}{|G|}} = h(\theta) \exp\Big(\frac{1}{|G|} \sum_{g\in G} \eta_g^\top T(\theta) - \frac{1}{|G|} \sum_{g\in G} A(\eta_g)\Big), \tag{97}$$

which after re-normalization yields a distribution in the standard exponential-family form with the stated $\tilde{\eta}$. Invariance follows because for any $g_0 \in G$, the set $\{\mathcal{T}_g \# q\}_{g\in G}$ is permuted under $\mathcal{T}_{g_0}$ and the product over G (with equal exponents) is unchanged. Hence $\tilde{q}$ is invariant and belongs to $\mathcal{Q}$. □

H THE TWO ORBIT AVERAGING STRATEGIES

In this section, we show how geometric averaging and projection naturally arise as two instances of orbit averaging. First, let us present in detail the canonical strategy of defining parameter space representations discussed in the main paper. Remember that a classical neural network (e.g. an MLP) consists of layers θ_ℓ of affine maps, whose domains and co-domains are vector spaces, say $\theta_\ell : \mathcal{V}_\ell \to \mathcal{V}_{\ell+1}$, with $\mathcal{V}_0 = \mathcal{X}$ and $\mathcal{V}_L = \mathcal{Y}$. One now introduces representations ρ_ℓ on each $\mathcal{V}_\ell$, with $\rho_0 = \rho_\mathcal{X}$ and $\rho_L = \rho_\mathcal{Y}$ to keep the construction compatible with the given in- and output representations, and defines

$$\rho_\ell(g)(\theta_\ell) = \rho_{\mathcal{V}_{\ell+1}}(g) \circ \theta_\ell \circ \rho_{\mathcal{V}_\ell}(g)^{-1} \tag{98}$$

If the activation functions are equivariant with respect to the intermediate representation, the above definition always yields a representation satisfying equation 2 – see Nordenfors et al. (2025).

For the networks that we are considering in Section 3.4, the input and intermediate spaces consist of *channels of images*. That is, for a space $\mathcal{I}$ of images, we set

$$\mathcal{V}_\ell = \mathcal{I}^{n_\ell}. \tag{99}$$

On the space $\mathcal{I}$, C_4 acts through a rotation operator $R : \mathcal{I} \to \mathcal{I}$. We can extend this to $\mathcal{V}_\ell$ in several ways. First, we can rotate channel-wise:

$$(\rho^{\text{ch}}_{\mathcal{V}_\ell}(i)v)_k = R^i v_k. \tag{100}$$

Alternatively, we can additionally act via 'rolling' along the channel dimension. More concretely, one divides the channels into groups of four, and then defines

$$(\rho^{\text{gcnn}}_{\mathcal{V}_\ell}(i)v)_{4n+k} = R^i v_{4n+(k+i \bmod 4)}. \tag{101}$$

The second representation is the one that leads to GCNNs.

Let us now apply equation 98 to a convolutional operator,

$$(\theta_\ell v)_k = \sum_j \eta^{kj} * v_j. \tag{102}$$

Note that rotation in the image space and convolutions are compatible, in the following sense:

$$(R\eta) * (Rv) = R(\eta * v), \tag{103}$$

Geometric average. For the channel-wise representation, we obtain:

$$(\rho_\ell^{\mathrm{ch}}(i)v)_k = \rho_{\mathcal{V}_{\ell+1}}^{\mathrm{ch}}(i)([\theta_\ell \circ \rho_{\mathcal{V}_\ell}^{\mathrm{ch}}(i)^{-1}]v)_k = R^i \sum_j \eta^{kj} * (\rho_{\mathcal{V}_\ell}^{\mathrm{ch}}(i)^{-1}v)_j \tag{104}$$

$$= R^i \sum_j \eta^{kj} * (R^{-i}v_j) = \sum_j (R^i\eta^{kj}) * v_j. \tag{105}$$

That is, ρ^{ch} acts via rotating each filter individually: $(\eta^{kj})_{k,j} \to (R^i\eta^{kj})_{k,j}$. Correspondingly, the orbit average averages all four rotations of each filter, yielding a filter-wise symmetrization.

Projection. For the GCNN-representation, let us restrict the calculation to the filters between two four-blocks, i.e. $n = 0$. We get

$$(\rho_\ell^{\mathrm{gcnn}}(i)v)_k = \rho_{\mathcal{V}_{\ell+1}}^{\mathrm{gcnn}}(i)([\theta_\ell \circ \rho_{\mathcal{V}_\ell}^{\mathrm{gcnn}}(i)^{-1}]v)_k = R^i \sum_j \eta^{(k+i \bmod 4),j} * (\rho_{\mathcal{V}_\ell}^{\mathrm{gcnn}}(i)^{-1}v)_j \tag{106}$$

$$= R^i \sum_j \eta^{(k+i \bmod 4),j} * (R^{-i}v_{j-i \bmod 4}) = \sum_j R^i\eta^{(k+i \bmod 4),(j+i \bmod 4)} * v_j. \tag{107}$$

Thus, $\rho_\ell^{\mathrm{gcnn}}$ acts through both rotating and rolling the filters along the diagonals of each four-by-four block. The resulting average operation will calculate an average of rotated filters along each such diagonal, and then insert it into the block in rotated versions, giving 4×4-blocks as in Figure 2.

I Expansion along orbit: filter arrangement ablation

Let us now discuss the orbit expansion strategy. The idea is to construct a set of equivariant filters, starting from a smaller set of generic filters. To keep the exposition light, we start with a single filter η and expand it to a single four-times-four block – the construction naturally generalizes to filter banks.

Since there are several notions of equivariance, there are different ways to achieve the equivariance of the larger filterbank. In Appendix H, we have already derived conditions for equivariance when both the input and output representations are equal to ρ^{ch} and ρ^{gcnn}, respectively. What are the conditions when the representations are unequal? Let us say that the representation on the layer $\mathcal{V}_1$ is ρ^{gcnn} and the one on $\mathcal{V}_2$ is ρ^{ch}. The lifted representation then becomes

$$(\rho_1(i)v)_k = \rho_{\mathcal{V}_2}^{\mathrm{ch}}(i)([\theta_\ell \circ \rho_{\mathcal{V}_1}^{\mathrm{gcnn}}(i)^{-1}]v)_k = R^i \sum_j \eta^{kj} * (\rho_{\mathcal{V}_1}^{\mathrm{gcnn}}(i)^{-1}v)_j$$

$$= R^i \sum_j \eta^{kj} * (R^{-i}v_{j-i}) = \sum_j (R^i\eta^{k,j+i}) * v_j.$$

Consequently, a layer is equivariant if and only if

$$R^i\eta^{k,j+i} = \eta^{k,j} \text{ for all } i, j, k. \tag{108}$$

This means that within each row, the filters should be rotations of each other. In Figure 5, we plot three ways to construct filters that fulfill this condition. We take the filter η and rotate it to form the first row, and expand that structure into an entire block in three different ways

- In the **row-constant channel-wise expansion**, we simply copy the first row to the others.

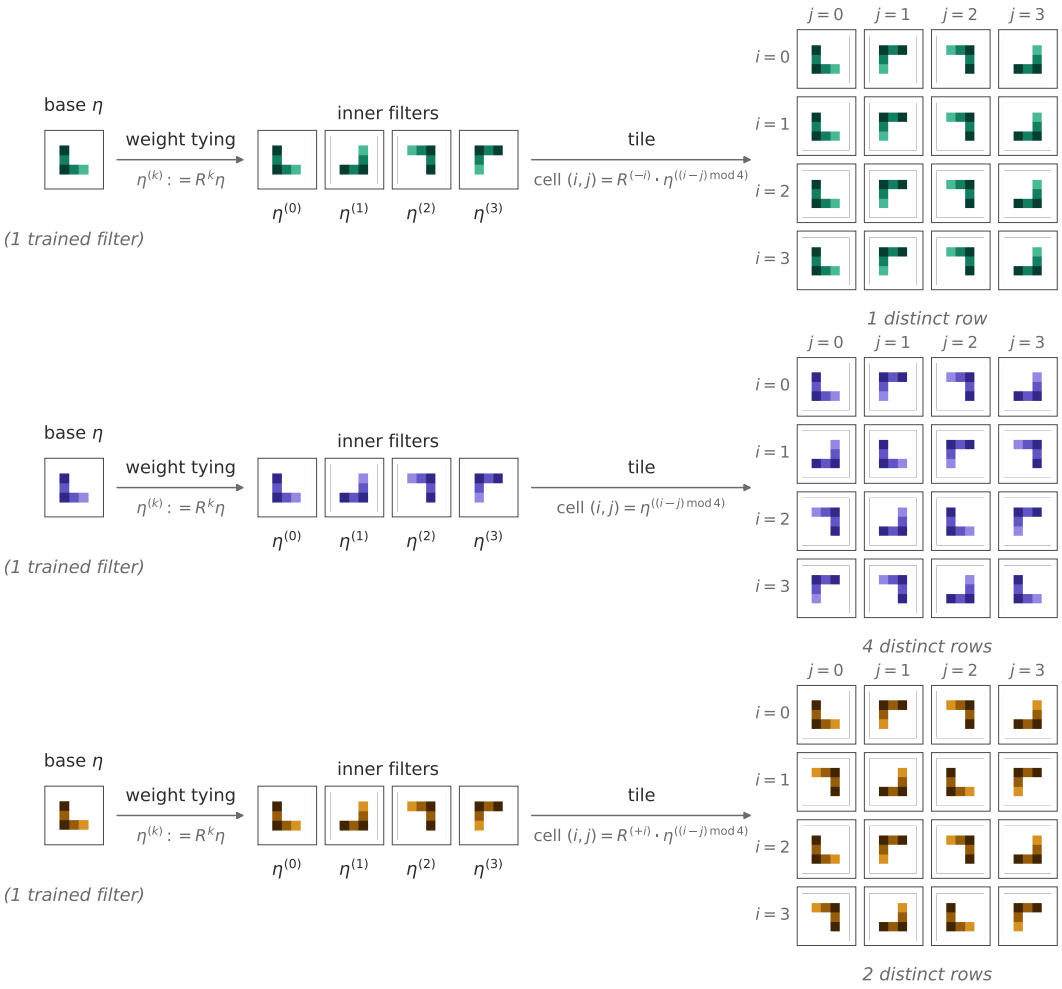


Figure 5: Three filter arrangements for orbit expansion under C_4, starting from a single base filter η trained in the width-$1/|G|$ Stage 1 network. All three share the same first row, and only differ in how that row is expanded to the other rows. **Top:** Row-constant channel-wise expansion **Middle:** Row-rolled channel-wise expansion **Bottom:** GCNN-expansion

- In the **row-rolled channel-wise expansion**, we roll the filters along the row dimension to define the new rows.
- In the **GCNN expansion**, we roll the filters along the row dimension and rotate them to define the new rows.

All these constructions yield per se valid structures of equivariance, since there is a freedom to choose the representation on the second feature space. At the same time, only the third construction is equivariant in the 'ρ^{gcnn}-to-ρ^{gcnn} sense', so it is special. See Figure 6 for a summary.

Remark 1. *One could also let ρ^{ch} act on the space $\mathcal{V}_1$, but this would necessitate using rotation-symmetric filters in the first layer, hurting expressiveness severely. We therefore do not consider that construction.*

Remark 2. *Note that the middle construction is interesting in that it has an additional partial equivariance – it is equivariant when C_4 acts* only *through rolling on both input and output space,*

Table 2: Effect of filter arrangement and trigger timing in orbit expansion (FashionMNIST, $N_0 = 5\,000$, C_4 augmentation, SGD, 5 seeds), evaluated at the best-accuracy checkpoint. **Bold**: best in each column. Shaded : best equivariance-accuracy trade-off.

Variant	Epoch	Acc. (%) ↑	OSP ↑	Sym. KL ↓
Row-constant channel-wise expansion	0	$80.2_{\pm 0.3}$	$0.957_{\pm .004}$	$0.0078_{\pm .0011}$
	20	$80.3_{\pm 0.5}$	$0.956_{\pm .004}$	$0.0084_{\pm .0014}$
	100	$80.1_{\pm 0.5}$	$0.955_{\pm .003}$	$0.0086_{\pm .0006}$
Row-rolled channel-wise expansion	0	$80.2_{\pm 0.3}$	$0.956_{\pm .004}$	$0.0084_{\pm .0010}$
	20	$80.3_{\pm 0.5}$	$0.956_{\pm .004}$	$0.0088_{\pm .0013}$
	100	$80.1_{\pm 0.5}$	$0.954_{\pm .003}$	$0.0086_{\pm .0006}$
GCNN-expansion	0	$\mathbf{80.4}_{\pm 0.4}$	$0.963_{\pm .003}$	$0.0057_{\pm .0004}$
	20	$80.4_{\pm 0.6}$	$0.966_{\pm .004}$	$0.0048_{\pm .0009}$
	100	$80.0_{\pm 0.5}$	$\mathbf{0.968}_{\pm .003}$	$\mathbf{0.0042}_{\pm .0007}$

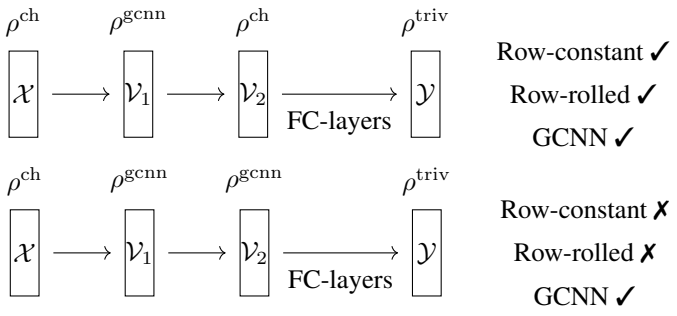


Figure 6: Two ways of choosing the intermediate representations acting on the filter banks. All of the orbit expansion strategies are equivariant with respect to the top choice, but only the GCNN-strategy is equivariant with respect to the bottom choice.

i.e. $(\rho^{\mathrm{roll}}(i)v)_k = v_{k-i}$. *The corresponding calculation for the lifted representation becomes*

$$(\rho_1(i)v)_k = \rho_{\mathcal{V}_2}^{\mathrm{roll}}(i)([\theta_\ell \circ \rho_{\mathcal{V}_\ell}^{\mathrm{roll}}(i)^{-1}]v)_k = \sum_j \eta^{(k+i),j} * (\rho_{\mathcal{V}_\ell}^{\mathrm{gcnn}}(i)^{-1}v)_j$$
$$= \sum_j \eta^{(k+i),j} * (v_{j-i}) = \sum_j \eta^{k+i,j+i} * v_j \,,$$

and the equivariance condition becomes $\eta^{k+i,j+i} = \eta^{k,j}$ *– i.e. that the filters along each diagonal are equal.*

Experiment. We evaluate all three variants with trigger epochs at 0, 20, and 100, using the same setup as Section 4.2 (FashionMNIST, $N_0 = 5\,000$, C_4 augmentation, SGD, 5 seeds). Table 2 reports metrics at the best-accuracy checkpoint (the convention used throughout the paper).

The GCNN-expansion arrangement systematically outperformed the row-constant channel-wise expansion and row-rolled channel-wise expansion. This is not surprising given that they are the only ones that are equivariant when the ρ^{gcnn}-representation is acting on all layers, which is the most expressive action in this setting. As we have seen, there are more filters that fulfill the weaker symmetry condition imposed by being equivariant from ρ^{gcnn} to ρ^{ch}, intuitively leading to a weaker implicit bias.

One can also additionally note that in the GCNN-mode, the outer rotation R^{-i} and the inner shift index $(i-j)$ are oriented so that i cancels in their composition, collapsing all rows to a single pattern. This leads to a lack of row-wise diversity, which seems to limit the effective degrees of freedom.

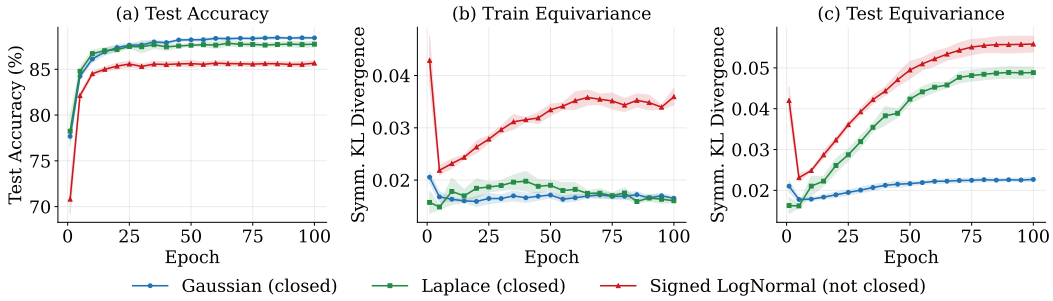


Figure 7: Trajectories for three variational families under C_4 augmentation (5 seeds, shaded region $= \pm$ std). **(a)** both closed families reach $\geq 87.7\%$ test accuracy; the non-closed signed log-normal plateaus at $\approx 85.7\%$. **(b)** Gaussian and Laplace converge to the same low train equivariance defect (≈ 0.016), confirming that closure enables the optimizer to find a near-equivariant posterior on training data, while the signed log-normal remains $> 2\times$ higher. **(c)** the test equivariance defect of all families grows during training as posteriors concentrate, but the ordering Gaussian $<$ Laplace $<$ log-normal is maintained throughout.

J VARIATIONAL FAMILY CLOSURE ABLATION

Gaussian and Laplace families satisfy the conditions in Theorem 1 for linear group representations ρ, whereas a signed log-normal family with random, fixed sign patterns does not. The reason is that the sign assignments break the required affine structure of the sufficient statistics.

We design a controlled experiment to test whether the algebraic distinction has observable consequences during augmented variational training.

Setup. We train Bayesian CNNs on FashionMNIST ($N_0 = 20000$ training samples) with full C_4 orbit augmentation, using SGD for 100 epochs. Three variational families are compared: **(i)** Gaussian (mean-field, closed under C_4), **(ii)** Laplace (mean-field, closed under C_4), and **(iii)** signed log-normal with random fixed signs (not closed under C_4). All other hyper-parameters (architecture, learning rate, KL weights) are shared. Each configuration is run with 5 random seeds.

Metrics. We measure the accuracy and symmetric KL divergence (Symm. KL) between the predictive distributions at x and gx, evaluated on both train and test sets.

Results. Table 3 and Figure 7 summarize the findings.

Table 3: Effect of variational-family closure on equivariance and accuracy under C_4 augmentation (epoch 100, mean $\pm$ std over 5 seeds). Symmetric KL divergence measures the equivariance defect (lower = more equivariant).

		Accuracy (%)		Symm. KL (↓)	
Family	Closed?	Train	Test	Train	Test
Gaussian	✓	90.79 ± 0.17	$\mathbf{88.43 \pm 0.06}$	0.017 ± 0.000	$\mathbf{0.023 \pm 0.001}$
Laplace	✓	96.58 ± 0.12	87.74 ± 0.30	$\mathbf{0.016 \pm 0.000}$	0.049 ± 0.001
Signed LogNorm	✗	90.31 ± 0.13	85.67 ± 0.28	0.036 ± 0.002	0.056 ± 0.002

Both closed families converge to the same low train symmetric KL divergence (≈ 0.016), despite differing in distributional shape and training-set accuracy (Gaussian 90.79%, Laplace 96.58%). The signed log-normal, which is not closed, converges to a $2.2\times$ higher symmetric KL divergence (0.036), with non-overlapping seed ranges. This confirms the theorem's prediction: closure enables the optimizer to move the posterior toward a (near-)equivariant solution within the family, whereas a non-closed family faces a structural barrier.

Table 4: **Trigger-timing ablation, full results.** Symmetrization mechanism × optimizer × trigger epoch on FashionMNIST ($N_0 = 5\,000$, C_4 augmentation, $10\,000$ test samples). For *Geometric avg.* and *Projection*, Epoch is the epoch at which the one-shot symmetrization is applied; for *Orbit expansion* it is the number of stage-1 epochs before expansion (stage-2 runs for the remaining $500 -$ Epoch epochs). All values are mean ± std over 5 seeds, taken at the best test-accuracy checkpoint. **Bold**: best across the whole table per column. Shaded : best equivariance–accuracy trade-off.

Method	Optimizer	Epoch	Acc. (%) ↑	OSP ↑	Sym. KL ↓
Baseline	AdamW	—	79.1±0.6	.923±.004	.0243±.0022
	SGD	—	79.0±0.6	.927±.003	.0217±.0028
Geometric avg.	AdamW	0	77.7±1.0	.944±.002	.0118±.0009
		20	77.9±0.4	.949±.005	.0096±.0011
		50	78.2±0.5	.953±.008	.0090±.0033
		100	77.3±0.9	.955±.008	.0076±.0025
		150	76.8±0.5	.955±.009	.0074±.0026
	SGD	0	78.6±0.4	.950±.003	.0094±.0009
		20	78.6±0.8	.952±.005	.0094±.0020
		50	78.0±0.4	.947±.006	.0116±.0031
		100	78.0±1.1	.946±.005	.0112±.0020
		150	77.9±1.3	.944±.006	.0118±.0025
Projection	AdamW	0	77.9±0.8	.922±.002	.0215±.0017
		20	77.8±0.6	.924±.004	.0215±.0016
		50	76.8±0.2	.926±.006	.0200±.0023
		100	76.2±0.6	.928±.004	.0185±.0020
		150	76.0±0.2	.928±.006	.0170±.0011
	SGD	0	78.4±0.8	.931±.004	.0174±.0026
		20	78.5±0.2	.931±.004	.0180±.0026
		50	77.8±0.9	.927±.004	.0194±.0020
		100	77.2±0.9	.929±.012	.0182±.0046
		150	76.3±1.3	.921±.006	.0208±.0023
Orbit expansion	AdamW	0	77.8±1.0	.932±.004	.0170±.0025
		20	78.0±0.6	.937±.010	.0145±.0041
		50	77.9±0.2	.944±.004	.0114±.0013
		100	77.7±0.4	.950±.004	.0090±.0015
		150	77.5±0.5	.949±.006	.0085±.0015
	SGD	0	**80.4**±0.4	.963±.003	.0057±.0004
		20	80.4±0.6	.966±.004	.0048±.0009
		50	80.2±0.5	.966±.003	.0050±.0008
		100	80.0±0.5	**.968**±.003	**.0042**±.0007
		150	79.9±0.5	.967±.002	.0044±.0003

We further find that equivariance acts as an implicit regularizer. Both closed families generalize better than the non-closed one: the Gaussian generalization gap is 2.4 pp (train 90.79% vs. test 88.43%), roughly half the 4.6 pp gap of the signed log-normal (train 90.31% vs. test 85.67%) at comparable training accuracy. This is consistent with equivariance reducing the effective degrees of freedom in the posterior.

While Laplace's train equivariance matches Gaussian, its test equivariance is noticeably worse (0.049 vs. 0.023). We attribute this to the heavier tails of the Laplace distribution. The posterior satisfies the closure condition on training data, but the tail behaviour introduces additional prediction variance on unseen inputs, partially offsetting the equivariance benefit. This suggests that closure under group action is necessary but not sufficient for strong test-time equivariance; the posterior's concentration properties also play a role.

K TRIGGER TIMING AND OPTIMIZER ABLATION

We empirically study the trigger-timing trade-off discussed in Section 3.4, and the role of the optimizer. The setup is identical to Section 4.2 (FashionMNIST $N_0 = 5000$, C_4 augmentation, 5 seeds; the orbit-expansion variant is the GCNN-expansion of Appendix I), extended to the trigger epoch set $\{0, 20, 50, 100, 150\}$. After the trigger, all methods continue training without symmetry constraints up to epoch 500. We consider both SGD as well as AdamW.

We compare the equivariance properties of the best model (highest test accuracy) of each method in Table 4. There are several aspects of the results to be commented on.

AdamW vs. SGD behavior. First, the results from the main paper for the SGD runs are confirmed: an earlier trigger timing generally leads to a higher accuracy and equivariance (with the exception of the orbit expansion method, which however operates in a region of high equivariance). The AdamW runs behave differently: while their accuracies are still raised by the increased training time post re-initialization, the equivariance is generally hurt. This effect is better illustrated in Figure 8. This confirms the relevance of our theory, since Theorem 2(ii) applies only to SGD. One should also note that although the AdamW-runs seem to drift more from the equivariant positions they are put in by our strategies, they are still made more equivariant than the baseline.

Drift away from H_G. In Figure 10, we plot the evolution of the symmetric KL-divergence for the example of orbit expansion. From this plot it is clear that the runs drift away from equivariance after the re-initialization. This effect persists across all runs and strategies, both for AdamW and SGD. Note that this is not in contradiction to our theory. Indeed, our results are technically only valid for gradient descent applied to the loss associated with the *true predictive* maps. In the practical runs, we are using stochastic gradient descent applied to Monte Carlo approximations of the true predictive map. Importantly, the drift is a lot slower for SGD, further confirming the role of Theorem 2(ii).

Trigger timing. The way the trigger timing influences the drift is subtle. In Figure 9, we plot the accuracy and equivariance metrics for the geometric averaging and projection metrics for two trigger timings. We see that both the immediately projected and the later projected runs drift – but that the earlier triggered runs drift less. We hypothesize that when the projection is applied later in training, it will project from a point further from the typical point a random initialization yields. It is therefore likely that it is projected to a point far away from the optimal region, with large gradients as a result. This should intuitively make the discretization effects more severe, and the drift faster as a result. This is indeed what we observe in the figures: The orbit averaging methods affect the accuracy negatively to a point where it can't recover over the remaining epochs, and also drift away from equivariance quickly. We also speculate that this is one of the reasons of the performance difference of projection and geometric averaging – it is plausible that the more non-local nature of the projection yield even more unlikely points post projection. We leave the task of providing a more complete examination of this phenomenon to future work.

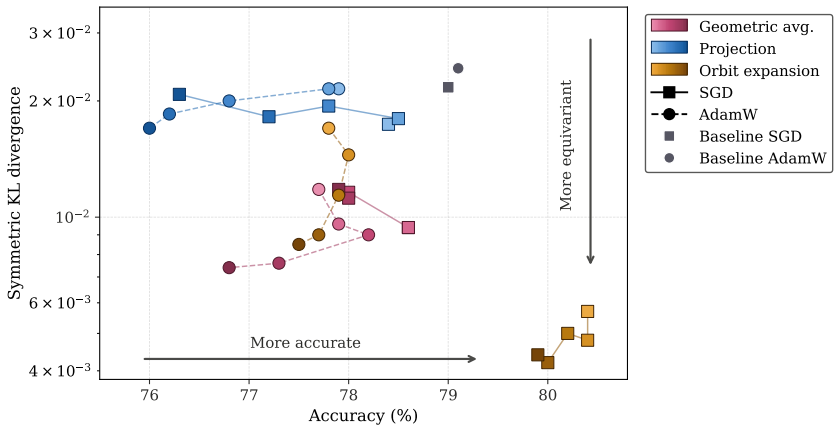


Figure 8: **Accuracy versus equivariance across methods, optimizers, and trigger timings** (FashionMNIST, $N_0 = 5000$, C_4; best-accuracy checkpoint, mean over 5 seeds). The x-axis is test accuracy and the y-axis is symmetric KL divergence on a log scale, so the best region is the *lower right* (more accurate *and* more equivariant). Hue encodes the method: rose for geometric averaging, blue for projection, amber for orbit expansion. Within each method the marker shade runs light→dark for trigger epochs $\{0, 20, 50, 100, 150\}$ (legend). Marker shape and line style encode the optimizer: squares with solid lines for SGD, circles with dashed lines for AdamW. Grey markers are the no-symmetrization baselines (square: SGD, circle: AdamW). Orbit expansion under SGD sits in the lower-right corner, attaining the best accuracy and equivariance simultaneously; SGD matches or exceeds AdamW on accuracy throughout, and projection is the least equivariant of the three mechanisms.

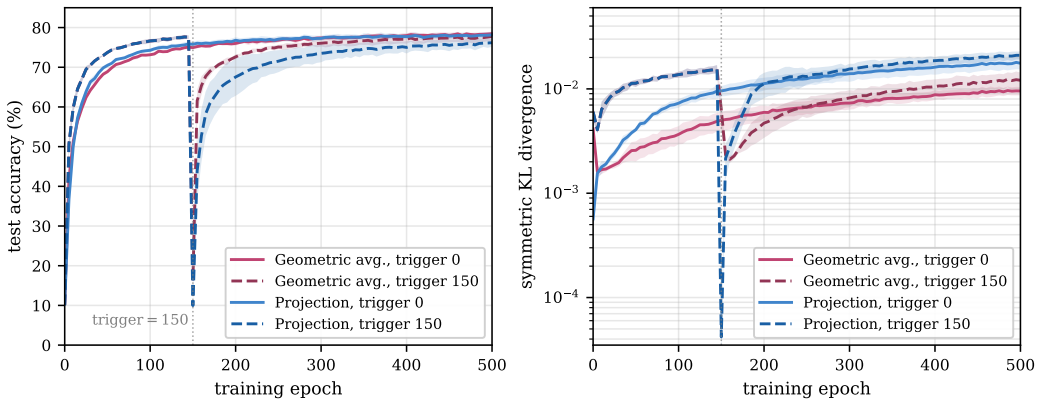


Figure 9: **Training trajectories under projection vs. geometric averaging with early vs. late triggers** (FashionMNIST, $N_0 = 5000$, C_4, SGD; single seed, representative of the 5-seed runs in Table 4). **Left:** test accuracy. At trigger 150 (dotted line) both methods cause a brief collapse and recover, but geometric averaging returns almost back to its trigger-0 level while projection does not. **Right:** symmetric KL divergence on a log scale. The post-trigger drift out of H_G has a similar shape and end value for both trigger settings within each method, consistent with the equivariance metric being controlled mainly by the drift dynamics rather than by trigger timing.

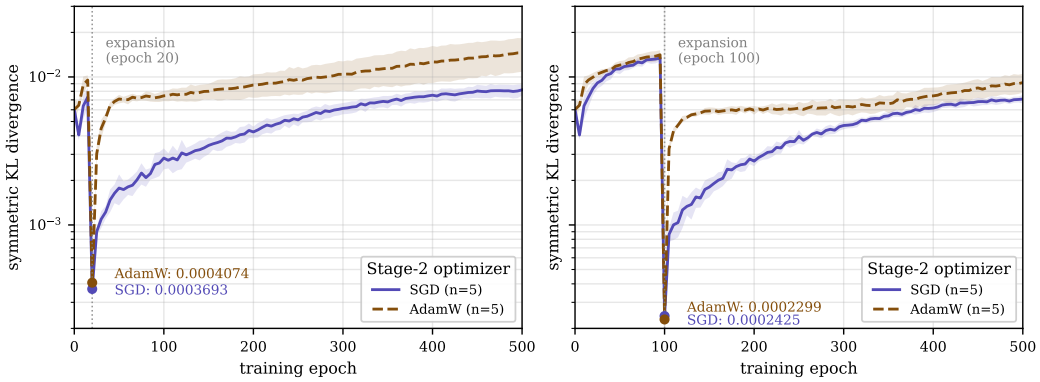


Figure 10: **Drift out of H_G under SGD vs. AdamW for orbit expansion**, at two expansion times (FashionMNIST, $N_0 = 5000$, C_4; mean over 5 seeds, shaded bands ± 1 s.d.). **Left:** Stage-1 $= 20$ epochs; **Right:** Stage-1 $= 100$ epochs. Before the dotted line both optimizers train the same width-$1/|G|$ base network and their curves coincide. At expansion the symmetric KL of both runs drops to the same near-zero floor. During Stage 2 the curves separate: SGD drifts up slowly, AdamW faster and with larger seed variance. Curves trace the full trajectory to epoch 500; Table 4 values are at the best-accuracy checkpoint, which for SGD precedes the late drift and is therefore lower.