Variational Inference Failures Under Model Symmetries: Permutation Invariant Posteriors for Bayesian Neural Networks

Yoav Gelberg [1]   **Tycho F.A. van der Ouderaa** [2][1]   **Mark van der Wilk** [1]   **Yarin Gal** [1]

Abstract

Weight space symmetries in neural network architectures, such as permutation symmetries in MLPs, give rise to Bayesian neural network (BNN) posteriors with many equivalent modes. This multimodality poses a challenge for variational inference (VI) techniques, which typically rely on approximating the posterior with a unimodal distribution. In this work, we investigate the impact of weight space permutation symmetries on VI. We demonstrate, both theoretically and empirically, that these symmetries lead to biases in the approximate posterior, which degrade predictive performance and posterior fit if not explicitly accounted for. To mitigate this behavior, we leverage the symmetric structure of the posterior and devise a *symmetrization* mechanism for constructing permutation invariant variational posteriors. We show that the symmetrized distribution has a strictly better fit to the true posterior, and that it can be trained using the original ELBO objective with a modified KL regularization term. We demonstrate experimentally that our approach mitigates the aforementioned biases and results in improved predictions and a higher ELBO.

1. Introduction

Bayesian neural networks (BNNs) (Tishby et al., 1989; Hinton & van Camp, 1993; Neal, 1995) model neural networks probabilistically by inferring a posterior distribution over their weights. This approach offers robustness to overfitting, epistemic uncertainty estimates, and the ability to learn from limited data, but is typically intractable due to the complexity of the posterior distribution. To use
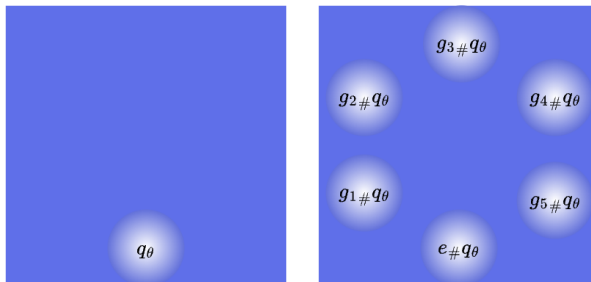

Figure 1. **Symmetrization of the variational posterior.** Given a variational distribution $q_\theta(\boldsymbol{\omega})$, its symmetrization with respect to a group G acting on the underlying space is the average of the pushforwards of $q_\theta(\boldsymbol{\omega})$ over all group elements $g \in G$.

BNNs in practice, posterior estimates are obtained through approximate inference techniques such as variational inference (VI) (Blundell et al., 2015).

While VI facilitates practical posterior estimation, it often relies on unimodal variational posterior families, limiting the accuracy of approximating the typically highly multimodal posterior. Multimodality of the posterior arises from various sources, including the data, task, and model choice. In BNNs specifically, inherent redundancies in the parametric representation of the model, caused by architectural symmetries (Hecht-Nielsen, 1990; Pourzanjani et al., 2017; Rossi et al., 2023; Laurent et al., 2024), are key sources of posterior multimodality. This phenomenon, sometimes referred to as non-identifiability, leads to the emergence of equivalent modes in the posterior that represent networks with identical functionality but different weight configurations.

This work investigates the behavior of unimodal VI approximations in the presence of equivalent modes. We show that unimodal variational distributions tend to interpolate between closely spaced modes, causing a shift in mean and variance. In the case of equivalent modes of the posterior, interpolation results in worse predictive performance and contributes to the underfitting tendencies of variational BNNs (Wenzel et al., 2020; Ghosh et al., 2018; Dusenberry et al., 2020). We demonstrate this effect in a VI setup with a Gaussian mixture target distribution and in a simple BNN with a tractable posterior. Additionally,

[1]University of Oxford, Oxford, UK [2]Imperial College London, London, UK. Correspondence to: Yoav Gelberg <yoav@robots.ox.ac.uk>.

we provide theoretical analysis on the spacing of equivalent modes in BNN posteriors, suggesting the severity of the problem grows with model width.

To address these limitations, we leverage the symmetric structure of the posterior and propose an approach that "bakes in" permutation invariance into the variational distribution. We devise a generic symmetrization mechanism that averages a black-box variational posterior family over the orbits of the permutation symmetry group. This mechanism results in a permutation invariant variational posterior which has a *provably* better fit to the true posterior. We show that training the symmetrized distribution can be done by modifying the prior KL term in the ELBO associated with the original distribution. Evaluating this new KL term requires computing the entropy of a symmetric mixture of densities, for which we derive a computationally efficient estimator. This is significant as it eliminates the need for computationally intractable averaging over *all* weight-space permutations. Instead, symmetry is accounted for by stochastically adapting the KL regularization term. We conduct empirical evaluations comparing our symmetrization method to standard mean-field VI. We demonstrate that our approach achieves better predictive performance and a higher ELBO.

Contributions. This paper makes the following contributions: (1) It demonstrates that in the presence of weight space symmetries, the use of unimodal approximate posterior distributions may result in interpolation between equivalent modes, which degrades predictive performance and posterior fit. (2) It devises a symmetrization procedure for approximate posterior families, proves that it strictly improves posterior fit, and derives a tractable ELBO objective for it by constructing a novel entropy estimator for symmetric mixture of densities. (3) It demonstrates, experimentally, that VI using symmetrized variational posteriors prevents equivalent mode interpolation and results in better predictions and posterior fit, improving overall performance over mean-field VI.

Related work. Several studies have looked into the effects of non-identifiability on approximate inference in BNNs. Kurle et al. (2022) formally discussed the effects of invariances of the likelihood function on mean-field VI and proposed a mitigation for data-dependent translation symmetries in Bayesian linear regression and BNNs. Pourzanjani et al. (2017) and Kurle et al. (2021) suggested constraining the model architecture to lower the number of weight space symmetries. Other works proposed approximating the posterior predictive directly (Sun et al., 2019; Rudner et al., 2022), thereby performing VI in function space. In this work, we focus on the effects of permutation symmetries, and mitigate the problem without constraining the architecture. Instead, we construct a permutation invariant variational posterior and derive an estimator for its VI objective. To the best of our knowledge, this is the first work to develop *trainable* permutation invariant variational posteriors for BNNs.

2. Preliminaries

Notation. We use bold lower case letters (e.g. $\mathbf{x}$) for vectors, bold upper case letters (e.g. $\mathbf{X}$) for matrices, and standard letters (e.g. x) for scalars. g is reserved for group elements and e denotes the identity. $\boldsymbol{\omega} \in \mathcal{W}$ represents model parameters and $\mathbf{f}^{\boldsymbol{\omega}}$ denotes the function parameterized by $\boldsymbol{\omega}$. $\mathcal{W}$ is referred to as the weight space of $\mathbf{f}^{\boldsymbol{\omega}}$. Throughout the paper we assume all discussed distributions admit a density function, and adopt the convention of identifying probability distributions with their densities (e.g. p denotes both a probability measure $p(\mathbf{x} \in A)$ and its density $p(\mathbf{x})$).

Bayesian deep learning. Let $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ be a dataset, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ represents the inputs and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ represents the corresponding outputs. Bayesian deep learning seeks parameters $\boldsymbol{\omega} \in \mathcal{W}$ of a function $\mathbf{y} = \mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x})$ that are most *likely* to have generated the observed outputs $\mathbf{Y}$. To do so, we place a prior distribution $p(\boldsymbol{\omega})$ over $\mathcal{W}$ which encodes our initial beliefs about the likelihood of the parameters before observing any data. Upon observing the data, the prior is updated to the posterior distribution $p(\boldsymbol{\omega} \mid \mathcal{D})$ using Bayes' theorem

$$p(\boldsymbol{\omega} \mid \mathcal{D}) = \frac{p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})p(\boldsymbol{\omega})}{p(\mathbf{Y} \mid \mathbf{X})} \tag{1}$$

Here, $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})$ denotes the likelihood, which describes the probability of observing the outputs $\mathbf{Y}$ given the inputs $\mathbf{X}$ and parameters $\boldsymbol{\omega}$. Assuming i.i.d data, $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega}) = \prod_{i=1}^{N} p(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\omega}) = \prod_{i=1}^{N} p(\mathbf{y}_i \mid \mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}_i))$. The choice of the likelihood function $p(\mathbf{y} \mid \mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}))$ depends on the task:

1. For classification tasks, a common choice is the softmax likelihood $p(y = k \mid \mathbf{x}, \boldsymbol{\omega}) = \text{softmax}(\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}))_k$.

2. In regression tasks, likelihood is usually assumed to be Gaussian, i.e. $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\omega}) = \mathcal{N}(\mathbf{y}; \mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}), \sigma^{-1}\mathbf{I})$.

A key component in Equation (1) is the normalizing constant, $p(\mathbf{Y} \mid \mathbf{X})$, also known as the *model evidence* or *marginal likelihood*. $p(\mathbf{Y} \mid \mathbf{X})$ integrates over all possible parameter values according to the prior

$$p(\mathbf{Y} \mid \mathbf{X}) = \int p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})p(\boldsymbol{\omega})d\boldsymbol{\omega} \tag{2}$$

While analytical computation of the marginal likelihood (and thus the posterior) is possible for simple models, it becomes intractable for complex models such as deep neural networks.

Variational inference. In cases where exact posterior evaluation is not feasible, variational inference provides an approach to approximate the true posterior $p(\boldsymbol{\omega} \mid \mathcal{D})$ using a tractable variational distribution $q_\theta(\boldsymbol{\omega})$. To find the best approximate posterior distribution, the Kullback–Leibler

divergence (KL) between $q_\theta(\boldsymbol{\omega})$ and the true posterior is minimized

$$
\begin{aligned}
\mathrm{KL}\left(q_\theta(\boldsymbol{\omega}) \,\|\, p(\boldsymbol{\omega} \mid \mathcal{D})\right) &= \int q_\theta(\boldsymbol{\omega}) \log\left(\frac{q_\theta(\boldsymbol{\omega})}{p(\boldsymbol{\omega} \mid \mathcal{D})}\right) d\boldsymbol{\omega} \\
&= \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta}\left(\log\left(\frac{q_\theta(\boldsymbol{\omega})}{p(\boldsymbol{\omega} \mid \mathcal{D})}\right)\right)
\end{aligned}
$$

Since the true posterior is intractable for large models, the above KL divergence is intractable in general. However, it turns out that minimizing KL divergence is equivalent to maximizing the evidence lower bound (ELBO), defined by

$$
\begin{aligned}
\mathcal{L}_{\mathrm{VI}}(\theta) := \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})}\left(\log\left(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})\right)\right) \\
- \mathrm{KL}\left(q_\theta(\boldsymbol{\omega}) \,\|\, p(\boldsymbol{\omega})\right)
\end{aligned}
\tag{3}
$$

The first term of the ELBO is referred to as the *expected log-likelihood* term and the second term is referred to as the *prior KL* term. Maximizing the expected log-likelihood term encourages $q_\theta(\boldsymbol{\omega})$ to fit the data, while minimizing the prior KL term encourages $q_\theta(\boldsymbol{\omega})$ to remain close to the prior. The term evidence lower bound comes from the identity

$$
\begin{aligned}
\mathcal{L}_{\mathrm{VI}}(\theta) &= \log\left(p\left(\mathbf{Y} \mid \mathbf{X}\right)\right) - \mathrm{KL}\left(q_\theta(\boldsymbol{\omega}) \,\|\, p(\boldsymbol{\omega} \mid \mathcal{D})\right) \\
&\leq \log(p(\mathbf{Y} \mid \mathbf{X}))
\end{aligned}
$$

Therefore, since the marginal likelihood is constant w.r.t θ, maximizing the ELBO is equivalent to minimizing KL with the posterior. For more details, see (Blundell et al., 2015).

Group representations. Given a group G and a vector space $\mathcal{V}$, a representation of G is a homomorphism $\rho : G \to GL(\mathcal{V})$. We call ρ an *orthonormal* representation if $\mathcal{V}$ admits an inner product and $\rho(G) \subseteq O(\mathcal{V})$, i.e. if $\rho(g)$ are norm-preserving linear transformations. When ρ is clear from the context, we identify g with its representation $\rho(g)$ and write $g \cdot \mathbf{v} := \rho(g)\mathbf{v}$, $\det(g) := \det(\rho(g))$.

Group invariant functions and measures. A function $f : \mathcal{V} \to \mathcal{Y}$ is called G-invariant if $\forall g \in G, \mathbf{v} \in \mathcal{V}$, $f(g \cdot \mathbf{v}) = f(\mathbf{v})$. A measure ν on $\mathcal{V}$ is called G-invariant if $\forall g \in G$ and for every measurable set $A \subseteq \mathcal{V}, \nu\left(g^{-1} \cdot A\right) = \nu(A)$. The distribution defined by $g_\# \nu(A) := \nu(g^{-1} \cdot A)$ is called the g-pushforward measure. Under this definition, ν is G-invariant iff $g_\# \nu = \nu$ for all $g \in G$. If ν has a density, f_ν, i.e. $\nu(A) = \int_A f_\nu(\mathbf{v}) \, d\mathbf{v}$, it transforms under pushforward and becomes $f_{g_\# \nu}(\mathbf{v}) = \frac{f_\nu(g^{-1} \cdot \mathbf{v})}{|\det(g)|}$. If the representation is orthonormal (thus $\det(g) = \pm 1$) then $f_{g_\# \nu}(\mathbf{v}) = f_\nu(g^{-1} \cdot \mathbf{v})$.

Multilayer perceptrons. Multilayer Perceptrons (MLPs) are feedforward neural networks with fully connected layers. Formally, an L-layer MLP is a parametric function $\mathbf{f} = \mathbf{f}^{\boldsymbol{\omega}}$ defined recursively by

$$
\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}) = \mathbf{x}_L, \quad \mathbf{x}_{l+1} = \sigma(\mathbf{W}_{l+1}\mathbf{x}_l + \mathbf{b}_{l+1}), \quad \mathbf{x}_0 = \mathbf{x}
$$

where $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and $\mathbf{b}_l \in \mathbb{R}^{d_l}$. $\boldsymbol{\omega} = (\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L)$ represents the MLP's parameters referred to as weights. $d_1, \ldots, d_{L-1}$ are called the *hidden dimensions*, d_0 is the *input dimension*, and d_L is the *output dimension*. Note that the weight space of an MLP is a real vector space and can be identified with $\bigoplus_{l=1}^{L}\left(\mathbb{R}^{d_l \times d_{l-1}} \oplus \mathbb{R}^{d_l}\right)$.

Permutation symmetries of MLPs. MLPs exhibit well-documented permutation invariance properties (Hecht-Nielsen, 1990; Pourzanjani et al., 2017; Simsek et al., 2020; Ainsworth et al., 2023; Navon et al., 2023; Zhou et al., 2024): permuting the neurons of any hidden layer, while keeping track of the connections to the neighboring layers, preserves the function represented by the MLP. More formally, this invariance can be described by a group representation.

Definition 2.1 (MLP permutation symmetry group). Given an MLP with hidden dimensions $d_1, \ldots d_{L-1}$, its permutation symmetry group is

$$
G = S_{d_1} \times \cdots \times S_{d_{L-1}}
$$

Given $\boldsymbol{\omega} = (\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L) \in \mathcal{W}$ and $g = (\tau_1, \ldots, \tau_{L-1}) \in G$, g's action on $\boldsymbol{\omega}$ is defined by $g \cdot \boldsymbol{\omega} = \boldsymbol{\omega}' = (\mathbf{W}_1', \ldots, \mathbf{W}_L', \mathbf{b}_1', \ldots, \mathbf{b}_L')$ with

$$
\begin{aligned}
\mathbf{W}_1' &= \mathbf{P}_{\tau_1}^\top \mathbf{W}_1, & \mathbf{b}_1' &= \mathbf{P}_{\tau_1}^\top \mathbf{b}_1 \\
\mathbf{W}_l' &= \mathbf{P}_{\tau_l}^\top \mathbf{W}_l \mathbf{P}_{\tau_{l-1}}, & \mathbf{b}_l' &= \mathbf{P}_{\tau_l}^\top \mathbf{b}_l, & l \in \{2, \ldots, L-1\} \\
\mathbf{W}_L' &= \mathbf{W}_L \mathbf{P}_{\tau_{L-1}}, & \mathbf{b}_L' &= \mathbf{b}_L
\end{aligned}
$$

where $\mathbf{P}_{\tau_l} \in \mathbb{R}^{d_l \times d_l}$ is the permutation matrix associated with $\tau_l \in S_{d_l}$. It's straightforward to verify that the mapping $\boldsymbol{\omega} \mapsto g \cdot \boldsymbol{\omega}$ is linear and norm-preserving and that the action of G defines an orthonormal representation on $\mathcal{W}$. Additionally, for any point-wise non-linearity, the transformed weights under g represent the same function as the original weights, i.e. $\mathbf{f}^{g \cdot \boldsymbol{\omega}} \equiv \mathbf{f}^{\boldsymbol{\omega}}$. In other words, the mapping $\boldsymbol{\omega} \mapsto \mathbf{f}^{\boldsymbol{\omega}}$ is G-invariant.

Permutation invariance of the posterior. Throughout the paper, we assume a G-invariant prior on $\mathcal{W}$. This assumption is reasonable as apriori all weight configurations representing the same function are equally likely. Additionally, the most commonly used priors over BNNs (such as the isotropic Gaussian prior, the de-facto standard (Fortuin et al., 2022)), are invariant to weight space permutations. Since the likelihood $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\omega}) = p(\mathbf{y} \mid \mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}))$ is a function of $\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x})$, it too is G-invariant ($p(\mathbf{Y}|\mathbf{X}, g^{-1} \cdot \boldsymbol{\omega}) = p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\omega})$). Therefore, from Equation (1) we get:

Proposition 2.2. $p(\boldsymbol{\omega} \mid \mathcal{D})$ *is G-invariant.*

This invariance property causes mode multiplicity in the posterior. If $\boldsymbol{\omega}^*$ is a mode of $p(\boldsymbol{\omega} \mid \mathcal{D})$, so is $g \cdot \boldsymbol{\omega}^*$ for all $g \in G$. This means that, for every mode of the posterior, there are up to $|G| = d_1! \cdots d_{L-1}!$ *equivalent* modes. The significant

number of equivalent modes poses a challenge for standard VI techniques, as we will explore in the next section.

It should be noted that MLPs and other deep learning architectures exhibit other weight space symmetries that are not considered in this work, such as scaling transformations (Badrinarayanan et al., 2015; Phuong & Lampert, 2020; Entezari et al., 2021) and data dependent symmetries.

3. Effects of Equivalent Modes in the Posterior on Unimodal VI Approximations

Mode-seeking behavior of the KL divergence. Variational Inference often employs unimodal approximate posteriors, such as the common mean-field approximation with a Gaussian variational posterior. A commonly cited justification for this approach is the mode-seeking behavior of the KL divergence (Murphy, 2022). When approximating a fixed distribution $p(\boldsymbol{\omega})$ with a simpler variational distribution $q_\theta(\boldsymbol{\omega})$ via KL minimization, we expect to see a different behavior depending on the direction of the KL loss.

1. **Forward KL.** Minimizing $\mathrm{KL}(p(\boldsymbol{\omega}) \,||\, q_\theta(\boldsymbol{\omega}))$ forces $q_\theta(\boldsymbol{\omega})$ to include all areas for which $p(\boldsymbol{\omega})$ assigns non-zero mass. This behavior is called *mode-covering*.

2. **Reverse KL.** Minimizing $\mathrm{KL}(q_\theta(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega}))$ forces $q_\theta(\boldsymbol{\omega})$ to exclude all the areas for which $p(\boldsymbol{\omega})$ is zero. This results in $q_\theta(\boldsymbol{\omega})$ assigning mass to few parts of the space, near $p(\boldsymbol{\omega})$'s modes. This behavior is called *mode-seeking*.

As seen in Equation (3), variational inference over BNNs employs reverse KL minimization. Therefore, we might intuitively expect a unimodal variational posterior to converge to one of the equivalent modes in the true posterior. Our experiments reveal a potential pitfall of this intuition. We demonstrate that for certain configurations of modes in the target distribution, a unimodal approximation can become "stuck" between two modes, even when using reverse KL minimization.

We approximate the Gaussian mixture $p_\alpha(\mathbf{x}) = \frac{1}{2}\left(\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}) + \mathcal{N}(\mathbf{x}; \alpha\mathbf{u}, \mathbf{I})\right)$, where $\mathbf{u} \in \mathbb{R}^d$ is a unit vector and α parameterizes the distance between $p_\alpha(\mathbf{x})$'s mixture components. We use a unimodal Gaussian variational approximation $q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with variational mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. We initialize optimization with $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = (\mathbf{0}, \mathbf{I})$, corresponding to $p_\alpha(\mathbf{x})$'s mixture component, in order to further encourage mode-seeking behavior. See additional details in Appendix A. The mode of the true optimum lies on the line segment $[\mathbf{0}, \alpha\mathbf{u}]$. Therefore, denoting the result of optimization by $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, we interpret $\alpha^{-1}\|\boldsymbol{\mu}^*\|_2$ as a measure of the solution's *interpolation* between $p_\alpha(\mathbf{x})$'s components. $\alpha^{-1}\|\boldsymbol{\mu}^*\|_2 = \frac{1}{2}$ corresponds to perfect mid-point interpolation and $\alpha^{-1}\|\boldsymbol{\mu}^*\|_2 = 0$ means no interpolation.
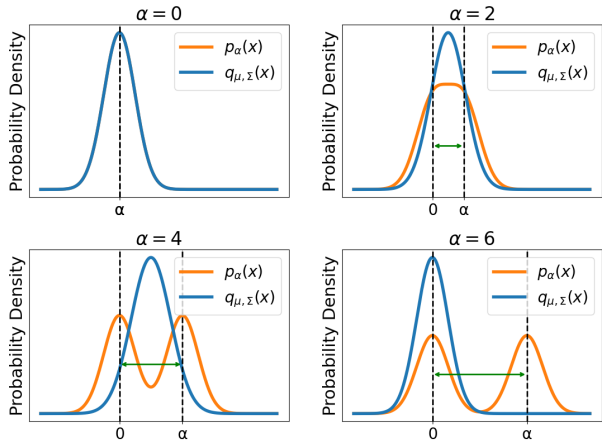

Figure 2. **Unimodal reverse KL minimizer of a bimodal target distribution.** Plots of the target distribution $q_\alpha(x)$ and the unimodal reverse KL minimizer $q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(x)$ in one dimension.

Our results, summarized in Figure 3, suggest the existence of a threshold α^* (in this experiment $\alpha^* \approx 5$), controlling mode-seeking behavior. For mode discrepancy $\alpha \leq \alpha^*$, the optimization results are in fact *mean-seeking* in the sense that $\boldsymbol{\mu}^*$ is a mid-point interpolation between between 0 and $\alpha\mathbf{u}$. Only for $\alpha \geq \alpha^*$, the behavior becomes mode-seeking. We empirically find that the threshold for mode-seeking behavior scales with the standard deviation of $p_\alpha(\mathbf{x})$'s mixture components. For details, see Appendix A.

Interpolating equivalent modes of the posterior. What are the effects of mode interpolation in the context of VI for BNNs? While interpolating posterior modes that arise from *different* functions fitting the data might be desirable in some scenarios, in the case of *equivalent* modes (i.e. ones that represent the same function) interpolation becomes detrimental. Interpolating equivalent modes causes the approximate posterior to assign a higher likelihood to weights that lie between multiple equivalent modes. These weights may correspond to less likely functions compared to those that are closer to a single mode. This means that sampling from the variational posterior results in less likely functions, underfitting the data and degrading predictive performance. This intuition is verified empirically in Section 5. Additionally, as demonstrated in Figure 3, interpolating solutions have higher variance than each of the mixture components, suggesting uncertainty may be overestimated.

Equivalent mode proximity. The above discussion suggests that unimodal approximate distributions would tend to interpolate closely spaced equivalent modes in the posterior, which can lead to underfitting. This raises the question: *are equivalent modes in the posterior close to one another?* The following theorem answers this question affirmatively for wide, single hidden layer MLPs.
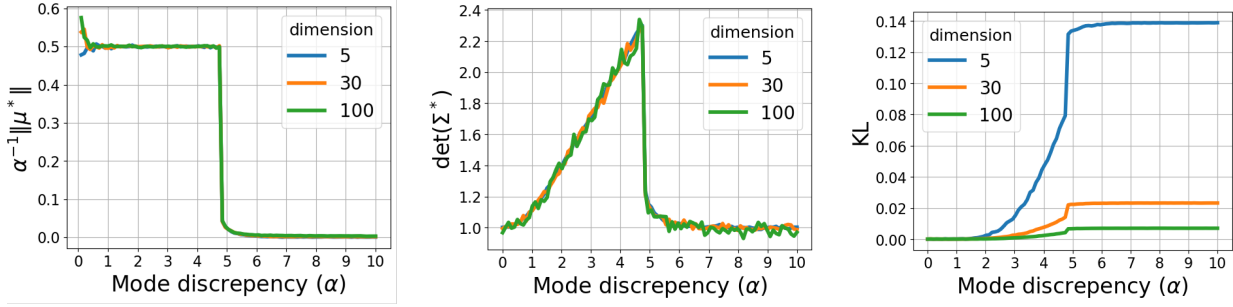
Figure 3. **behavior of the unimodal reverse KL minimizer.** We optimize $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ to minimize $\mathrm{KL}(q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{x}) \; \| \; p_\alpha(\mathbf{x}))$, where $q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a Gaussian distribution and $p_\alpha(\mathbf{x})$ is a mixture of two standard Gaussians with one component centered at $\mathbf{0}$ and another centered α away from $\mathbf{0}$. We plot the results of the optimization as a function of α. From left to right, we plot: $\alpha^{-1} \|\boldsymbol{\mu}^*\|_2$ (quantifying the amount of interpolation between $p_\alpha(\mathbf{x})$'s components), $\det(\boldsymbol{\Sigma}^*)$, and the optimal $\hat{\mathrm{KL}}$ reached by optimization.

Theorem 3.1. *Given an MLP with a single hidden layer of width d_h, input dimension d_i and output dimension d_o, for every weight configuration $\boldsymbol{\omega} \in \mathcal{W}$ there exists a non-trivial weight space permutation $g \in G$ such that*

$$\|\boldsymbol{\omega} - g \cdot \boldsymbol{\omega}\|_2 = \mathcal{O}\left(\frac{d_i + d_o}{\log(d_h)} \|\boldsymbol{\omega}\|_2\right)$$

See proof in Appendix B. As a result, in wide neural networks (e.g. fixing d_i, d_o and taking $d_h \to \infty$), the distance between a mode $\boldsymbol{\omega}^*$ and the closest equivalent mode, relative to $\|\boldsymbol{\omega}^*\|_2$, goes to 0. This suggests higher relative mode proximity in wide networks.

4. Symmetrization of the Variational Posterior

To address the challenge of approximating a permutation invariant posterior with a large number of equivalent modes, we propose using a G-invariant variational posterior. There are several approaches for designing invariant models, including the use of *invariant architectures* (Cohen & Welling, 2016; Zaheer et al., 2017; Hartford et al., 2018; Maron et al., 2018; Navon et al., 2023) and transforming the input to a *canonical form* (Kaba et al., 2022; Tai et al., 2019; Kofinas et al., 2021). In this work, we opt for a simple, architecture-agnostic strategy: averaging a non-invariant model over the orbits of the symmetry group (Puny et al., 2021; Yarotsky, 2018; Benton et al., 2020). We call this process *G-symmetrization*. Concretely, given an approximate posterior family $q_\theta(\boldsymbol{\omega})$, its G-symmetrization is

$$q_\theta^G(\boldsymbol{\omega}) = \frac{1}{|G|} \sum_{g \in G} q_\theta(g^{-1} \cdot \boldsymbol{\omega}) \quad (4)$$

See illustration in Figure 1. $q_\theta^G(\boldsymbol{\omega})$ can be viewed as a coupling of two random variables, $g \in G$ and $\boldsymbol{\omega} \in \mathcal{W}$, with

$$q_\theta^G(g, \boldsymbol{\omega}) = q_\theta^G(\boldsymbol{\omega} \mid g) \cdot q_\theta^G(g) = q_\theta(g^{-1} \cdot \boldsymbol{\omega}) \cdot \frac{1}{|G|}$$

and $q_\theta^G(\boldsymbol{\omega}) = \mathbb{E}_{g \sim q_\theta^G(g)}(q_\theta^G(\boldsymbol{\omega} \mid g))$. Additionally, since the representation of G is orthonormal, $g_\# q_\theta(\boldsymbol{\omega}) = q_\theta(g^{-1} \cdot \boldsymbol{\omega})$ (see Section 2). Thus, Equation (4) can be written as $q_\theta^G(\boldsymbol{\omega}) = \mathbb{E}_{g \sim \mathrm{uniform}(G)}(g_\# q_\theta(\boldsymbol{\omega}))$. For a Gaussian variational posterior $q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ the symmetrization takes the form of a symmetric Gaussian mixture $q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}^G(\boldsymbol{\omega}) = \frac{1}{|G|} \sum_{g \in G} \mathcal{N}(\boldsymbol{\omega}; g \cdot \boldsymbol{\mu}, g \cdot \boldsymbol{\Sigma})$, where $g \cdot \boldsymbol{\Sigma} := \rho(g)^\top \boldsymbol{\Sigma} \rho(g)$. In the case of mean-field VI, the covariance matrix is diagonal $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma})$, so this becomes $g \cdot \mathrm{diag}(\boldsymbol{\sigma}) = \mathrm{diag}(g \cdot \boldsymbol{\sigma})$.

Unless $q_\theta(\boldsymbol{\omega})$ is itself G-invariant (in which case $q_\theta^G \equiv q_\theta$), its symmetrization $q_\theta^G(\boldsymbol{\omega})$ has a *strictly better* fit to the posterior (see Corollary 4.2). However, sampling a model from $q_\theta^G(\boldsymbol{\omega})$ (i.e. $\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x})$, $\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})$) is equivalent to sampling a model from $q_\theta(\boldsymbol{\omega})$. This is because drawing a sample from $q_\theta^G(\boldsymbol{\omega})$ is equivalent to sampling $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$ and transforming it by a random $g \sim \mathrm{uniform}(G)$. Since $\mathbf{f}^{\boldsymbol{\omega}} \equiv \mathbf{f}^{g \cdot \boldsymbol{\omega}}$, the resulting function space distributions are the same (for a more elborate discussion, see appendix C). Therefore, the only difference in performing VI using $q_\theta^G(\boldsymbol{\omega})$ lies in training for the optimal variational parameters θ by maximizing the ELBO objective associated with $q_\theta^G(\boldsymbol{\omega})$. We denote $q_\theta^G(\boldsymbol{\omega})$'s ELBO by $\mathcal{L}_{\mathrm{VI}}^G(\theta)$ and refer to it as the *symmetrized ELBO*. Importantly, even though we train via $\mathcal{L}_{\mathrm{VI}}^G(\theta)$, we can still use the original distribution $q_\theta(\boldsymbol{\omega})$ for making predictions.

Structure of the symmetrized ELBO. While the vast number of weight space permutations induced by the symmetry group G might initially suggest that training using the fully symmetrized ELBO $\mathcal{L}_{\mathrm{VI}}^G(\theta)$ is intractable, the inherent symmetry of $q_\theta^G(\boldsymbol{\omega})$ offers a key advantage and simplifies the structure of $\mathcal{L}_{\mathrm{VI}}^G(\theta)$.

Theorem 4.1. *Given a variational posterior $q_\theta(\boldsymbol{\omega})$*

$$\begin{aligned} \mathcal{L}_{\mathit{VI}}^G(\theta) &= \mathcal{L}_{\mathit{VI}}(\theta) + H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(\boldsymbol{\omega})) \\ &= \mathcal{L}_{\mathit{VI}}(\theta) + I(g; \boldsymbol{\omega}) \end{aligned} \quad (5)$$

where $H(q_\theta(\boldsymbol{\omega}))$ is the entropy of $q_\theta(\boldsymbol{\omega})$, $H(q_\theta^G(\boldsymbol{\omega}))$ is the entropy of $q_\theta^G(\boldsymbol{\omega})$, and $I(g; \boldsymbol{\omega})$ is the mutual information of g and $\boldsymbol{\omega}$ under $q_\theta^G(g, \boldsymbol{\omega})$.

See proof in Appendix D. Theorem 4.1 serves two purposes. In the next section, we use Equation (5) to evaluate $\mathcal{L}_{\mathrm{VI}}^G(\theta)$ in practice by estimating the mutual information $I(g; \boldsymbol{\omega})$. In addition, Theorem 4.1 offers theoretical insight, detailed below, on the posterior fit of the symmetrized variational approximation.

Corollary 4.2. *For all variational parameter configurations θ, $\mathcal{L}_{VI}^G(\theta) \geq \mathcal{L}_{VI}(\theta)$, with equality iff $q_\theta(\boldsymbol{\omega})$ is G-invariant.*

Proof of Corollary 4.2. Since mutual information is non-negative,

$$\mathcal{L}_{\mathrm{VI}}^G(\theta) = \mathcal{L}_{\mathrm{VI}}(\theta) + I(g; \boldsymbol{\omega}) \geq \mathcal{L}_{\mathrm{VI}}(\theta) \qquad (6)$$

Equality in (6) means that $I(g; \boldsymbol{\omega}) = 0$, and therefore that g and $\boldsymbol{\omega}$ are independent under $q_\theta^G(g, \boldsymbol{\omega})$. This, in turn, implies that $\forall g \in G$, $q_\theta^G(\boldsymbol{\omega} \mid g) = q_\theta^G(\boldsymbol{\omega})$, and in particular

$$\forall g \in G, \; q_\theta^G(\boldsymbol{\omega} \mid g) = q_\theta^G(\boldsymbol{\omega}) = q_\theta^G(\boldsymbol{\omega} \mid e) \qquad (7)$$

where $e \in G$ is the identity element. By definition, $q_\theta^G(\boldsymbol{\omega} \mid g) = q_\theta(g^{-1} \cdot \boldsymbol{\omega})$, so (7) becomes $\forall g \in G$, $q_\theta(g^{-1} \cdot \boldsymbol{\omega}) = q_\theta(e^{-1} \cdot \boldsymbol{\omega}) = q_\theta(\boldsymbol{\omega})$, which means $q_\theta(\boldsymbol{\omega})$ is G-invariant. In the other direction, if $q_\theta(\boldsymbol{\omega})$ is G-invariant, its symmetrization is trivial $q_\theta^G \equiv q_\theta$, thus $\mathcal{L}_{\mathrm{VI}}^G(\theta) = \mathcal{L}_{\mathrm{VI}}(\theta)$. $\qquad \square$

In other words, unless $q_\theta(\boldsymbol{\omega})$ is G-invariant to begin with, there is a strict gap between $\mathcal{L}_{\mathrm{VI}}^G(\theta)$ and $\mathcal{L}_{\mathrm{VI}}(\theta)$, implying $q_\theta^G(\boldsymbol{\omega})$ has a strictly better fit to the true posterior, as mentioned previously.

Note that given optimums $\theta^* \in \mathrm{argmax}_\theta \mathcal{L}_{\mathrm{VI}}(\theta)$ and $\theta_G^* \in \mathrm{argmax}_\theta \mathcal{L}_{\mathrm{VI}}^G(\theta)$ we might have $\mathcal{L}_{\mathrm{VI}}(\theta^*) \geq \mathcal{L}_{\mathrm{VI}}(\theta_G^*)$, but we'll always have $\mathcal{L}_{\mathrm{VI}}^G(\theta_G^*) \geq \mathcal{L}_{\mathrm{VI}}^G(\theta^*) > \mathcal{L}_{\mathrm{VI}}(\theta^*)$. Thus, symmetrized VI always results in a better posterior, further motivating the optimization of $\mathcal{L}_{\mathrm{VI}}^G(\theta)$. To tractably do so, Theorem 4.1 suggests focusing on the mutual information term $I(g; \boldsymbol{\omega}) = H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(\boldsymbol{\omega}))$.

Estimating the symmetrized ELBO. As is common in VI, we assume both $\mathcal{L}_{\mathrm{VI}}(\theta)$ and $H(q_\theta(\boldsymbol{\omega}))$ are tractable to evaluate (e.g. in the case of Gaussian prior and variational posterior). Therefore, in order to evaluate $\mathcal{L}_{\mathrm{VI}}^G(\theta)$, it is enough to evaluate the entropy of the symmetrized variational posterior $H(q_\theta^G(\boldsymbol{\omega}))$. The entropy of mixture models typically lacks a closed-form expression, which means we need to resort to an estimator. However, standard estimators for mixture model entropy (Huber et al., 2008; Kolchinsky & Tracey, 2017) are not applicable in this case due to the large number of mixture components in $q_\theta^G(\boldsymbol{\omega})$. Fortunately, we can leverage the symmetric structure of

$q_\theta^G(\boldsymbol{\omega})$ to construct an efficient estimator for the entire mutual information term, and $H(q_\theta^G(\boldsymbol{\omega}))$ in particular.

Mutual information estimation is core to many machine learning tasks such as representation learning (van den Oord et al., 2018; Wu et al., 2020), reinforcement learning (Nachum et al., 2018; Shu et al., 2020), and more. In order to estimate $I(g; \boldsymbol{\omega})$, we leverage the InfoNCE bound (van den Oord et al., 2018; Poole et al., 2019), which is extensively used in contrastive learning, and has well-understood bias and variance (Song & Ermon, 2020). Using InfoNCE, and leveraging the symmetric structure of $q_\theta^G(\boldsymbol{\omega})$ and the identity $H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(\boldsymbol{\omega})) = I(g; \boldsymbol{\omega})$, we obtain the following estimators for $H(q_\theta^G(\boldsymbol{\omega}))$.

Theorem 4.3. *Let $q_\theta(\boldsymbol{\omega})$ be a variational posterior. The estimator*

$$H^K(\theta) = \mathbb{E}\left(\log\left(-\frac{1}{K}\left(q_\theta(\boldsymbol{\omega}) + \sum_{j=1}^{K-1} q_\theta(g_j^{-1} \cdot \boldsymbol{\omega}) \right) \right) \right)$$

where the expectation is over $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$ and $g_1, \ldots, g_{K-1} \sim uniform(G)$ i.i.d. satisfies

$$H^K(\theta) \leq H(q_\theta^G(\boldsymbol{\omega})), \quad \lim_{K \to \infty} H^K(\theta) = H(q_\theta^G(\boldsymbol{\omega}))$$

See proof in Appendix E. The symmetric structure of $q_\theta^G(\boldsymbol{\omega})$ allows us to simplify the original InfoNCE estimator, reducing its computational complexity from $\mathcal{O}(K^2)$ to $\mathcal{O}(K)$. Since $H^K(\theta)$ becomes unbiased as $K \to \infty$, we can trade-off the number of samples, K, and the bias in our estimator. We therefore estimate $\mathcal{L}_{\mathrm{VI}}^G(\theta)$ as

$$\mathcal{L}_{\mathrm{VI}}^K(\theta) = \mathcal{L}_{\mathrm{VI}}(\theta) - H(q_\theta(\boldsymbol{\omega})) + H^K(\theta)$$

Combining Theorems 4.1 and 4.3 we get $\mathcal{L}_{\mathrm{VI}}^K(\theta) \leq \mathcal{L}_{\mathrm{VI}}^G(\theta)$ and $\lim_{K \to \infty} \mathcal{L}_{\mathrm{VI}}^K(\theta) = \mathcal{L}_{\mathrm{VI}}^G(\theta)$.

Optimizing the symmetrized ELBO. We use Monte Carlo estimation to approximate both the expected log-likelihood term and the entropy estimator $H^K(\theta)$ within $\mathcal{L}_{\mathrm{VI}}^K(\theta)$. We assume closed-form expressions for $\mathrm{KL}(q_\theta(\boldsymbol{\omega}) \| p(\boldsymbol{\omega}))$ and $H(q_\theta(\boldsymbol{\omega}))$, which is common in many VI setups. Given a mini-batch of M data points $\mathcal{B} = (\mathbf{X}_\mathcal{B}, \mathbf{Y}_\mathcal{B})$ drawn from a dataset of size N, we sample S weight configurations $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_S \sim q_\theta(\boldsymbol{\omega})$ and use them to estimate $\hat{\mathcal{L}}_{\mathrm{VI}}(\theta)$ as

$$\frac{N}{M}\frac{1}{S}\sum_{i=1}^{S} \log(p(\mathbf{Y}_\mathcal{B} \mid \mathbf{X}_\mathcal{B}, \boldsymbol{\omega}_i)) + \mathrm{KL}(q_\theta(\boldsymbol{\omega}) \| p(\boldsymbol{\omega}))$$

To estimate $H^K(\theta)$, we proceed by sampling $K-1$ random weight space permutations $g_1, \ldots, g_{K-1} \sim \mathrm{uniform}(G)$, computing $\hat{H}^K(\theta)$ as

$$-\frac{1}{S}\sum_{i=1}^{S} \log\left(\frac{1}{K}\left(q_\theta(\boldsymbol{\omega}_i) + \sum_{j=1}^{K-1} q_\theta(g_j^{-1} \cdot \boldsymbol{\omega}_i) \right) \right)$$

(a) Ground truth posterior

$\alpha = 0.05$ $\qquad$ $\alpha = 0.1$ $\qquad$ $\alpha = 0.15$ $\qquad$ $\alpha = 0.2$

(b) Mean-field VI with a Gaussian approximate posterior

$\alpha = 0.05$ $\qquad$ $\alpha = 0.1$ $\qquad$ $\alpha = 0.15$ $\qquad$ $\alpha = 0.2$

(c) Symmetric Gaussian mixture approximate posterior

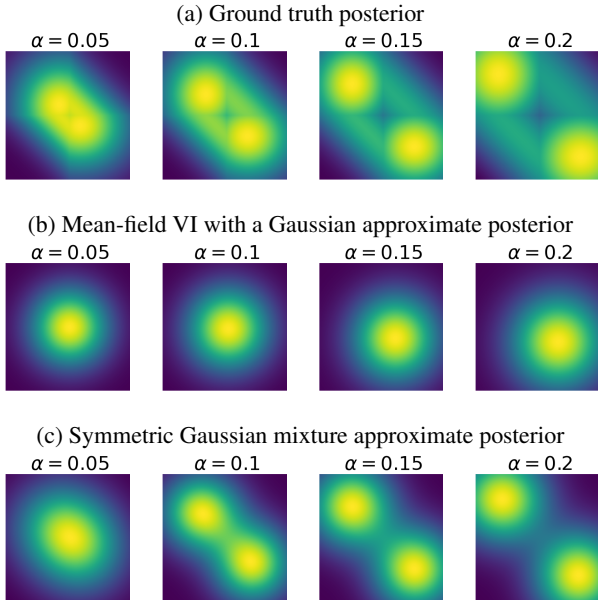$\alpha = 0.05$ $\qquad$ $\alpha = 0.1$ $\qquad$ $\alpha = 0.15$ $\qquad$ $\alpha = 0.2$

Figure 4. **Posterior plots for a tractable BNN.** Ground truth and approximate posterior plots for the model $\mathbf{f}^{w_1,w_2}(x) = \mathrm{ReLU}(w_1 x) + \mathrm{ReLU}(w_2 x)$ given data generated by $f_\alpha(x) = \alpha|x| = \mathrm{ReLU}(\alpha x) + \mathrm{ReLU}(-\alpha x)$.

Combining both estimators, we obtain the final estimator for the symmetrized ELBO objective:

$$\hat{\mathcal{L}}_{\mathrm{VI}}^K(\theta) = \hat{\mathcal{L}}_{\mathrm{VI}}(\theta) - H(q_\theta(\boldsymbol{\omega})) + \hat{H}^K(\theta)$$

We optimize $\hat{\mathcal{L}}_{\mathrm{VI}}^K(\theta)$ via gradient ascent, backpropagating gradients through the Monte Carlo estimators using the reparametrization trick (Kingma & Welling, 2013).

5. Experiments

In this section, we empirically demonstrate both the effects of weight space permutations and the performance of our proposed mitigation. To do so, we conduct two experiments. First, we look at a simple BNN, for which exact posterior computation is tractable. We compare its ground truth posterior to VI approximations given by both mean-field VI and our proposed symmetrization. Secondly, we train MLPs to classify MNIST digits (LeCun & Cortes, 2005) using both inference techniques. We show that our method achieves better predictive performance, and that the gap grows with model width, validating our intuition from Section 3.

In both experiments, we use mean-field VI with a Gaussian approximate posterior (referred to as MFVI) as the base variational posterior. That is, $q_\theta(\boldsymbol{\omega}) = q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}))$, with $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathcal{W}$. We train $q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\boldsymbol{\omega})$ using both $\hat{\mathcal{L}}_{\mathrm{VI}}(\theta)$ and $\hat{\mathcal{L}}_{\mathrm{VI}}^K(\theta)$. The latter corresponds to training a symmetric Gaussian mixture

variational posterior (referred to as SGM) of the form $q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\boldsymbol{\omega}) = \frac{1}{|G|} \sum_{g \in G} \mathcal{N}(\boldsymbol{\omega}; g \cdot \boldsymbol{\mu}, \mathrm{diag}(g \cdot \boldsymbol{\sigma}))$. We place an isotropic Gaussian prior on the weights and optimize using Adam (Kingma & Ba, 2014).

At test time, we take the average prediction of models sampled from the variational posterior, i.e. $\hat{\mathbf{f}}(\mathbf{x}) = \frac{1}{S} \sum_{i=1}^S \mathbf{f}^{\boldsymbol{\omega}_i}(\mathbf{x})$, with $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_S \sim q_\theta(\boldsymbol{\omega})$. We evaluate $\hat{\mathbf{f}}(\mathbf{x})$ on the test set with different random seeds, and report average MSE or accuracy $\pm$ standard deviation. We additionally report ELBO values for the training set, which can be interpreted as measuring the fit of our approximation to the true posterior.

Tractable posterior neural networks. Due to the intractability of ground truth posterior distributions for large models, our first experiment involves a small BNN for which we can analytically compute the posterior. We use a two-layer MLP with scalar input and output, two hidden neurons, ReLU activations, no biases, and a fixed output layer. This design yields a model with two trainable parameters $\boldsymbol{\omega} = (w_1, w_2)$ and a forward pass of $\mathbf{f}^{\boldsymbol{\omega}}(x) = \mathrm{ReLU}(w_1 x) + \mathrm{ReLU}(w_2 x)$. This is the smallest example that still exhibits weight space permutation symmetries, which in this case consist only of the identity permutation and the transposition $(w_1, w_2) \mapsto (w_2, w_1)$. This setup allows us to evaluate the posterior fit for both inference techniques with high accuracy.

We train on synthetic data generated by the function $f_\alpha(x) = \alpha|x| = \mathrm{ReLU}(\alpha x) + \mathrm{ReLU}(-\alpha x)$. The weight configurations $(w_1^*, w_2^*) = (\alpha, -\alpha)$ and $(w_1^*, w_2^*) = (-\alpha, \alpha)$ are equivalent optimal solutions that constitute the two equivalent modes of the true posterior, plotted in Figure 4a. The distance between the modes grows linearly with α. We use training and test sets of size $N = 100$, generated as $((x_1, f_\alpha(x_1)), \dots, (x_N, \mathbf{f}_\alpha(x_N)))$ with $x_1, \dots, x_N \sim \mathrm{uniform}([-10, 10])$ i.i.d. We train our approximate inference methods using mini-batches of 10 data points with learning rate $= 5 \times 10^{-3}$. Results are collected after 10 epochs.

We plot the posterior approximations given by both MFVI and SGM (with $K = 2$) in Figure 4. Figure 4b demonstrates our intuition from Section 3 and shows that the unimodal approximate posterior interpolates between the modes. This translates to poor predictive performance, as demonstrated in Table 2. On the other hand, Figure 4c demonstrates the more accurate posterior fit of the symmetrization. This leads to higher (symmetrized) ELBO values, detailed in Table 3, and better predictive performance, detailed in Table 2.

MNIST classifiers. For the second experiment, we train BNNs to classify MNIST digits using both inference methods. We use single hidden layer MLPs with increasing hidden dimension as the underlying models. We train for 10

Method	Hidden dimension			
	5	10	20	30
Mean-field VI with a Gaussian approximate posterior	$88.379 \pm 0.023\%$	$93.217 \pm 0.022\%$	$94.836 \pm 0.025\%$	$95.432 \pm 0.023\%$
Symmetric Gaussian mixture VI ($K = 5$)	$\mathbf{88.423} \pm 0.027\%$	$\mathbf{93.221} \pm 0.021\%$	$94.899 \pm 0.035\%$	$95.534 \pm 0.033\%$
Symmetric Gaussian mixture VI ($K = 10$)	$88.405 \pm 0.034\%$	$93.220 \pm 0.023\%$	$94.896 \pm 0.026\%$	$95.540 \pm 0.025\%$
Symmetric Gaussian mixture VI ($K = 20$)	$88.408 \pm 0.026\%$	$\mathbf{93.221} \pm 0.027\%$	$\mathbf{94.905} \pm 0.024\%$	$\mathbf{95.552} \pm 0.023\%$

Table 1. **Test accuracy on MNIST.** Predictions taken from an average of 1000 models sampled from the approximate posterior. Results are shown as mean $\pm$ standard deviation over 10 runs with different random seeds. Higher hidden dimension leads to an increase in the number of equivalent modes in the true posterior and thus a larger performance increase over mean-field VI.

Method	α			
	0.05	0.1	0.15	0.2
MFVI	0.11 ± 0.22	0.12 ± 0.16	0.13 ± 0.15	0.18 ± 0.19
SGM ($K = 2$)	$\mathbf{0.09} \pm 0.19$	$\mathbf{0.09} \pm 0.17$	$\mathbf{0.10} \pm 0.15$	$\mathbf{0.12} \pm 0.16$

Table 2. **Test MSE on synthetic dataset.** Predictions are taken from an average of 1000 models sampled from the approximate posterior. Results are shown as mean $\pm$ standard deviation over 10 runs with different random seeds.

Method	ELBO	α			
		0.05	0.1	0.15	0.2
MFVI	$\mathcal{L}_{\text{VI}}$	-0.022	-0.025	-0.032	-0.042
	$\mathcal{L}_{\text{VI}}^G$	-0.009	-0.013	-0.021	-0.031
SGM ($K = 2$)	$\mathcal{L}_{\text{VI}}$	-0.023	-0.029	-0.035	-0.043
	$\mathcal{L}_{\text{VI}}^G$	-0.004	-0.007	-0.012	-0.019

Table 3. **Normalized ELBO values on synthetic dataset.** $\mathcal{L}_{\text{VI}}^G$ is approximated as $\hat{\mathcal{L}}_{\text{VI}}^K$ with $K = 500$ and estimated using 1000 weight samples from $q_\theta(\boldsymbol{\omega})$. Higher values are better as they imply better posterior fit. Notice that for both inference methods, $\mathcal{L}_{\text{VI}} \leq \mathcal{L}_{\text{VI}}^G$. The symmetrized variational posterior has higher $\mathcal{L}_{\text{VI}}^G$ values and marginally lower $\mathcal{L}_{\text{VI}}$ values.

epochs with a batch size of 100 and learning rate $= 10^{-3}$. For the symmetric Gaussian mixture posterior, results are reported for $K = 5, 10, 20$. Accuracy results are displayed in Table 1. The improvement in accuracy achieved by the symmetrization increases with model width, verifying our intuition from Theorem 3.1.

We empirically noticed that the variance of $\hat{H}^K(\theta)$ grows with model size. This can be mitigated by increasing the value of K, which reduces the variance of $\hat{H}^K(\theta)$ and leads to more stable training. Ideally, in order to unlock the full potential of this approach for larger models, we want to be able to reduce the variance of $\hat{H}^K(\theta)$ without using large values of K, which increase the computational cost. This could be achieved in the future using variance reduction techniques, such as importance sampling or stratified sampling.

6. Conclusion

In this work, we demonstrate the effects of weight space permutation symmetries in MLPs on unimodal VI posterior approximations. We propose a symmetrization mechanism to mitigate these effects, prove that it improves the posterior fit of the variational distribution, and demonstrate its performance improvements over mean-field VI. This approach offers an architecture-agnostic way to construct trainable permutation invariant variational posteriors, and paves the way for incorporating other weight-space symmetries in VI.

Limitations. One limitation of our approach is that the variance of $\hat{H}^K(\theta)$ grows model size, which currently limits its effectiveness for larger architectures. This could be mitigated in the future using variance reduction techniques such as importance sampling. One possible mitigation, in the case of symmetric Gaussian mixtures, is importance sampling $g_1, \ldots, g_{K-1}$ with a proposal distribution that places lower mass on permutations associated with far modes, e.g. $f(g_j) \propto \exp\left(-\|g_j \cdot \boldsymbol{\mu} - \boldsymbol{\mu}\|_2\right)$. This gives more weight to permutations that lead to modes that are closely spaced, which are the ones influencing the entropy of $q_\theta^G(\boldsymbol{\omega})$ the most. Implementing this, or other variance reduction strategies, is left for future work.

Future directions. Our approach achieves invariance by averaging the density over group orbits. Exploring alternative avenues for building invariant variational posteriors is a worthwhile research direction. Promising approaches include: (1) *Invariant architectures.* Several invariant architectures for neural network weight spaces have recently been proposed (Navon et al., 2023; Zhou et al., 2024; Lim et al., 2024; Kofinas et al., 2024). Designing expressive weight space permutation invariant posteriors using these architectures is an interesting future direction. (2) *Input canonicalization.* Using weight space alignment techniques such as (Ainsworth et al., 2023; Guerrero Peña et al., 2023; Navon et al., 2024) to transform weights into a canonical form, we can potentially perform inference on the quotient space $\mathcal{W}/G$. Finally, extending our framework to other weight space symmetries is another interesting direction. Most of the analysis in this paper can be extended to general finite/compact symmetry groups of weight spaces.

7. Acknowledgments

References

Ainsworth, S., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T. 3, 8

Badrinarayanan, V., Mishra, B., and Cipolla, R. Understanding symmetries in deep networks. *ArXiv*, abs/1511.01029, 2015. URL https://api.semanticscholar.org/CorpusID:18720237. 4

Benton, G. W., Finzi, M., Izmailov, P., and Wilson, A. G. Learning invariances in neural networks. *ArXiv*, abs/2010.11882, 2020. URL https://api.semanticscholar.org/CorpusID:225039835. 5

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015. 1, 3

Chen, J., Gan, Z., Li, X., Guo, Q., Chen, L., Gao, S., Chung, T., Xu, Y., Zeng, B., Lu, W., Li, F., Carin, L., and Tao, C. Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce. *ArXiv*, abs/2107.01152, 2021. URL https://api.semanticscholar.org/CorpusID:235727265. 15

Cohen, T. and Welling, M. Group equivariant convolutional networks. *ArXiv*, abs/1602.07576, 2016. URL https://api.semanticscholar.org/CorpusID:609898. 5

Dusenberry, M. W., Jerfel, G., Wen, Y., Ma, Y.-A., Snoek, J., Heller, K. A., Lakshminarayanan, B., and Tran, D. Efficient and scalable bayesian neural nets with rank-1 factors. *ArXiv*, abs/2005.07186, 2020. URL https://api.semanticscholar.org/CorpusID:218628552. 1

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. *ArXiv*, abs/2110.06296, 2021. URL https://api.semanticscholar.org/CorpusID:238743980. 4

Fortuin, V., Garriga-Alonso, A., Ober, S. W., Wenzel, F., Ratsch, G., Turner, R. E., van der Wilk, M., and Aitchison, L. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xkjqJYqRJy. 3

Ghosh, S., Yao, J., and Doshi-Velez, F. Structured variational learning of Bayesian neural networks with horseshoe priors. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1744–1753. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ghosh18a.html. 1

Guerrero Peña, F. A., Medeiros, H., Dubail, T., Aminbeidokhti, M., Granger, E., and Pedersoli, M. Re-basin via implicit sinkhorn differentiation. pp. 20237–20246, 06 2023. doi: 10.1109/CVPR52729.2023.01938. 8

Hartford, J. S., Graham, D. R., Leyton-Brown, K., and Ravanbakhsh, S. Deep models of interactions across sets. In *International Conference on Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:3819852. 5

Hecht-Nielsen, R. On the algebraic structure of feed-forward network weight spaces. In ECKMILLER, R. (ed.), *Advanced Neural Computers*, pp. 129–135. North-Holland, Amsterdam, 1990. ISBN 978-0-444-88400-8. doi: https://doi.org/10.1016/B978-0-444-88400-8.50019-4. URL https://www.sciencedirect.com/science/article/pii/B9780444884008500194. 1, 3

Hinton, G. E. and van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pp. 5–13, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897916115. doi: 10.1145/168304.168306. URL https://doi.org/10.1145/168304.168306. 1

Huber, M. F., Bailey, T., Durrant-Whyte, H., and Hanebeck, U. D. On entropy approximation for gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 181–188, 2008. doi: 10.1109/MFI.2008.4648062. 6

Kaba, S.-O., Mondal, A. K., Zhang, Y., Bengio, Y., and Ravanbakhsh, S. Equivariance with learned canonicalization functions. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*,

2022. URL https://openreview.net/forum?id=pVD1k8ge25a. 5

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106. 7

Kingma, D. P. and Welling, M. Stochastic gradient vb and the variational auto-encoder. 2013. URL https://api.semanticscholar.org/CorpusID:51841161. 7

Kofinas, M., Nagaraja, N. S., and Gavves, E. Roto-translated local coordinate frames for interacting dynamical systems. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=c3RKZas9am. 5

Kofinas, M., Knyazev, B., Zhang, Y., Chen, Y., Burghouts, G. J., Gavves, E., Snoek, C. G. M., and Zhang, D. W. Graph neural networks for learning equivariant representations of neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oO6FsMyDBt. 8

Kolchinsky, A. and Tracey, B. D. Estimating mixture entropy with pairwise distances. *ArXiv*, abs/1706.02419, 2017. URL https://api.semanticscholar.org/CorpusID:21868289. 6

Kurle, R., Januschowski, T., Gasthaus, J., and Wang, Y. B. On symmetries in variational bayesian neural nets. In *NeurIPS 2021 Workshop on Bayesian Deep Learning*, 2021. 2

Kurle, R., Herbrich, R., Januschowski, T., Wang, B., and Gasthaus, J. On the detrimental effect of invariances in the likelihood for variational inference. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ft4xGJ8tIZH. 2

Laurent, O., Aldea, E., and Franchi, G. A symmetry-aware exploration of bayesian neural network posteriors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FOSBQuXgAq. 1

LeCun, Y. and Cortes, C. The mnist database of handwritten digits. 2005. URL https://api.semanticscholar.org/CorpusID:60282629. 7

Lim, D., Maron, H., Law, M. T., Lorraine, J., and Lucas, J. Graph metanetworks for processing diverse neural architectures. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ijK5hyxs0n. 8

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. *ArXiv*, abs/1812.09902, 2018. URL https://api.semanticscholar.org/CorpusID:56895597. 5

Murphy, K. P. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL probml.ai. 4

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Near-optimal representation learning for hierarchical reinforcement learning. *ArXiv*, abs/1810.01257, 2018. URL https://api.semanticscholar.org/CorpusID:52909341. 6

Navon, A., Shamsian, A., Achituve, I., Fetaya, E., Chechik, G., and Maron, H. Equivariant architectures for learning in deep weight spaces. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. 3, 5, 8

Navon, A., Shamsian, A., Fetaya, E., Chechik, G., Dym, N., and Maron, H. Equivariant deep weight space alignment, 2024. URL https://openreview.net/forum?id=iT1ttQXwOg. 8

Neal, R. M. Bayesian learning for neural networks. 1995. URL https://api.semanticscholar.org/CorpusID:60809283. 1

Phuong, M. and Lampert, C. H. Functional vs. parametric equivalence of re{lu} networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Bylx-TNKvH. 4

Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., and Tucker, G. On variational bounds of mutual information. *ArXiv*, abs/1905.06922, 2019. URL https://api.semanticscholar.org/CorpusID:155100374. 6, 15

Pourzanjani, A. A., Jiang, R. M., and Petzold, L. R. Improving the identifiability of neural networks for bayesian inference. 2017. URL https://api.semanticscholar.org/CorpusID:46932278. 1, 2, 3

Puny, O., Atzmon, M., Ben-Hamu, H., Smith, E. J., Misra, I., Grover, A., and Lipman, Y. Frame averaging for invariant and equivariant network design. *ArXiv*, abs/2110.03336, 2021. URL https://api.semanticscholar.org/CorpusID:238419638. 5

Rossi, S., Singh, A., and Hannagan, T. On permutation symmetries in bayesian neural network posteriors: a variational perspective. *ArXiv*, abs/2310.10171, 2023. URL https://api.semanticscholar.org/CorpusID:264146533. 1

Rudner, T. G. J., Chen, Z., Teh, Y. W., and Gal, Y. Tractable Function-Space Variational Inference in Bayesian Neural Networks. In *Advances in Neural Information Processing Systems 35*, 2022. 2

Shu, R., Nguyen, T. D., Chow, Y., Pham, T., Than, K., Ghavamzadeh, M., Ermon, S., and Bui, H. H. Predictive coding for locally-linear control. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:211678253. 6

Simsek, B., Brea, J., Illing, B., and Gerstner, W. Weight-space symmetry in neural network loss landscapes revisited, 2020. URL https://openreview.net/forum?id=rkxmPgrKwB. 3

Song, J. and Ermon, S. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1x62TNtDS. 6, 15

Sun, S., Zhang, G., Shi, J., and Grosse, R. FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkxacs0qY7. 2

Tai, K. S., Bailis, P. D., and Valiant, G. Equivariant transformer networks. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:59523684. 5

Tishby, Levin, and Solla. Consistent inference of probabilities in layered networks: predictions and generalizations. In *International 1989 Joint Conference on Neural Networks*, pp. 403–409 vol.2, 1989. doi: 10.1109/IJCNN.1989.118274. 1

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925. 6

Wenzel, F., Roth, K., Veeling, B. S., Swiatkowski, J. B., Tran, L. H., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. Appendix for: How good is the bayes posterior in deep neural networks really?

2020. URL https://api.semanticscholar.org/CorpusID:211044145. 1

Wu, M., Zhuang, C., Mossé, M., Yamins, D. L. K., and Goodman, N. D. On mutual information in contrastive learning for visual representations. *ArXiv*, abs/2005.13149, 2020. URL https://api.semanticscholar.org/CorpusID:218900934. 6

Yarotsky, D. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55:407–474, 2018. URL https://api.semanticscholar.org/CorpusID:13745401. 5

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf. 5

Zhou, A., Finn, C., and Harrison, J. Universal neural functionals. *ArXiv*, abs/2402.05232, 2024. URL https://api.semanticscholar.org/CorpusID:267547546. 3, 8

A. Mode Interpolation: Experimental Details and Additional Results

Experimental details. Since $\text{KL}(q_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) \,\|\, p_\alpha(\mathbf{x}))$ does not have a closed-form expression, we minimize the empirical reverse KL which is estimated by sampling S points $\mathbf{x}_1, \ldots, \mathbf{x}_S$ from $q_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x})$ and computing:

$$\hat{\text{KL}} = \frac{1}{S} \sum_{i=1}^{S} \log \left(\frac{q_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}_i)}{p_\alpha(\mathbf{x}_i)} \right) \tag{8}$$

Using the reparametrization trick to compute gradients w.r.t $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we minimize $\hat{\text{KL}}$ via stochastic gradient descent. We use $S = 5 \times 10^3$ samples to approximate $\hat{\text{KL}}$ and perform 3×10^3 training steps with learning rate $= 10^{-2}$, empirically observing convergence of the loss.

Scaling of the threshold for mode-seeking behavior. As mentioned in the main text, the threshold for mean-seeking vs. mode-seeking behavior in the experimental setup described in Section 3 scales with the standard deviation of $p_\alpha(\mathbf{x})$'s mixture components. To see that, we define $p_\alpha^\sigma(\mathbf{x}) = \frac{1}{2} \left(\mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma^2 \mathbf{I}) + \mathcal{N}(\mathbf{x}; \alpha\mathbf{u}, \sigma^2 \mathbf{I}) \right)$, and repeat the experiment presented in Figure 3 with different values of σ. The following are plots of the results for $\sigma = 0.25, 0.5, 1, 2$.
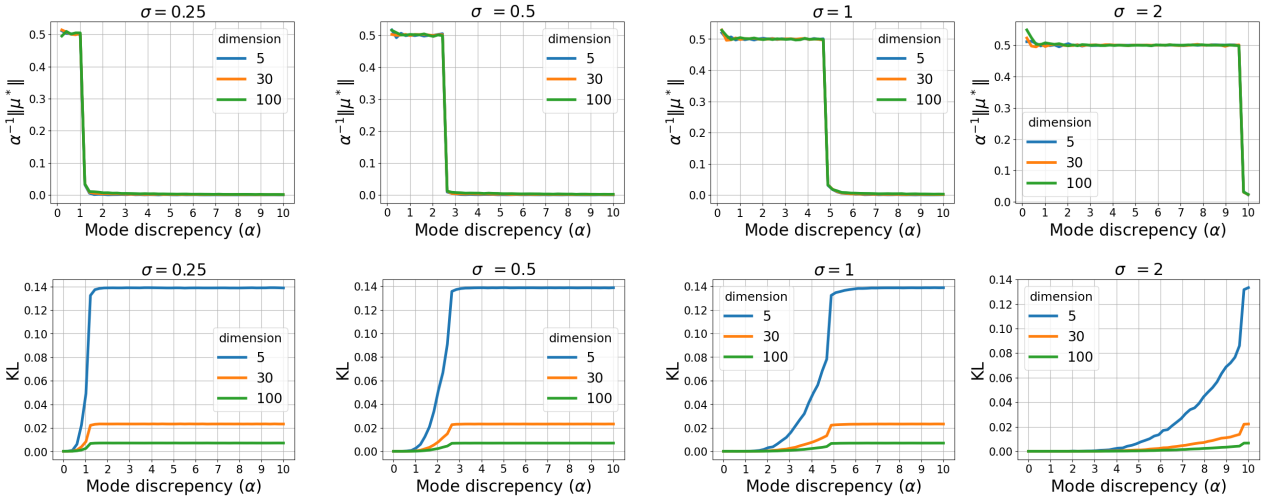

Figure 5. **Threshold for mode-seeking behavior.** We plot $\alpha^{-1}\|\boldsymbol{\mu}^*\|_2$ (interpreted as quantifying the amount of interpolation between $p_\alpha^\sigma(\mathbf{x})$'s modes) and the optimal $\hat{\text{KL}}$ reached by optimization. Plots are provided for $\sigma = 0.25, 0.5, 1, 2$. As can be seen from the plots, the corresponding thresholds $\alpha^* \approx 1.25, 2.5, 5, 10$, scale linearly with σ.

B. Mode Proximity Proof

Proof of Theorem 3.1. Let $\boldsymbol{\omega} \neq \mathbf{0}$ be a weight configuration $\in \mathcal{W}$ (the result holds trivially for $\boldsymbol{\omega} = \mathbf{0}$). Denote $D = \dim(\mathcal{W}) = d_i d_h + d_h + d_h d_o + d_o$ and fix $R = \frac{D}{\log(|G|)}\|\boldsymbol{\omega}\|_2$. Define $B(\boldsymbol{\omega}', r) = \{\boldsymbol{\omega} \in \mathcal{W} \mid \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2 \leq r\}$ to be the ball of radius $r \in \mathbb{R}$ around $\boldsymbol{\omega}' \in \mathcal{W}$. Assume, by way of contradiction, that for all non-trivial $g \in G$, $\|\boldsymbol{\omega} - g \cdot \boldsymbol{\omega}\|_2 > 2R$. Since the action of G is norm-preserving, this implies that $\forall g_1 \neq g_2 \in G$:

$$\|g_1 \cdot \boldsymbol{\omega} - g_2 \cdot \boldsymbol{\omega}\|_2 = \|g_1 \cdot (\boldsymbol{\omega} - g_1^{-1}g_2 \cdot w_2)\|_2 = \|\boldsymbol{\omega} - g_1^{-1}g_2 \cdot \boldsymbol{\omega}\|_2 > 2R$$

Therefore, $\forall g_1 \neq g_2 \in G$, $B(g_1 \cdot \boldsymbol{\omega}, R) \cap B(g_2 \cdot \boldsymbol{\omega}, R) = \emptyset$. Additionally, if $\boldsymbol{\omega}' \in B(g \cdot \boldsymbol{\omega}, R)$ we have

$$\|\boldsymbol{\omega}'\|_2 = \|\boldsymbol{\omega}' - g \cdot \boldsymbol{\omega} + g \cdot \boldsymbol{\omega}\|_2 \leq \|\boldsymbol{\omega}' - g \cdot \boldsymbol{\omega}\|_2 + \|g \cdot \boldsymbol{\omega}\|_2 \leq R + \|\boldsymbol{\omega}\|_2$$

so $B(g \cdot \boldsymbol{\omega}, R) \subseteq B(\mathbf{0}, \|\boldsymbol{\omega}\|_2 + R)$. Combining these observations, we get that $\bigsqcup_{g \in G} B(g \cdot \boldsymbol{\omega}, R) \subseteq B(\mathbf{0}, \|\boldsymbol{\omega}\|_2 + R)$. Therefore, looking at the volume of these subsets

$$\sum_{g \in G} \text{vol}(B(g \cdot \boldsymbol{\omega}, R)) = \text{vol}\left(\bigsqcup_{g \in G} B(g \cdot \boldsymbol{\omega}, R) \right) \leq \text{vol}(B(\mathbf{0}, \|\boldsymbol{\omega}\|_2 + R)) \tag{9}$$

12

The volume of a D-dimensional ball of radius r is a constant times r^D, so Equation (9) becomes

$$|G|R^D \leq (\|\boldsymbol{\omega}\|_2 + R)^D \tag{10}$$

Rearranging and taking the log of both sides gives $\log(|G|) \leq D \log\left(\frac{\|\boldsymbol{\omega}\|_2}{R} + 1\right)$. Using the fact that $\log(1+x) < x$ for $x \neq 0$, we get $R < \frac{D}{\log(|G|)}\|\boldsymbol{\omega}\|_2$, which contradicts the definition of R. Therefore, the contradicting hypothesis is false, and $\exists g \neq e$ such that $\|\boldsymbol{\omega} - g \cdot \boldsymbol{\omega}\|_2 \leq 2R$. In order to analyze the resulting bound, recall that $D = d_i d_h + d_h + d_h d_o + d_o$ and $\log(|G|) = \log(d_h!) \geq \frac{d_h}{2}\log(\frac{d_h}{2})$. This results in the bound

$$\|\boldsymbol{\omega} - g \cdot \boldsymbol{\omega}\|_2 \leq 2R = \frac{2\left(d_i + d_o + 1 + \frac{1}{d_h}\right)}{\log(\frac{d_h}{2})}\|\boldsymbol{\omega}\|_2 = \mathcal{O}\left(\frac{d_i + d_o}{\log(d_h)}\|\boldsymbol{\omega}\|_2\right) \tag{11}$$

as in Theorem 3.1. $\qquad\square$

C. Sampling From $q_\theta(\boldsymbol{\omega})$ vs. Sampling From $q_\theta^G(\boldsymbol{\omega})$

In Section 4 we introduced the symmetrized variational posterior $q_\theta^G(\boldsymbol{\omega})$. We mention that we can sample from $q_\theta^G(\boldsymbol{\omega})$ by sampling $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$ and transforming it using a random $g \sim \text{Uniform}(G)$, and that the induced function space distribution is the same as the one induced by sampling weights from $q_\theta(\boldsymbol{\omega})$. The following elaborates why this is the case.

Since $q_\theta^G(\boldsymbol{\omega})$ is a uniform mixture of the distributions $\{g_\# q_\theta(\boldsymbol{\omega})\}_{g \in G}$, we can sample from it by sampling $g \sim \text{uniform}(G)$ and then taking $\boldsymbol{\omega} \sim g_\# q_\theta(\boldsymbol{\omega})$. By the definition of the pushforward measure, if $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$, then $\boldsymbol{\omega}' = g \cdot \boldsymbol{\omega} \sim g_\# q_\theta(\boldsymbol{\omega}')$. Therefore, we can sample from $q_\theta^G(\boldsymbol{\omega}')$ by independently sampling $g \sim \text{uniform}(G)$, $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$ and taking $\boldsymbol{\omega}' = g \cdot \boldsymbol{\omega}$. We now consider the functions $\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x})$ and $\mathbf{f}^{\boldsymbol{\omega}'}(\mathbf{x})$, where $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$ and $\boldsymbol{\omega}' = g \cdot \boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega}')$. Since the mapping $\boldsymbol{\omega} \mapsto \mathbf{f}^{\boldsymbol{\omega}}$ is G-invariant (as mentioned in Section 2), the function represented by $\boldsymbol{\omega}'$ is $\mathbf{f}^{\boldsymbol{\omega}'} \equiv \mathbf{f}^{g \cdot \boldsymbol{\omega}} \equiv \mathbf{f}^{\boldsymbol{\omega}}$. In other words the function space distribution induced by $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$ and $\boldsymbol{\omega}' \sim q_\theta^G(\boldsymbol{\omega}')$ are identical.

D. Symmetrization of the ELBO

In Section 4 we introduced the symmetrized ELBO objective $\mathcal{L}_{\text{VI}}^G(\theta)$. Theorem 4.1 establishes a relationship between $\mathcal{L}_{\text{VI}}^G(\theta)$ and $\mathcal{L}_{\text{VI}}(\theta)$ and states that

1. $\mathcal{L}_{\text{VI}}^G(\theta) = \mathcal{L}_{\text{VI}}(\theta) + H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(w))$.

2. $H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(\boldsymbol{\omega})) = I(g; \boldsymbol{\omega})$, where $I(g; \boldsymbol{\omega})$ is the mutual information of $g, \boldsymbol{\omega} \sim q_\theta^G(g, \boldsymbol{\omega})$.

Proof of Theorem 4.1. Recall that

$$\mathcal{L}_{\text{VI}}^G(\theta) = \underbrace{\mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})}\left(\log\left(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})\right)\right)}_{\text{expected log-likelihood term}} - \underbrace{\text{KL}\left(q_\theta^G(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega})\right)}_{\text{prior KL term}} \tag{12}$$

We start by analyzing the expected log-likelihood term.

$$\mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})}\left(\log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega}))\right) = \int_{\mathcal{W}} \log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega}))\frac{1}{|G|}\sum_{g \in G} q_\theta(g^{-1} \cdot \boldsymbol{\omega})\, d\boldsymbol{\omega}$$

$$= \frac{1}{|G|}\sum_{g \in G}\int_{\mathcal{W}} \log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega}))q_\theta(g^{-1} \cdot \boldsymbol{\omega})\, d\boldsymbol{\omega} \tag{13}$$

$$= \frac{1}{|G|}\sum_{g \in G}\int_{g^{-1} \cdot \mathcal{W}} \log(p(\mathbf{Y} \mid \mathbf{X}, g \cdot \boldsymbol{\omega}))q_\theta(\boldsymbol{\omega})|\det(g)|\, d\boldsymbol{\omega}$$

As mentioned in the discussion leading to Proposition 2.2, $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})$ is G-invariant. Additionally, $|\det(g)| = 1$ from the

orthonormality of the representation, and $g^{-1} \cdot \mathcal{W} = \mathcal{W}$ since g^{-1} is invertible. Therefore,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})} \left(\log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})) \right) &= \frac{1}{|G|} \sum_{g \in G} \int_{\mathcal{W}} \log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})) q_\theta(\boldsymbol{\omega}) \, d\boldsymbol{\omega} \\
&= \int_{\mathcal{W}} \log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})) q_\theta(\boldsymbol{\omega}) \, d\boldsymbol{\omega} \\
&= \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})} \left(\log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})) \right)
\end{aligned}
\tag{14}
$$

Thus, $\mathcal{L}_{\mathrm{VI}}^G(\theta)$ and $\mathcal{L}_{\mathrm{VI}}(\theta)$ have identical expected log-likelihood terms. Moving on to the prior KL term

$$
\begin{aligned}
\mathrm{KL}(q_\theta^G(\boldsymbol{\omega}) \,\|\, p(\boldsymbol{\omega})) &= \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})} \left(\log\left(\frac{q_\theta^G(\boldsymbol{\omega})}{p(\boldsymbol{\omega})} \right) \right) \\
&= \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})} \left(-\log(p(\boldsymbol{\omega})) \right) - H(q_\theta^G(\boldsymbol{\omega})) \\
&= -\int \log(p(\boldsymbol{\omega})) \frac{1}{|G|} \sum_{g \in G} q_\theta(g^{-1} \cdot \boldsymbol{\omega}) \, d\boldsymbol{\omega} - H(q_\theta^G(\boldsymbol{\omega})) \\
&= -\frac{1}{|G|} \sum_{g \in G} \int \log(p(\boldsymbol{\omega})) q_\theta(g^{-1} \cdot \boldsymbol{\omega}) \, d\boldsymbol{\omega} - H(q_\theta^G(\boldsymbol{\omega})) \\
&= -\frac{1}{|G|} \sum_{g \in G} \int \log(p(g \cdot \boldsymbol{\omega})) q_\theta(\boldsymbol{\omega}) |\det(g)| \, d\boldsymbol{\omega} - H(q_\theta^G(\boldsymbol{\omega})) \\
&= -\frac{1}{|G|} \sum_{g \in G} \int \log(p(\boldsymbol{\omega})) q_\theta(\boldsymbol{\omega}) \, d\boldsymbol{\omega} - H(q_\theta^G(\boldsymbol{\omega})) \\
&= -\int \log(p(\boldsymbol{\omega})) q_\theta(\boldsymbol{\omega}) \, d\boldsymbol{\omega} - H(q_\theta^G(\boldsymbol{\omega})) \\
&= \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})} \left(-\log(p(\boldsymbol{\omega})) \right) - H(q_\theta^G(\boldsymbol{\omega}))
\end{aligned}
$$

Therefore, since the prior KL term of $\mathcal{L}_{\mathrm{VI}}(\theta)$ is $\mathrm{KL}(q_\theta(\boldsymbol{\omega}) \,\|\, p(\boldsymbol{\omega})) = E_{\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})} \left(-\log(p(\boldsymbol{\omega})) \right) - H(q_\theta(\boldsymbol{\omega}))$, we have

$$
\mathrm{KL}(q_\theta^G(\boldsymbol{\omega}) \,\|\, p(\boldsymbol{\omega})) = \mathrm{KL}(q_\theta(\boldsymbol{\omega}) \,\|\, p(\boldsymbol{\omega})) - H(q_\theta^G(\boldsymbol{\omega})) + H(q_\theta(\boldsymbol{\omega}))
\tag{15}
$$

Putting it all together, we get

$$
\mathcal{L}_{\mathrm{VI}}^G(\theta) = \mathcal{L}_{\mathrm{VI}}(\theta) + H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(\boldsymbol{\omega}))
\tag{16}
$$

which finishes the proof of statement 1. All that is left, is to show that indeed $I(g; \boldsymbol{\omega}) = H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(\boldsymbol{\omega}))$. In order to do so, we need the following lemma

Lemma D.1. *Let $\mathcal{V}$ be a finite dimensional inner product space. If $p(\mathbf{v})$ is a probability measure on $\mathcal{V}$ and $T \in O(\mathcal{V})$ is a norm-preserving linear operator on $\mathcal{V}$, then $H(T_\# p(\mathbf{v})) = H(p(\mathbf{v}))$.*

Lemma D.1 is proved later in Appendix D. Recall that G acts on $\mathcal{W}$ via an orthonormal representation and that that $q_\theta^G(\boldsymbol{\omega} \mid g) = g_\# q_\theta(\boldsymbol{\omega})$. Therefore, by the definition of mutual information, and using Lemma D.1, we get

$$
\begin{aligned}
I(g; \boldsymbol{\omega}) &= \mathbb{E}_{g, \boldsymbol{\omega} \sim q_\theta^G(g, \boldsymbol{\omega})} \left(\log\left(\frac{q_\theta^G(\boldsymbol{\omega} \mid g)}{q_\theta^G(\boldsymbol{\omega})} \right) \right) \\
&= \mathbb{E}_{g, \boldsymbol{\omega} \sim q_\theta^G(g, \boldsymbol{\omega})} \left(\log\left(q_\theta^G(\boldsymbol{\omega} \mid g) \right) \right) - \mathbb{E}_{g, \boldsymbol{\omega} \sim q_\theta^G(g, \boldsymbol{\omega})} \left(\log\left(q_\theta^G(\boldsymbol{\omega}) \right) \right) \\
&= \mathbb{E}_{g \sim q_\theta^G(g)} \left(\mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega}|g)} \left(\log\left(q_\theta^G(\boldsymbol{\omega} \mid g) \right) \right) \right) - \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})} \left(\mathbb{E}_{g \sim q_\theta^G(g|\boldsymbol{\omega})} \left(\log\left(q_\theta^G(\boldsymbol{\omega}) \right) \right) \right) \\
&= -\mathbb{E}_{g \sim q_\theta^G(g)} \left(H(q_\theta^G(\boldsymbol{\omega} \mid g)) \right) - \mathbb{E}_{\boldsymbol{\omega} \sim q_\theta^G(\boldsymbol{\omega})} \left(\log\left(q_\theta^G(\boldsymbol{\omega}) \right) \right) \\
&= -\mathbb{E}_{g \sim q_\theta^G(g)} \left(H(g_\# q_\theta(\boldsymbol{\omega})) \right) + H(q_\theta^G(\boldsymbol{\omega})) \\
&= -H(q_\theta(\boldsymbol{\omega})) + H(q_\theta^G(\boldsymbol{\omega}))
\end{aligned}
$$

$\square$

All that is left is to prove Lemma D.1.

Proof of Lemma D.1. With the notations of lemma D.1, and using the change of variable formula for pushforward measures:

$$
\begin{aligned}
H(T_{\#}p(\mathbf{v})) &= \int_{\mathcal{V}} p\left(T^{-1}(\mathbf{v})\right) \log\left(p\left(T^{-1}(\mathbf{v})\right)\right) \, d\mathbf{v} \\
&= \int_{T(\mathcal{V})} p(\mathbf{v}) \log\left(p(\mathbf{v})\right) \left|\det(D_{\mathbf{v}}T(\mathbf{v}))\right| d\mathbf{v} \\
&= \int_{\mathcal{V}} p(\mathbf{v}) \log\left(p(\mathbf{v})\right) \left|\det(T)\right| d\mathbf{v} \\
&= \int_{\mathcal{V}} p(\mathbf{v}) \log\left(p(\mathbf{v})\right) \, d\mathbf{v} \\
&= H(p(\mathbf{v}))
\end{aligned}
$$

where we used the fact that $T \in O(\mathcal{V}) \Rightarrow T(\mathcal{V}) = \mathcal{V}$ and $\det(T) = \pm 1$. $\qquad\square$

E. Mutual Information Estimation

Mutual information (MI) is a fundamental concept in information theory that measures the dependence between two random variables. It plays a crucial role in various areas of machine learning, including representation learning, reinforcement learning, deep learning theory, and more. Given random variables $\mathbf{x}, \mathbf{y}$, with a joint probability distribution $p(\mathbf{x}, \mathbf{y})$, their mutual information is defined as:

$$
I(\mathbf{x}; \mathbf{y}) = \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})} \left(\frac{p(\mathbf{x} \mid \mathbf{y})}{p(\mathbf{x})} \right) = \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})} \left(\frac{p(\mathbf{y} \mid \mathbf{x})}{p(\mathbf{y})} \right) \tag{17}
$$

Depending on the problem, estimating MI directly can be challenging. Therefore, various variational bounds have been established to approximate MI. See (Poole et al., 2019) for a survey. In our case, in order to estimate $I(g; \boldsymbol{\omega})$, we leverage the fact that conditional $q_\theta^G(\boldsymbol{\omega} \mid g) = q_\theta(g^{-1} \cdot \boldsymbol{\omega})$ is tractable, and use the InfoNCE bound.

Theorem E.1 (InfoNCE bound with optimal critic (Poole et al., 2019)). *Given two random variables $\mathbf{x}, \mathbf{y}$ distributed by a joint probability function $p(\mathbf{x}, \mathbf{y})$ the mutual information $I(\mathbf{x}; \mathbf{y})$ satisfies*

$$
I(\mathbf{x}; \mathbf{y}) \geq I^K(\mathbf{x}; \mathbf{y}) = \mathbb{E} \left(\frac{1}{K} \sum_{i=1}^{K} \log \left(\frac{p(\mathbf{x}_i \mid \mathbf{y}_i)}{\frac{1}{K} \sum_{j=1}^{K} p(\mathbf{x}_i \mid \mathbf{y}_j)} \right) \right) \tag{18}
$$

Where the expectation is over K i.i.d samples from $p(\mathbf{x}, \mathbf{y})$, i.e. over $\Pi_{i=1}^{K} p(\mathbf{x}_i, \mathbf{y}_i)$. Moreover, $\lim_{K \to \infty} I^K(\mathbf{x}; \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$.

Proof for $\lim_{K \to \infty} I^K(\mathbf{x}, \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$ can be found in (Chen et al., 2021). The InfoNCE bound offers a good trade-off between bias and variance compared to other MI estimation methods (Song & Ermon, 2020). However, reducing bias requires increasing the number of samples K, which has quadratic cost in K. Fortunately, in our case, the specific structure of $q_\theta^G(\boldsymbol{\omega})$ allows us to simplify Equation (18) and obtain a more efficient estimator, presented in Theorem 4.3.

Proof of Theorem 4.3. Plugging in $q_\theta^G(g, \boldsymbol{\omega})$ to Theorem E.1 yields

$$
I^K(\theta) = I^K(g; \boldsymbol{\omega}) = \mathbb{E} \left(\frac{1}{K} \sum_{i=1}^{K} \log \left(\frac{q_\theta^G(\boldsymbol{\omega}_i \mid g_i)}{\frac{1}{K} \sum_{j=1}^{K} q_\theta^G(\boldsymbol{\omega}_i \mid g_j)} \right) \right) \tag{19}
$$

where the expectation is over $g_1, \ldots, g_K \sim \text{uniform}(G)$ i.i.d and $\boldsymbol{\omega}_i \sim q_\theta^G(\boldsymbol{\omega} \mid g_i)$. Note that we can sample $\boldsymbol{\omega}_i \sim q_\theta^G(\boldsymbol{\omega} \mid g_i)$ by sampling $\boldsymbol{\omega}_i' \sim q_\theta(\boldsymbol{\omega})$ and taking $\boldsymbol{\omega}_i = g_i \cdot \boldsymbol{\omega}_i'$. Using this fact, Equation (19) becomes

$$
\begin{aligned}
I^K(\theta) &= \mathbb{E} \left(\frac{1}{K} \sum_{i=1}^{K} \log \left(\frac{q_\theta(\boldsymbol{\omega}_i')}{\frac{1}{K} \sum_{j=1}^{K} q_\theta(g_j^{-1} g_i \cdot \boldsymbol{\omega}_i')} \right) \right) \\
&= -\mathbb{E} \left(\frac{1}{K} \sum_{i=1}^{K} \log \left(\frac{1}{K} \sum_{j=1}^{K} q_\theta(g_j^{-1} g_i \cdot \boldsymbol{\omega}_i') \right) \right) - H(q_\theta(\boldsymbol{\omega}))
\end{aligned} \tag{20}
$$

15

where the expectation is over $g_1, \ldots, g_K \sim \text{uniform}(G)$ i.i.d and $\boldsymbol{\omega}'_1, \ldots, \boldsymbol{\omega}'_K \sim q_\theta(\boldsymbol{\omega})$ i.i.d. We can further simplify $I^K(\theta)$ using the group structure of G, to that end we write

$$
I^K(\theta) = -H(q_\theta(\boldsymbol{\omega})) - \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\omega}'_i, g_1, \ldots, g_K} \left(\log \left(\frac{1}{K} \left(q_\theta(\boldsymbol{\omega}'_i) + \sum_{j \neq i} q_\theta(g_j^{-1} g_i \cdot \boldsymbol{\omega}'_i) \right) \right) \right)
$$

$$
= -H(q_\theta(\boldsymbol{\omega})) - \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{g_i} \left(\overbrace{\mathbb{E}_{\boldsymbol{\omega}'_i, g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_K} \left(\log \left(\frac{1}{K} \left(q_\theta(\boldsymbol{\omega}'_i) + \sum_{j \neq i} q_\theta(g_j^{-1} g_i \cdot \boldsymbol{\omega}'_i) \right) \right) \right)}^{A(g_i)} \right) \tag{21}
$$

Since G is a group, inverses and right multiplication are bijective transformations. This implies that for a given $i \in \{1, \ldots, K\}$, the random variables $\{g_j^{-1} g_i \mid j \neq i\}$ are i.i.d uniformly on G conditioned on g_i. Therefore, $A(g_i)$ can be written as

$$
A(g_i) = \mathbb{E} \left(\log \left(\frac{1}{K} \left(q_\theta(\boldsymbol{\omega}'_i) + \sum_{j=1}^{K-1} q_\theta(g'_j \cdot \boldsymbol{\omega}'_i) \right) \right) \right) \tag{22}
$$

where the expectation is over $\boldsymbol{\omega}'_i \sim q_\theta(\boldsymbol{\omega})$ and $g'_1, \ldots, g'_{K-1} \sim \text{uniform}(G)$ i.i.d. Notice that $A(g_i)$ is independent of g_i! Therefore $I^K(\theta)$ simplifies to

$$
I^K(\theta) = -H(q_\theta(\boldsymbol{\omega})) - \mathbb{E} \left(\log \left(\frac{1}{K} \left(q_\theta(\boldsymbol{\omega}) + \sum_{j=1}^{K-1} q_\theta(g_j \cdot \boldsymbol{\omega}) \right) \right) \right) \tag{23}
$$

Where the expectation is over $\boldsymbol{\omega} \sim q_\theta(\boldsymbol{\omega})$ and $g_1, \ldots, g_{K-1} \sim \text{uniform}(G)$. Again using the fact that inversion in bijective in groups, this can be rewritten as in Theorem 4.3

$$
I^K(\theta) = -H(q_\theta(\boldsymbol{\omega})) - \mathbb{E} \left(\log \left(\frac{1}{K} \left(q_\theta(\boldsymbol{\omega}) + \sum_{j=1}^{K-1} q_\theta(g_j^{-1} \cdot \boldsymbol{\omega}) \right) \right) \right) \tag{24}
$$

Theorem 4.3, which is phrased in terms of $H^K(\theta)$, can be recovered from the above, using the fact that $I(g; \boldsymbol{\omega}) = H(q_\theta^G(\boldsymbol{\omega})) - H(q_\theta(\boldsymbol{\omega}))$. $\qquad \square$