
VARIATIONAL BAYESIAN NEURAL NETWORKS VIA RESOLUTION OF SINGULARITIES

A PREPRINT

 **Susan Wei**

School of Mathematics and Statistics
University of Melbourne
Melbourne, Victoria
susan.wei@unimelb.edu.au

 **Edmund Lau**

School of Mathematics and Statistics
University of Melbourne
Melbourne, Victoria
elau1@student.unimelb.edu.au

ABSTRACT

In this work, we advocate for the importance of singular learning theory (SLT) as it pertains to the theory and practice of variational inference in Bayesian neural networks (BNNs). To begin, using SLT, we lay to rest some of the confusion surrounding discrepancies between downstream predictive performance measured via e.g., the test log predictive density, and the variational objective. Next, we use the SLT-corrected asymptotic form for singular posterior distributions to inform the design of the variational family itself. Specifically, we build upon the idealized variational family introduced in Bhattacharya et al. [2020] which is theoretically appealing but practically intractable. Our proposal takes shape as a normalizing flow where the base distribution is a carefully-initialized generalized gamma. We conduct experiments comparing this to the canonical Gaussian base distribution and show improvements in terms of variational free energy and variational generalization error.

Keywords Normalizing Flow · Real Log Canonical Threshold · Singular Learning Theory · Singular Models · Test log-likelihood · Variational Free Energy · Variational Inference · Variational Generalization Error

1 Introduction

A Bayesian neural network (BNN) Mackay [1995] is a neural network endowed with a prior distribution φ on its weights w . Despite their theoretical appeal Lampinen and Vehtari [2001], Wang and Yeung [2020], applying BNNs in practice is not without significant challenges. MCMC and its variants, while widely considered the gold standard, can be prohibitively expensive in terms of computation. On the other hand, fast alternatives such as variational inference may result in *uncontrolled* approximations.

In this work, we mine insights from **singular learning theory** (SLT) Watanabe [2009] to explain and improve upon certain aspects of BNNs. Roughly speaking, a model is (strictly) **singular** if the parameter-to-model mapping is not one-to-one and the likelihood function does not look Gaussian¹. That neural networks are singular is well documented Sussmann [1992], Watanabe [2000, 2001], Fukumizu [2003], Watanabe [2007]. We refer the readers to Wei et al. [2022] for a detailed proof in the case of a standard feedforward network. The singular nature of BNNs has interesting implications for the posterior distribution, see Figure 1.

Let (x, y) denote the input-target pair modeled jointly as $p(x, y|w) = p(y|x, w)p(x)$ where $w \in \mathbb{R}^d$ is the model parameter. Let $p(y|x, w)$ be a neural network model with functional model f , by which we mean $y = f(x, w) + \epsilon$ where ϵ is some random variable. For example, if we have Gaussian additive noise ϵ , the conditional distribution could be modelled as $\mathcal{N}(y|f(x, w), \sigma^2 I)$ where f is a feedforward ReLU network with weights w .

The central quantity of interest in BNNs is the intractable posterior distribution over the neural network weights,

$$p(w|\mathcal{D}_n) = \frac{\prod_{i=1}^n p(y_i|x_i, w)\varphi(w)}{Z(n)},$$

¹These features should not be viewed as pathological, see “Deep learning is singular and that’s good” by Wei et al. [2022].

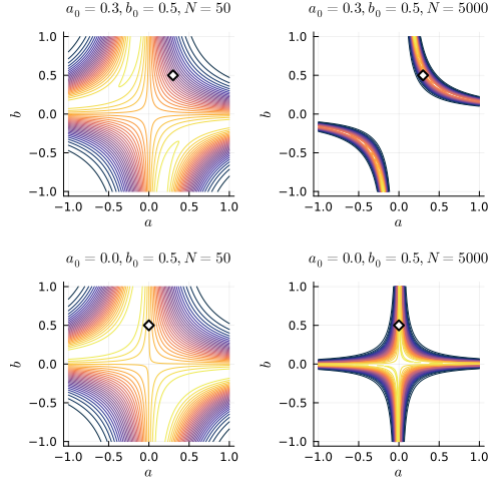


Figure 1: Posterior density contour plot for a 2D tanh-regression model, $p(y|x, a, b) \propto \exp(y - a \tanh(bx))$. The white diamond marks the true parameter (a_0, b_0) used to generate the dataset \mathcal{D}_n . Each row shows a different true distribution, while each column shows a different sample size n . When $a_0 b_0 = 0$ as in the second row, the set of true parameters W_0 is not a singleton and contains a singularity at the origin. It is worth noticing that, for a singular model, even when the truth is not at a singularity (first row), the posterior is still far from being locally Gaussian even at sample size $n = 5000$.

where $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ is a dataset of n input-output pairs. The normalizing constant,

$$Z(n) = \int \prod_{i=1}^n p(y_i|x_i, w) \varphi(w) dw,$$

is variously known as the **model evidence** and the **marginal likelihood**. Define the **empirical entropy** of the training data,

$$S_n = -\frac{1}{n} \sum_{i=1}^n \log p_0(y_i|x_i).$$

We shall call

$$\bar{Z}(n) = -\log Z(n) - nS_n$$

the **normalized evidence**. Let us call $F(n) := -\log Z(n)$ the **Bayes free energy** and $\bar{F}(n) := -\log \bar{Z}(n)$ its normalized version.

Unlike prediction in traditional neural networks, prediction in BNNs proceeds by marginalization, i.e., averaging over all possible values of the network weights. Namely, prediction in BNNs makes use of the **Bayes posterior predictive distribution**,

$$p(y|x, \mathcal{D}_n) := \int p(y|x, w) p(w|\mathcal{D}_n) dw. \quad (1)$$

With (1), we can calculate prediction uncertainties as well as obtain better calibrated predictions Heek [2018], Osawa et al. [2019], Maddox et al. [2019].

In Section 3, we recapitulate from the perspective of SLT the predictive advantages of BNNs over traditional neural networks. Specifically, SLT shows that the Bayes posterior predictive distribution in (1) has lower generalization error compared to MLE or MAP point estimates.

Despite compelling arguments for employing BNNs, we must reckon with the fact that they can only ever be applied *approximately*. Among approximate techniques, a major class is represented by scaling classic MCMC to modern settings of large datasets and deep neural networks Welling and Teh [2011], Chen et al. [2014], Zhang et al. [2020]. In this paper, we instead turn our focus to variational inference, which is particularly suited to scaling BNNs to large datasets.

All variational inference techniques are characterized by two ingredients. First, a family of densities \mathcal{Q} , often called the variational family, is posited. Second, some $q^* \in \mathcal{Q}$ is found via optimization according to some criterion that measures

closeness to the desired target density. In this work, we seek to approximate the posterior density and we will employ the conventional Kullback-Leibler divergence. This leads to the optimization problem,

$$\min_{q \in \mathcal{Q}} \text{KL}(q(w) \parallel p(w|\mathcal{D}_n)). \quad (2)$$

This is equivalent to minimizing the so-called **normalized² variational free energy (VFE)**,

$$\bar{F}_{vb}(n) := \mathbb{E}_q n K_n(w) + \text{KL}(q(w) \parallel \varphi(w)).$$

It is easy to see that $\bar{F}_{vb}(n) \geq \bar{F}(n)$ with equality if and only if the variational distribution is exactly equal to the posterior. Readers are likely more familiar with the variational objective of maximizing the so-called **evidence lower bound (ELBO)** which is simply related to the (normalized) VFE via $\text{ELBO} = -\bar{F}_{vb}(n)$.

Let $q^* \in \mathcal{Q}$ be a minimizer of (2). Let us call the variational approximation to (1) given by

$$p_{vb}(y|x, \mathcal{D}_n) := \int p(y|x, w) q^*(w) dw, \quad (3)$$

the **induced predictive distribution**. We can measure the predictive accuracy of p_{vb} using once again the KL divergence, i.e.,

$$G_n(p_{vb}(y|x, \mathcal{D}_n)) := \text{KL}(p_0(y|x) \parallel p_{vb}(y|x, \mathcal{D}_n)),$$

which we shall call the **variational generalization error (VGE)**. Per the discussion in Section 3, this is, up to a constant and a sign flip, nothing more than the typical **test log predictive density** Gelman et al. [2014] commonly employed in variational inference evaluation.

We shall see in Section 4 that, surprisingly, the VGE may be arbitrarily high even for a variational family whose minimum VFE is close to optimality. In other words, it is not guaranteed that minimizing (2) results in good downstream predictive performance. The outlook is not entirely bleak. Depending on the relationship between two critical quantities of variational inference – the **MVFE coefficient** λ_{vfe} and the **VGE coefficient** λ_{vge} – the generalization error of the induced predictive distribution may be controllable via minimizing the VFE.

Clarification of the relationship between the two variational coefficients for most common variational learning problems is an open problem, which we leave aside for future work. We will assume the variational coefficients are related *favorably*, in a manner which will be made clear in Section 4, and proceed to design a variational family whose **variational approximation gap** is small. The proposal is predicated on an important SLT result which states that, roughly speaking, the posterior distribution over the parameters of a singular model is not asymptotically Gaussian, but can still be put into an explicit standard form via the **resolution of singularities**.

2 Singular learning theory

In this section, we give a succinct overview of key concepts from SLT. We focus in particular on what SLT has to say about the behavior of the posterior distribution in strictly singular models. Let us assume the parameter space W is a compact set in \mathbb{R}^d and $p_0(x, y) = p_0(y|x)p(x)$ is the true data-generating mechanism. Throughout, we suppose there exists $w_0 \in W$ such that $p_0(y|x) = p(y|x, w_0)$. In the parlance of SLT, this condition is known as **realizability**. Let $\varphi(w)$ be a compactly-supported prior. We shall refer to $(p(\cdot, \cdot), p_0(\cdot, \cdot), \varphi(\cdot))$ as a **model-truth-prior triplet**. The roles played by compactness and realizability in singular learning theory are discussed in Appendix A.

Define $K(w)$ to be the Kullback-Leibler divergence between the truth and the model, i.e.,

$$K(w) := \text{KL}(p_0(x, y) \parallel p(x, y|w)).$$

Following Watanabe [2009], we say a model is **regular** if 1) it is identifiable, i.e., the map $w \mapsto p(\cdot, \cdot|w)$ from parameter to model is one-to-one and 2) its Fisher information matrix $I(w)$ is positive definite for arbitrary $w \in W$. We call a model **strictly singular** if it is not regular. The term singular will refer to either regular or strictly singular models. See Figure 1 for an example of a strictly singular model with two truth settings. This figure illustrates an important lesson: for strictly singular models, even when the true parameter set $W_0 := \{w : K(w) = 0\}$ does not contain singularities, the posterior distribution is still far from Gaussian.

The following theorem from Watanabe [2009], adapted for notational consistency, gives precise conditions for the existence of **resolution maps**, algebraic-geometrical transformations which enables $K(w)$ to be locally written as a

²Throughout this paper, we work with normalized quantities for ease of exposition. The asymptotics presented hold equally for the unnormalized counterparts.

monomial, i.e., a product of powers of variables such as in the right-hand-side of (4). The result is itself based on Hironaka's resolution of singularities, a celebrated result in modern algebraic geometry.

To prepare, let $W_\epsilon = \{w \in W : K(w) \leq \epsilon\}$ for some small positive constant ϵ and $W_\epsilon^{(R)}$ be some real open set such that $W_\epsilon \subset W_\epsilon^{(R)}$. The theorem below will make use of the multi-index notation: for a given $\xi = (\xi_1, \dots, \xi_d) \in \mathbb{R}^d$, define $w^{\mathbf{k}} := w_1^{k_1} \cdots w_d^{k_d}$ where the multi-index $\mathbf{k} = (k_1, \dots, k_d)$ with each k_j a nonnegative integer. Due to space constraints, Fundamental Conditions I and II required below are stated and discussed in Appendix A.

Theorem 2.1 (Theorem 6.5 of Watanabe [2009]). *Suppose the model-truth-prior triplet (p, p_0, φ) satisfies Fundamental Conditions I and II with $s = 2$. We can find a real analytic manifold $M^{(R)}$ and a proper and real analytic map $g : M^{(R)} \rightarrow W_\epsilon^{(R)}$ such that*

1. $M = g^{-1}(W_\epsilon)$ is covered by a finite set $M = \cup_\alpha M_\alpha$ where $M_\alpha = [0, b]^d$.

2. In each M_α ,

$$K(g(\xi)) = \xi^{2\mathbf{k}} = \xi_1^{2k_1} \cdots \xi_d^{2k_d}, \quad (4)$$

where $k_j \in \mathbb{N}, j = 1, \dots, d$ are such that not all k_j are zero.

3. There exists a C^∞ function $b(\xi)$ such that

$$\varphi(g(\xi))|g'(\xi)| = \xi^{\mathbf{h}} b(\xi) = \xi_1^{h_1} \cdots \xi_d^{h_d} b(\xi), \quad (5)$$

where $h_j \in \mathbb{N}, j = 1, \dots, d$, $|g'(\xi)|$ is the absolute value of the determinant of the Jacobian and $b(\xi) > c > 0$ for $\xi \in [0, b]^d$.

In Theorem 2.1 we have suppressed the dependency on the manifold chart index α , but the reader should keep in mind that the maps g and the multi-indices are all indexed by α . It is also important to recognize that none of these said quantities are unique for a given triplet (p, p_0, φ) .

A crucial quantity that appears in SLT is a rational number in $(0, d/2]$ known as the **real log canonical threshold** (RLCT). Let $\{M_\alpha : \alpha\}$ be as in Theorem 2.1 and define

$$\lambda_j = \frac{h_j + 1}{2k_j}, j = 1, \dots, d$$

where h_j and k_j are the entries of the multi-indices \mathbf{h} and \mathbf{k} in a local coordinate M_α . When $k_j = 0$, λ_j is taken to be infinity.

Uniquely associated to a triplet (p, p_0, φ) are its real log canonical threshold (RLCT) and its multiplicity defined, respectively, as

$$\lambda = \min_\alpha \min_{j \in \{1, \dots, d\}} \lambda_j, \quad m = \max_\alpha \#\{j : \lambda_j = \lambda\}. \quad (6)$$

Let $\{\alpha^*\}$ be the set of those local coordinates in which both the min and max in (6) are attained. Watanabe [2009] calls this set the **essential coordinates** and the corresponding collection $\{M_\alpha\}$ the **essential charts**.

If $\{w : K(w) = 0, \varphi(w) > 0\}$ is not the empty set, the RLCT of a model-truth-prior triplet is *at most* $d/2$ [Watanabe, 2009, Theorem 7.2]. When the model is regular, the RLCT is *exactly equal* to $d/2$ and the multiplicity $m = 1$ [Watanabe, 2009, Remark 1.15]. In fact, (twice the) RLCT may be regarded as the effective degrees of freedom in strictly singular models [Wei et al., 2022]. The RLCT also shows up in important asymptotic results, see (10) and (11).

Henceforth, to make clear that the RLCT and multiplicity are invariants of the model-truth-prior triplet, we shall write $\lambda(p, p_0, \varphi)$ and $m(p, p_0, \varphi)$ to mark this dependence. In Appendix B, we recall a simple toy network, a two-parameter tanh network, where the resolution map, the RLCT, and the multiplicity can be calculated explicitly.

2.1 Posterior distribution in singular models

The posterior distribution in strictly singular models is decidedly not Gaussian. The correct asymptotic form can be derived using SLT. For a particular manifold chart index α , let us apply the transformation $g_\alpha(\xi) = w$ and rewrite the posterior distribution in the new coordinate ξ ,

$$p(\xi | \mathcal{D}_n) = \frac{\exp(-nK_n(g_\alpha(\xi)))\varphi(g_\alpha(\xi))|g'_\alpha(\xi)|}{\bar{Z}(n)}, \quad (7)$$

with

$$K_n(w) = \frac{1}{n} \sum_{i=1}^n \log \frac{p_0(y_i|x_i)}{p(y_i|x_i, w)}$$

denoting the sample average log likelihood ratio. Note $K_n(w)$ is the empirical counterpart to $K(w)$.

By (cheekily) substituting (4) (5) into $p(\xi|\mathcal{D}_n)$ in (7), we obtain that the posterior distribution for large n , in the chart M_α , is described as a so-called **standard form** Watanabe [2018]:

$$\exp(-n\xi_1^{2k_1}\xi_2^{2k_2}\dots\xi_d^{2k_d})|\xi_1^{h_1}\dots\xi_d^{h_d}|b(\xi).$$

In other words, the posterior distribution over the parameters of a singular model can be transformed into a mixture of standard forms, asymptotically. In Figure 1 we display the singular posterior density contour plot for a toy 2D tanh-neural network in two settings of the true distribution.

3 The Bayes posterior predictive distribution

Let the generalization error of some predictive distribution $\hat{p}_n(y|x)$, estimated from a training set \mathcal{D}_n , be measured using the KL divergence:

$$G_n(\hat{p}_n(y|x)) := \text{KL}(p_0(y|x)p(x) \parallel \hat{p}_n(y|x)p(x)) \quad (8)$$

In the machine learning community, this goes by another name: $G_n(\cdot)$ is, up to a constant and a sign flip, the population counterpart to the commonly reported test log-likelihood, aka the predictive log-likelihood or **test log-predictive density**. This can be seen by writing

$$\hat{G}_n = -\frac{1}{n'} \sum_{(x,y) \in \mathcal{D}_{n'}} (\log p_0(y|x) - \log \hat{p}_n(y|x)) \quad (9)$$

where $\mathcal{D}_{n'}$ is an independent dataset.¹ According to Theorems 1.2 and 7.2 in Watanabe [2009], we have, for the Bayes posterior predictive distribution (1),

$$\mathbb{E}G_n(p(y|x, \mathcal{D}_n)) = \lambda(p, p_0, \varphi)/n + o(1/n) \quad (10)$$

where the expectation is taken with respect to \mathcal{D}_n . We will call the left hand side (10) the expected **Bayes generalization error**. This can be contrasted to the expected generalization error of MLE (and similarly of MAP), which Theorem 6.4 of Watanabe [2009] shows to be $\mathbb{E}G_n(p(y|x, \hat{w}_{mle})) = S/n + o(1/n)$ where S , the maximum of a Gaussian process, can be much larger than $\lambda(p, p_0, \varphi)$. The situation is markedly different for regular models, where differences between the three estimators become negligible in the large- n regime.

We briefly outline the derivation of (10) as it will inform the narrative on the VGE in the next section. First, for the normalized Bayes free energy, under the Fundamental Conditions I and II discussed in A, it was proven in [Watanabe, 2009, Main Theorem 6.2] that the following asymptotic expansion holds

$$\bar{F}(n) = \lambda(p, p_0, \varphi) \log n + (m-1) \log \log n + O_P(1). \quad (11)$$

The result in (10) is then proven using the above expansion together with the well known relationship between the Bayes generalization error and the (normalized) Bayes free energy [Watanabe, 2009, Theorem 1.2]:

$$\mathbb{E}G_n(p(y|x, \mathcal{D}_n)) = \mathbb{E}\bar{F}(n+1) - \mathbb{E}\bar{F}(n). \quad (12)$$

where on the right-hand side, the first expectation is with respect to dataset \mathcal{D}_{n+1} and the second \mathcal{D}_n . Due to this relationship, the Bayes free energy shares the *same* coefficient as the Bayes generalization error.

4 A tale of two variational coefficients

Most applications of variational inference in BNNs labor under the following implicit assumptions: 1) optimizers of the variational objective in (2) have good induced predictive distributions, and 2) two variational families can be compared according to the performance of their induced predictive distributions. A look at the experimental sections of various works on variational BNNs reveal that these assumptions underlie standard practice Blundell et al. [2015], Rezende and Mohamed [2015], Louizos and Welling [2016, 2017], Osawa et al. [2019], Swiatkowski et al. [2020]. We shall see in this section that these two assumptions do not always hold.

Let us associate to a variational family \mathcal{Q} its **normalized minimum variational free energy (MVFE)**,

$$\bar{F}_{vb}^*(n) := \min_{q \in \mathcal{Q}} \bar{F}_{vb}(n).$$

Asymptotics for the MVFE have so far been addressed on a case-by-case basis for certain models and certain variational families, e.g., Gaussian mean-field variational families for reduced rank regression Nakajima and Watanabe [2007], nonnegative matrix factorization Kohjima and Watanabe [2017], Hayashi [2020], normal mixture model Watanabe and Watanabe [2006], hidden Markov model Hosino et al. [2005]. In all the cited instances above, the asymptotic expansion of the **average normalized MVFE** takes the form

$$\mathbb{E}\bar{F}_{vb}^*(n) = \lambda_{vfe} \log n + o(\log n) \quad (13)$$

where the expectation is taken over datasets \mathcal{D}_n . Note that $\lambda_{vfe} \geq \lambda(p, p_0, \varphi)$ necessarily Nakajima and Watanabe [2007]. Because the **variational approximation gap**,

$$\mathcal{G} := \bar{F}_{vb}^*(n) - \bar{F}(n), \quad (14)$$

is the difference of the (normalized) MVFE and the (normalized) Bayes free energy, the gap boils down to the difference between two coefficients:

$$\mathcal{G} \approx (\lambda_{vfe} - \lambda(p, p_0, \varphi)) \log n.$$

Now, under some natural conditions³, the VGE admits the asymptotic expansion,

$$\mathbb{E}G_n(p_{vb}(y|x, \mathcal{D}_n)) = \lambda_{vge}/n + o(1/n). \quad (15)$$

Importantly, $\lambda_{vge} \neq \lambda_{vfe}$ in general, e.g., Nakajima and Watanabe [2007]. This is in contrast to the Bayesian posterior predictive distribution in (1), where the coefficient of the leading $O(1/n)$ term is precisely the RLCT, $\lambda(p, p_0, \varphi)$. That $\lambda_{vge} \neq \lambda_{vfe}$ results from the fact that the relationship (12) is not valid when a variational approximation to the posterior is employed.

In Figure 2a, we illustrate the three possible configurations of the coefficients $\lambda(p, p_0, \varphi)$, λ_{vfe} , λ_{vge} for a given variational family \mathcal{Q} and a model-truth-prior triplet. When $\lambda_{vfe} > \lambda_{vge}$, we call the setting **favorable** since minimizing the VFE offers control over the VGE. When $\lambda_{vfe} < \lambda_{vge}$, we call this **unfavorable** since achieving even a small variational approximation gap could result in an induced predictive distribution with high generalization error. The distribution of favorable versus unfavorable settings in practice is unclear, as the exact relationship between λ_{vfe} and λ_{vge} has been derived in a limited number of works. The results in Nakajima and Watanabe [2007] on linear neural networks, aka reduced rank regression, show there are both favorable and unfavorable settings depending on the input and output dimension, the number of hidden units, and a rank measurement on the truth.

Note that even in favorable settings, we must be careful when comparing two variational families \mathcal{Q}_1 and \mathcal{Q}_2 . Figure 2b illustrates a scenario where the family \mathcal{Q}_1 incurs a smaller variational approximation gap than \mathcal{Q}_2 , but the induced predictive distribution of \mathcal{Q}_1 has λ_{vge} higher than that of \mathcal{Q}_2 . This shows that comparing different variational approximations by their test log predictive density is fraught with potential misinterpretations. In order to control the downstream predictive performance, it is thus important to find a variational family with a small approximation gap, so that we can inherit (and sometimes even beat!) the predictive advantages of the exact Bayes posterior predictive distribution (1), i.e., achieve $\lambda_{vge} < \lambda(p, p_0, \varphi)$.

5 Related work

Although the perspective on offer here – that the discrepancy between test log predictive density and the variational objective *amounts to the relationship between two variational coefficients* – is novel, we are not the first to point out this general phenomenon in variational inference Yao et al. [2018], Huggins et al. [2020], Deshpande et al. [2022], Dhaka et al. [2020]. This phenomenon is also documented in the specific setting of variational inference for BNNs Heek [2018], Yao et al. [2019], Krishnan and Tickoo [2020], Foong et al. [2020]. For instance, Foong et al. [2020] demonstrated in experiments that optimizing the ELBO may not lead to accurate predictive means or variances.

Another area of active research in variational BNNs is the design of the variational family itself. For the large part, the mean-field family of fully factorized Gaussian distributions is still predominant in the general practice of variational inference [Graves, 2011, Blundell et al., 2015, Hernandez-Lobato et al., 2016, Li and Turner, 2016, Khan et al., 2018, Sun et al., 2019]. The mean-field assumption is mostly adopted for computational ease, though the limitations are well known [MacKay, 1992, Coker et al., 2022]. Moving beyond mean-field Gaussian, we can find works that make use of more realistic covariance structures [Louizos and Welling, 2016, Zhang et al., 2018] or more expressive approximating families, e.g., via normalizing flows [Louizos and Welling, 2017, Papamakarios et al., 2021].

Finally, we note there have been a few recent works that recognize the non-identifiability of deep learning models Moore [2016], Pourzanjani et al. [2017], Kurle et al. [2022]. These works however seem to treat the non-identifiability as an issue to be fixed.

³The predictive distribution should be consistent as n goes to infinity, see the discussion in Chapter 13 of Nakajima et al. [2019]

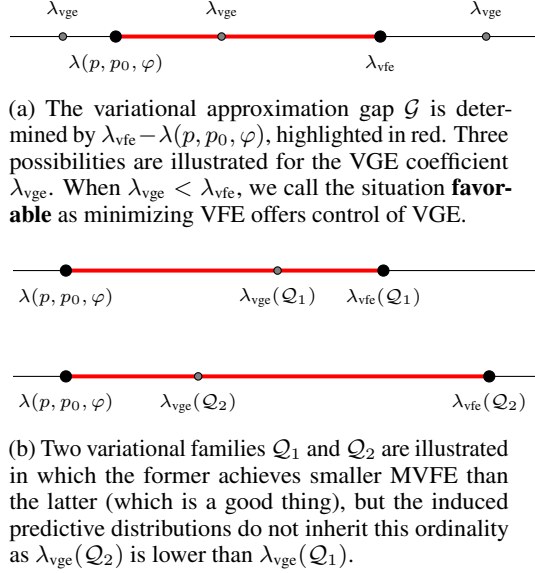


Figure 2: We show in these schematics that evaluating variational approximations to BNNs according to their induced predictive distribution is fraught with potential misinterpretations.

6 Methodology

To achieve a good variational approximation, conventional wisdom says to make \mathcal{Q} as “expressive” as possible. We will approach the design of the variational family in a more principled manner using SLT. To this end, we rely on recent work in Bhattacharya et al. [2020] which leveraged SLT to produce an *idealized* variational family as follows. Let \mathcal{Q}_0 be a family consisting of generalized gamma distributions in \mathbb{R}^d :

$$\mathcal{Q}_0 = \{q_0(\xi | \boldsymbol{\lambda}, \mathbf{k}, \boldsymbol{\beta}) = \prod_{j=1}^d q_0^j(\xi_j | \lambda_j, k_j, \beta_j)\} \quad (16)$$

where

$$q_0^j(\xi_j | \lambda_j, k_j, \beta_j) \propto \xi_j^{2k_j \lambda_j - 1} \exp(-\beta_j \xi_j^{2k_j}) 1_{[0,1]}(\xi_j)$$

for $\boldsymbol{\lambda} \in \mathbb{R}_{>0}^d$, $\mathbf{k} \in \mathbb{R}_{>0}^d$, $\boldsymbol{\beta} \in (0, \infty)^d$. **Henceforth, let $g := g_\alpha$ where α is such that M_α is an essential chart.** In other words, we are fixing a resolution map g , working in a fixed essential chart domain, and a coordinate ξ on that domain that makes $K(g(\xi))$ a monomial as a function from $\mathbb{R}^d \rightarrow \mathbb{R}^d$. The idealized variational family of Bhattacharya et al. [2020] is given as the pushforward of base distributions $q_0 \in \mathcal{Q}_0$ by said map g :

$$\mathcal{Q} = \{g\#q_0 : q_0 \in \mathcal{Q}_0\}. \quad (17)$$

We refer to this as an *idealized* variational family for the simple fact that the resolution map g , though its existence is guaranteed, is almost never tractable except in the simplest model-truth-prior triplets. Also note that although the family \mathcal{Q}_0 is mean-field, (17) is *not*.

To study the variational approximation gap incurred by the idealized family (17), we will first introduce some definitions to help us rewrite the gap \mathcal{G} in notation that is consistent with Bhattacharya et al. [2020]. Define

$$\Psi_n(q_0) = -\mathbb{E}_{q_0} n K_n(g(\xi)) - \text{KL}(q_0(\xi) \parallel \varphi(g(\xi)) | g'(\xi)) \quad (18)$$

See Appendix C for the derivation that the variational approximation gap in (14) is equivalent to

$$\mathcal{G} = \log \bar{Z}(n) - \sup_{q_0 \in \mathcal{Q}_0} \Psi_n(q_0). \quad (19)$$

Following Bhattacharya et al. [2020], we consider the deterministic approximation gap corresponding to (19). This is accomplished by replacing K_n with K , leading to

$$\Psi(q_0) := -\mathbb{E}_{q_0} n K(g(\xi)) - \text{KL}(q_0(\xi) \parallel \varphi(g(\xi)) | g'(\xi)). \quad (20)$$

and

$$\bar{Z}_K(n) := \int_W e^{-nK(w)} \varphi(w) dw.$$

For our theoretical investigation, we shall concern ourselves with the *deterministic* variational approximation gap,

$$\mathcal{G}_K := \log \bar{Z}_K(n) - \sup_{q_0 \in \mathcal{Q}_0} \Psi(q_0). \quad (21)$$

Techniques for generalizing the main result Theorem 6.1 which concerns \mathcal{G}_K to the stochastic world can be found in Plummer [2021, Section 5.3.3].

We will appeal to large- n asymptotics to study the behavior of (21). Note that the study and deployment of BNNs is no stranger to large- n asymptotics, both in early MacKay [1992] and recent Ritter et al. [2018] works. We proceed under this tradition, but deviate from the crude (and incorrect) Laplace approximation that is often employed and instead use the correct asymptotics provided by SLT.

6.1 Model evidence in singular models

To study the gap in (21), we begin by examining the asymptotic behavior of $\bar{Z}_K(n)$. When the model is regular, we need not bother with SLT and may find to leading order, $\bar{Z}_K(n) = \varphi(w_0) \sqrt{\frac{(2\pi)^d}{\det H(w_0)}} n^{-d/2}$ via the Laplace approximation. This approximation, however, is egregiously inappropriate for strictly singular models, in particular neural networks Wei et al. [2022]. Nonetheless, perhaps due to a sense that no tractable alternatives exist, the Laplace approximation is seeing a resurgence of application in Bayesian deep learning Ritter et al. [2018], Immer et al. [2021].

For strictly singular models, the quantities $Z(n)$, $\bar{Z}(n)$ and $\bar{Z}_K(n)$ manifest as singular integrals, i.e., integrals of the form $\int_W e^{-nf(w)} \varphi(w) dw$ where $W \subset \mathbb{R}^d$ is a compact semi-analytic subset, and f and φ are real analytic functions. The behavior of a singular integral depends critically on the zeros of f . According to Theorem 6.7 in Watanabe [2009], we find to leading order:

$$\bar{Z}_K(n) = C(p, p_0, \varphi) n^{-\lambda(p, p_0, \varphi)} (\log n)^{m(p, p_0, \varphi) - 1}, \quad (22)$$

where $C(p, p_0, \varphi)$ is a constant independent of n that we shall call the **leading coefficient** following the terminology of Lin [2011]. Note that since $\lambda(p, p_0, \varphi) = d/2$ and $m(p, p_0, \varphi) = 1$ in regular models, (22) is a true generalization of the Laplace approximation, holding for both regular and strictly singular models.

6.2 Bounding \mathcal{G}_K

We show in Lemma D.2 in Appendix D, that for large n , the following bound holds

$$\sup_{q_0 \in \mathcal{Q}_0} \Psi(q_0) \geq -\lambda(p, p_0, \varphi) \log n + C \quad (23)$$

where C is the constant free of n in Lemma D.2. This result is in the same spirit as [Bhattacharya et al., 2020, Theorem 3.1], except that we have improved on the tightness of their lower bound, which in turn allows us to devise better initialization of the variational parameters. With Lemma D.2, we are now in a position to characterize the (deterministic) variational approximation gap, \mathcal{G}_K .

Theorem 6.1 (Deterministic variational approximation gap). *Suppose the model-truth-prior triplet (p, p_0, φ) is such that Theorem 2.1 holds. Let $g = g_\alpha$ where α is such that M_α is an essential chart. On this essential chart, write the local RLCTs $\tilde{\lambda}_j = \frac{\tilde{h}_j + 1}{2k_j}$, $j = 1, \dots, d$ in descending order so that $\tilde{\lambda}_1$ is the RLCT of the triplet (p, p_0, φ) , i.e., $\tilde{\lambda}_1 = \lambda(p, p_0, \varphi)$. If the multiplicity of the triplet is 1, we have, for n large, $\mathcal{G}_K \leq \log C(p, p_0, \varphi) - C + o(1)$, where the constant C is as given in Lemma D.2.*

All that is needed for the proof of Theorem 6.1 is to put together the lower bound in Lemma D.2 with the fact that $\bar{Z}_K(n)$ admits the asymptotic expansion in (22). Even when $m(p, p_0, \varphi) \neq 1$, there may be finite n situations when the two terms $(m(p, p_0, \varphi) - 1) \log \log n$ and $\log C(p, p_0, \varphi) - C$ are comparable. In such settings, the idealized variational family \mathcal{Q} in (17) could still perform well.

6.3 Learning to desingularize

In the preceding section, we studied the deterministic variational approximation gap of an idealized variational family. Although Hironaka proved the existence of a resolution map and showed that it can be found by recursive blow up, known algorithms for finding such resolutions, other than a few exceptional cases (such as those for toric resolutions),

have complexity that vastly exceed existing computational capabilities. Thus we are precluded from directly applying the idealized variational family.

This leads us to consider *learning* the resolution map g using an invertible architecture G_θ resulting in the variational family

$$\hat{\mathcal{Q}} = \{G_\theta \# q_0(\boldsymbol{\lambda}, \mathbf{k}, \boldsymbol{\beta}) : \boldsymbol{\beta} = (n, \beta_2, \dots, \beta_d)\}. \quad (24)$$

If the network is expressive enough, we can hope that $g \in \{G_\theta : \theta\}$, which would lead $\hat{\mathcal{Q}}$ to enjoy the theoretical guarantee provided in Theorem 6.1. Note in (24) the first coordinate of $\boldsymbol{\beta}$ has been set to the sample size n . The proof of Lemma D.2 reveals why we do so. Specifically, it is shown that the following parameters in q_0 can achieve $\Psi(q_0) = -\lambda(p, p_0, \varphi) \log n + C$:

$$\lambda_1 = \lambda(p, p_0, \varphi), \quad k_1 = \tilde{k}_1, \quad \beta_1 = n$$

where \tilde{k}_1 is as in Theorem 6.1. Note that $\lambda(p, p_0, \varphi)$ and \tilde{k}_1 are unknown, but n is certainly known.

It might be readily apparent at this point that we have in $\hat{\mathcal{Q}}$ a standard normalizing flow, albeit with the base distribution given by the generalized gamma distribution. To ease the computational cost, we fix the variational parameters $\boldsymbol{\lambda}, \mathbf{k}, \boldsymbol{\beta}_{[-1]}$ and absorb the learning of their optimal values into the invertible transformation G_θ . Note that this is in line with standard practice, whereby normalizing flows adopt parameter-less base distributions.

To summarize, recognizing that the variational approximation gap can be theoretically studied using SLT allowed for the design of a principled variational family which incurs a variational approximation gap that is independent of sample size n , to leading order. To the best of our knowledge, no existing works on normalizing flows for BNNs theoretically address the variational approximation gap. Furthermore, our results offer a new perspective on the benefits of using normalizing flows for variational inference in BNNs.

7 Experiments

In the following set of experiments⁴, we will isolate and examine the effect of the base distribution. Specifically, we compare the *generalized gamma base distribution* to the commonly-adopted *Gaussian base distribution*, holding the architecture of G_θ fixed when we do so. At the outset, we expect that when G_θ is expressive enough, the effect of the base distribution will be small. However, when G_θ is more limited (and thus less computationally expensive), we conjecture the generalized gamma base distribution can “pick up the slack” and outperform the Gaussian base distribution.

Table 1: The various model-truth-prior triplets considered in experiments. The truth is realizable. The prior over network weights is standard Gaussian. The RLCT is only known in some of the cases.

model	H	\dim_w	$\lambda(p, p_0, \varphi)$	\dim_x	\dim_y
ffrelu	3	42	-	13	1
	7	98	-	13	1
	16	224	-	13	1
	40	560	-	13	1
reducedrank	2	14	5.0	5	2
	7	119	35.0	10	7
	10	230	65.0	13	10
	16	560	152.0	19	16
tanh	15	30	-	1	1
	50	100	-	1	1
	115	230	-	1	1
	280	560	-	1	1
tanh (zero mean)	15	30	1.93	1	1
	50	100	3.53	1	1
	115	230	5.36	1	1
	280	560	8.36	1	1

In line with our earlier discussion, the parameters of the base distributions are frozen throughout training, see Appendix E for the initialization used. The invertible network G_θ is implemented as a sequence of affine coupling transformations. We denote by `base_numcouplingpairs_numhidden` the variational family that results from pushing forward the base distribution through G_θ with the said configuration, see Appendix E for a complete description of the implementation. We consider a total of four different expressivity levels of G_θ from least to most: `2_4`, `2_16`, `4_4`, `4_16`.

The expression for the ELBO objective corresponding to each of the base distributions is given in (29) and (30) in Appendix E. Details of the training procedure such as epochs, learning rate, and optimizer are also given there. Let \hat{q}^*

⁴The code to reproduce our results is available at https://github.com/suswei/BNN_via_SLT.

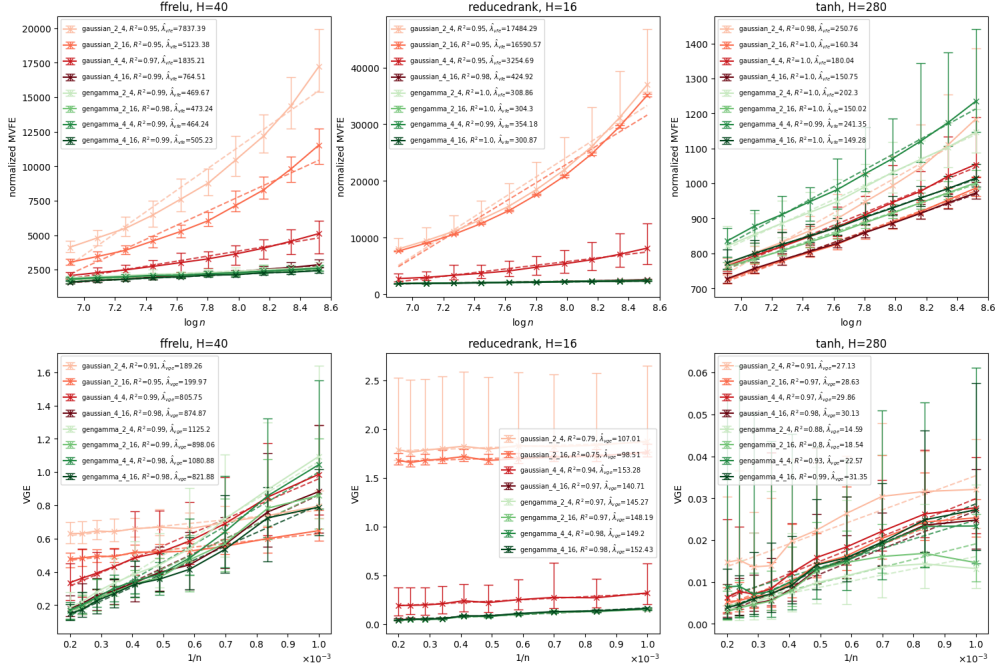


Figure 3: MVFE versus $\log n$ is displayed in the first column and VGE versus $1/n$ is displayed in the second. Each row corresponds to a different model-truth-prior triplet. Line color indicates the expressiveness of the network G_θ , darker being more expressive. Error bars represent mean, min and max over 30 draws of the training set \mathcal{D}_n . The dashed line is the least squares fit with λ_{vfe} and λ_{vge} coefficients and their R^2 values displayed in legend.

be the variational distribution obtained at the end of training. Comparison of the base distributions, and hence the two different normalizing flows, will be made according to normalized MVFE, $\bar{F}_{vb}^*(n)$, and VGE, $G_n(p_{vb}(y|x, \mathcal{D}_n))$. (For both, the lower, the better.) We will also estimate the coefficients λ_{vfe} in (13) and λ_{vge} in (15), see Appendix E.

We consider four model-truth-prior triplets, summarized in Table 1, in which the truth is always realizable. In all four triplets, the prior over the neural network weights is chosen to be the standard Gaussian following conventional practice in BNNs Neal [1996], Bishop [2006]. Note, priors for BNNs are notoriously difficult to design and is an area under active research Sun et al. [2019], Nalisnick et al. [2021].

7.1 Results

Due to space constraints, we only show a subset of the results in Figure 3; complete results can be found in Appendix E. In the first column of Figure 3, we plot $\log n$ versus the normalized MVFE. First, we observe that when G_θ is not very expressive, the generalized gamma resoundingly outperforms the Gaussian base distribution for the reduced rank and ReLU experiments across all values of H in terms of achieving lower MVFE. (This can be better seen in Figure 8 in Appendix E.) On the other hand, as conjectured, when G_θ is most expressive at the 4_16 configuration, the distinction in MVFE between the base distributions is still discernible but less dramatic, see Figure 10. Interestingly, for the tanh triplet, the Gaussian base distribution sometimes achieves lower MVFE depending on the configuration of G_θ .

In the second column of Figure 3, we plot $1/n$ versus the VGE. The results empirically verify the issues we highlighted in Section 4. In terms of VGE, the generalized gamma is not uniformly better than the Gaussian base distribution for the ReLU experiment, contrary to what the corresponding MVFE plots suggest. Only for the reduced rank experiment do we see one-to-one correspondence between MVFE and VGE. Note that the VGE fit is particularly poor for the Gaussian 2_4 and 2_16 configurations because these variational approximations are themselves poor. Next, note the scenario in Figure 2b is borne out by some of the tanh experiments. Take for instance tanh at $H = 115$ for the 2_4 configuration. Judging by MVFE alone the generalized gamma base is worse than Gaussian base, but the corresponding VGE curves show the opposite, see (3,3) subplot in Figures 8 and 9.

8 Discussion

We conclude by discussing some of the limitations of the current work. On the empirical front, the reader may have noticed that our experiments did not involve truly deep BNNs. Strictly speaking this is not a limitation of the proposed method but rather a limitation of the scalability of normalizing flows for approximating deep BNNs. We expect the proposed methodology to benefit from orthogonal research advances in normalizing flow architectures.

On the theoretical side, it may be of interest to flush out the magnitude of $\log C(p, p_0, \varphi) - C$ in Theorem 6.1. The general expression for $C(p, p_0, \varphi)$, although known in special cases [Lin, 2011, Corollary 5.9], has complex dependency on $K(w)$ and the prior. However, we do expect that the leading coefficient can be bounded with some effort. Relatedly, it is important to recognize that Theorem 6.1 only concerns the variational approximation gap of the idealized family in (17). Deriving an analogous result for the Gaussian base distribution would make for interesting future work.

We are optimistic that natural conditions on the model-truth-prior triplet and the variational family should allow for general statements about MVFE asymptotic expansions. Further efforts into studying the asymptotics of the MVFE will also advance knowledge of the relationship between λ_{vfe} and λ_{vge} . In its place, our results here show that it is all the more important to pay attention to the variational approximation gap if we wish to have useful downstream predictions.

Acknowledgements

We thank Daniel Murfet for helpful discussions. SW was supported by the ARC Discovery Early Career Researcher Award (DE200101253). This material is also based on work that is partially funded by an unrestricted gift from Google.

References

- Anirban Bhattacharya, Debdeep Pati, and Sean Plummer. Evidence bounds in singular models: probabilistic and variational perspectives, August 2020. URL <http://arxiv.org/abs/2008.04537>. arXiv: 2008.04537.
- David J C Mackay. Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, January 1995. URL https://doi.org/10.1088/0954-898X_6_3_011.
- Jouko Lampinen and Aki Vehtari. Bayesian approach for neural networks—review and case studies. *Neural Networks*, 14(3):257–274, April 2001. ISSN 0893-6080. doi:10.1016/S0893-6080(00)00098-8. URL <https://www.sciencedirect.com/science/article/pii/S0893608000000988>.
- Hao Wang and Dit-Yan Yeung. A Survey on Bayesian Deep Learning. *ACM Computing Surveys*, 53(5):108:1–108:37, September 2020. ISSN 0360-0300. doi:10.1145/3409383. URL <https://doi.org/10.1145/3409383>.
- Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, USA, 2009.
- Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep Learning Is Singular, and That’s Good. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022. ISSN 2162-2388. doi:10.1109/TNNLS.2022.3167409. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.
- Héctor J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, July 1992. ISSN 0893-6080. doi:10.1016/S0893-6080(05)80037-1. URL <https://www.sciencedirect.com/science/article/pii/S0893608005800371>.
- Sumio Watanabe. On the generalization error by a layered statistical model with Bayesian estimation. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 83(6):95–106, 2000. ISSN 1520-6440. doi:10.1002/(SICI)1520-6440(200006)83:6<95::AID-ECJC11>3.0.CO;2-B. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291520-6440%28200006%2983%3A6%3C95%3A%3AAID-ECJC11%3E3.0.CO%3B2-B>.
- S. Watanabe. Learning efficiency of redundant neural networks in Bayesian estimation. *IEEE Transactions on Neural Networks*, 12(6):1475–1486, November 2001. ISSN 1941-0093. doi:10.1109/72.963783. Conference Name: IEEE Transactions on Neural Networks.
- Kenji Fukumizu. Likelihood ratio of unidentifiable models and multilayer neural networks. *The Annals of Statistics*, 31(3):833–851, June 2003. ISSN 0090-5364. doi:10.1214/aos/1056562464. URL <http://projecteuclid.org/euclid.aos/1056562464>.
- Sumio Watanabe. Almost All Learning Machines are Singular. In *2007 IEEE Symposium on Foundations of Computational Intelligence*, pages 383–388, April 2007. doi:10.1109/FOCI.2007.371500.
- Jonathan Heck. *Well-Calibrated Bayesian Neural Networks*. PhD thesis, University of Cambridge, 2018.
- Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical Deep Learning with Bayesian Principles. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/b53477c2821c1bf0da5d40e57b870d35-Abstract.html>.
- Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in neural information processing systems*, 2019. URL <https://openreview.net/pdf/628ff0351ad95e51cf1aad6af16ae5b7928ec3ea.pdf>.
- M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, 2011. URL <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.441.3813&rep=rep1&type=pdf>.
- Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st international conference on machine learning*, June 2014. URL <http://proceedings.mlr.press/v32/chen14.html>.
- Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning. In *International Conference on Learning Representations*, 2020.
- Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, November 2014. ISSN 0960-3174, 1573-1375. doi:10.1007/s11222-013-9416-2. URL <http://link.springer.com/10.1007/s11222-013-9416-2>.
- Sumio Watanabe. *Mathematical Theory of Bayesian Statistics*. Chapman and Hall/CRC, 1st edition, 2018.

- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, June 2015. URL <http://proceedings.mlr.press/v37/blundell15.html>. ISSN: 1938-7228.
- Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15*, pages 1530–1538, Lille, France, July 2015. JMLR.org.
- Christos Louizos and Max Welling. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. In *International Conference on Machine Learning*, June 2016. URL <http://proceedings.mlr.press/v48/louizos16.html>.
- Christos Louizos and Max Welling. Multiplicative Normalizing Flows for Variational Bayesian Neural Networks. In *International Conference on Machine Learning*, 2017. URL <https://arxiv.org/abs/1703.01961>.
- Jakub Swiatkowski, Kevin Roth, Bastiaan Veeling, Linh Tran, Joshua Dillon, Jasper Snoek, Stephan Mandt, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. The k-tied Normal Distribution: A Compact Parameterization of Gaussian Mean Field Posteriors in Bayesian Neural Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9289–9299. PMLR, November 2020. URL <https://proceedings.mlr.press/v119/swiatkowski20a.html>. ISSN: 2640-3498.
- Shinichi Nakajima and Sumio Watanabe. Variational Bayes Solution of Linear Neural Networks and Its Generalization Performance. *Neural Computation*, 19(4):1112–53, 2007. URL https://www.researchgate.net/publication/6457767_Variational_Bayes_Solution_of_Linear_Neural_Networks_and_Its_Generalization_Performance.
- Masahiro Kohjima and Sumio Watanabe. Phase Transition Structure of Variational Bayesian Nonnegative Matrix Factorization. In Alessandra Lintas, Stefano Rovetta, Paul F.M.J. Verschure, and Alessandro E.P. Villa, editors, *Artificial Neural Networks and Machine Learning – ICANN 2017*, Lecture Notes in Computer Science, pages 146–154, Cham, 2017. Springer International Publishing.
- Naoki Hayashi. Variational approximation error in non-negative matrix factorization. *Neural Networks*, 126:65–75, June 2020. ISSN 0893-6080. doi:10.1016/j.neunet.2020.03.009. URL <https://www.sciencedirect.com/science/article/pii/S0893608020300861>.
- Kazuho Watanabe and Sumio Watanabe. Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation. *The Journal of Machine Learning Research*, 7:625–644, December 2006.
- T. Hosino, K. Watanabe, and S. Watanabe. Stochastic complexity of variational Bayesian hidden Markov models. In *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.*, volume 2, pages 1114–1119 vol. 2, 2005.
- Shinichi Nakajima, Kazuho Watanabe, and Masashi Sugiyama. *Variational Bayesian Learning Theory*. Cambridge University Press, Cambridge, 2019. ISBN 978-1-107-07615-0. doi:10.1017/9781139879354. URL <https://www.cambridge.org/core/books/variational-bayesian-learning-theory/0F6AABA050630E01E1B6EDA5E2CAFA05>.
- Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but Did It Work?: Evaluating Variational Inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5581–5590. PMLR, July 2018. URL <https://proceedings.mlr.press/v80/yao18a.html>. ISSN: 2640-3498.
- Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated Variational Inference via Practical Posterior Error Bounds. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR, June 2020. URL <https://proceedings.mlr.press/v108/huggins20a.html>. ISSN: 2640-3498.
- Sameer K. Deshpande, Soumya Ghosh, Tin D. Nguyen, and Tamara Broderick. Are you using test log-likelihood correctly? In *36th Conference on Neural Information Processing Systems*, November 2022.
- Akash Kumar Dhaka, Alejandro Catalina, Michael R Andersen, Må ns Magnusson, Jonathan Huggins, and Aki Vehtari. Robust, Accurate Stochastic Optimization for Variational Inference. In *Advances in Neural Information Processing Systems*, volume 33, pages 10961–10973. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/7cac11e2f46ed46c339ec3d569853759-Abstract.html>.
- J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez. Quality of Uncertainty Quantification for Bayesian Neural Network Inference. In *Proceedings at the International Conference on Machine Learning: Workshop on Uncertainty & Robustness in Deep Learning (ICML)*, 2019.

- Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 18237–18248. Curran Associates, Inc., 2020. URL <https://papers.nips.cc/paper/2020/hash/d3d9446802a44259755d38e6d163e820-Abstract.html>.
- Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the Expressiveness of Approximate Inference in Bayesian Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 15897–15908. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/b6dfd41875bc090bd31d0b1740eb5b1b-Abstract.html>.
- Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in neural information processing systems*, volume 24, pages 2348–2356. Curran Associates, Inc., 2011. URL <https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf>.
- Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, and Richard Turner. Black-box alpha divergence minimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd international conference on machine learning*, volume 48 of *Proceedings of machine learning research*, pages 1511–1520, New York, New York, USA, June 2016. PMLR. URL <http://proceedings.mlr.press/v48/hernandez-lobatob16.html>.
- Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in neural information processing systems*, volume 29, pages 1073–1081. Curran Associates, Inc., 2016. URL <https://proceedings.neurips.cc/paper/2016/file/7750ca3559e5b8e1f44210283368fc16-Paper.pdf>.
- Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *International conference on machine learning*, pages 2611–2620, 2018. tex.organization: PMLR.
- Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional Variational Bayesian Neural Networks. In *International Conference on Learning Representations*, 2019.
- David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3): 448–472, May 1992. URL <https://resolver.caltech.edu/CaltechAUTHORS:MACnc92b>.
- Beau Coker, Wessel P. Bruinsma, David R. Burt, Weiwei Pan, and Finale Doshi-Velez. Wide Mean-Field Bayesian Neural Networks Ignore the Data. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 5276–5333. PMLR, May 2022. URL <https://proceedings.mlr.press/v151/coker22a.html>. ISSN: 2640-3498.
- Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy Natural Gradient as Variational Inference. In *International Conference on Machine Learning*, pages 5852–5861. PMLR, July 2018. URL <http://proceedings.mlr.press/v80/zhang18l.html>. ISSN: 2640-3498.
- George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):1–64, 2021. ISSN 1532-4435.
- David A Moore. Symmetrized Variational Inference. In *NIPS Workshop on Advances in Approximate Bayesian Inference*, volume 4, page 8, 2016.
- Arya A Pourzanjani, Richard M Jiang, and Linda R Petzold. Improving the Identifiability of Neural Networks for Bayesian Inference. In *NIPS Workshop on Bayesian Deep Learning*, volume 4, page 5, 2017.
- Richard Kurle, Ralf Herbrich, Tim Januschowski, Yuyang Wang, and Jan Gasthaus. On the detrimental effect of invariances in the likelihood for variational inference. In *NeurIPS*, October 2022.
- Sean Plummer. *Statistical and Computational Properties of Variational Inference*. PhD thesis, Texas A&M University, 2021.
- Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th international conference on learning representations, ICLR 2018-Conference track proceedings*, 2018. URL <https://openreview.net/pdf?id=Skdvd2xAZ>.
- Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable Marginal Likelihood Estimation for Model Selection in Deep Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4563–4573. PMLR, July 2021. URL <https://proceedings.mlr.press/v139/immer21a.html>. ISSN: 2640-3498.

- Shaowei Lin. *Algebraic Methods for Evaluating Integrals in Bayesian Statistics*. PhD thesis, University of California Berkeley, 2011.
- Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1996. URL <http://link.springer.com/10.1007/978-1-4612-0745-0>.
- Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.
- Eric Nalisnick, Jonathan Gordon, and Jose Miguel Hernandez-Lobato. Predictive Complexity Priors. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 694–702. PMLR, March 2021. URL <https://proceedings.mlr.press/v130/nalisnick21a.html>. ISSN: 2640-3498.
- Sumio Watanabe. Asymptotic Learning Curve and Renormalizable Condition in Statistical Learning Theory. *Journal of Physics: Conference Series*, 233:012014, June 2010. ISSN 1742-6596. doi:10.1088/1742-6596/233/1/012014.
- Shuya Nagayasu and Sumio Watanabe. Asymptotic behavior of free energy when optimal probability distribution is not unique. *Neurocomputing*, 500:528–536, August 2022. ISSN 0925-2312. doi:10.1016/j.neucom.2022.05.071. URL <https://www.sciencedirect.com/science/article/pii/S092523122200652X>.
- Miki Aoyagi and Sumio Watanabe. Resolution of Singularities and the Generalization Error with Bayesian Estimation for Layered Neural Network. *IEICE Trans*, pages 2112–2124, 2006.
- Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18(7):924–933, September 2005. URL <https://linkinghub.elsevier.com/retrieve/pii/S0893608005000559>.

A SLT assumptions

Conventional learning theory studies parametric statistical models under the assumption that they satisfy certain regularity conditions. Unfortunately, most models employed in modern machine learning lack such regularity and exhibit behavior that is unaccounted for by conventional learning theory. The core observation of singular learning theory is that **singularities** of unidentifiable models have drastic impact on learning behavior. In Watanabe [2009] and Watanabe [2018], Watanabe carried out a rigorous investigation into singular statistical models from the Bayesian perspective, culminating in several cornerstone results including Theorem 2.1 and those described in Section 3.

Fundamental Conditions I and II given in Definitions 6.1 and 6.3 of Watanabe [2009], respectively, are a set of blanket conditions that Watanabe uses throughout the development of SLT; some components of these conditions are not actually relevant to the results we cite in this paper. Below, we simply present the parts of Fundamental Conditions I and II that are relevant to the SLT results we care about in this paper.

1. The model has compact parameter space $W \subset \mathbb{R}^d$ defined by real analytic inequalities.
2. The parameter space W is equipped with a prior distribution with semi-analytic density, i.e. the prior density can be expressed as $\varphi(w) = \varphi_0(w)\varphi_1(w)$ with φ_0 a positive smooth function and φ_1 a non-negative analytic function.
3. For all $w \in W$, $p(x|w)$ has the same support as the truth $p_0(x)$ ⁵
4. The true distribution $p_0(x)$ is realisable by the model $p(x|w)$. In other words, there exist a parameter $w_0 \in W$, such that $p_0(x) = p(x|w_0)$.
5. The log-likelihood ratio function $f(x, w) := \log \frac{p_0(x)}{p(x|w)}$ can be extended to a complex analytic function $W_{\mathbb{C}} \ni w \mapsto f(\cdot, w)$, taking value in the $L^s(p_0)$ with $s = 2$, i.e., the space of functions that are square integrable with respect to the true measure p_0 .

On compactness We require the parameter space W to be compact in Assumption 1. This is not required when the set of true parameters $W_0 = \{w : K(w) = 0\}$ is contained within a relatively compact neighborhood as contributions of parameters far from W_0 drops of exponentially. Even in the case where W_0 is not compact, we could consider compactification of $\overline{\mathbb{R}}^d \simeq \mathbb{R}^d \cup \{|w| = \infty\}$, but we will need to ensure that $f(x, w)$ extends to an analytic function in the neighborhood of infinity. In practical implementation however, it is common to have compact W due to machine implementation constraints.

On realizability Assumption 4 above required that the zero set of $K(w)$ be non-empty. Let's discuss how to deal with violations of this assumption. In unrealisable cases, we can still derive many SLT results by replacing $K(w)$ with $K(w) - K(w_0)$ where w_0 is any parameter that achieves the minimum of K and replacing $f(x, w)$ in Assumption 5 with $f(x, w) = \log \frac{p(x|w_0)}{p(x|w)}$. Then we can smoothly proceed with the theory in the usual manner by resolving singularities of $K(w) - K(w_0)$ in a neighbourhood of the optimal parameter set $W_0 = \{w : K(w) - K(w_0) = 0\}$, if we make an additional assumption known as the **renormalisability** condition [Watanabe, 2010]. Without renormalisability, we can still proceed but with considerably more difficult technical challenges Watanabe [2010], Nagayasu and Watanabe [2022].

On analyticity and integrability conditions The result in Theorem 2.1 is obtained through a direct application of Hironaka's resolution of singularities, simultaneously, to $K(w) = \int p_0(x)f(x, w)dx$ and the prior $\varphi(w)$. It only requires that the zero set of $K(w)$ is non-empty and both functions are analytic on an open neighbourhood of the zero set. The requirements can be further relaxed to have $K(w)$ and $\varphi(w)$ being semi-analytic and the resolution theorem applied to their analytic factors. The analyticity condition on $f(x, w)$ in Assumption 5 is usually sufficient to ensure analyticity on $K(w)$. Application of a resolution map $g(\xi) = w$ for $K(w)$, together with integrability conditions for $f(x, w)$ (Assumption 5) results in the discovery of the connection between geometry W_0 with free energy asymptotics 11 and 10 via the RLCT.

It should be noted, however, that even in cases where $f(x, w)$ is non-analytic, the model might still be amenable to the same treatment if an equivalent analytic representation can be found. For instance, [Watanabe, 2009, Section 7.8] shows how the non-analytic $f(x, w)$ for normal mixture models can be analysed in SLT.

⁵In the main text we work with the "supervised" setting and model the joint distribution $p(x, y|w) = p(x)p(y|x, w)$. Here, for easier exposition, we limit the discussion to the "unsupervised" setting $p(x|w)$.

B Toy example of RLCT calculation

We recall Example 27 from Watanabe [2018] to illustrate the concepts of resolution map, RLCT and multiplicity for a simple model-truth-prior triplet. For univariate input $x \in [0, 1]$ and univariate output $y \in \mathbb{R}$, consider the model with parameter $w = (a, b) \in [0, 1]^2$ given by

$$p(x, y|w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - a \tanh(bx))^2\right). \quad (25)$$

Suppose the prior is uniform, i.e., $\varphi(w) = 1$ and the truth is given by $p_0(x, y) = p(x, y|0, 0)$. Then we can easily see that

$$K(w) = b^2 a^2 \frac{1}{2} K_0(w),$$

where

$$K_0(w) = \int_0^1 \left(\frac{\tanh(bx)}{b}\right)^2 dx.$$

The following desingularization map puts the triplet in standard form:

$$\begin{aligned} \xi_1 &= \sqrt{\frac{K_0(w)}{2}} a \\ \xi_2 &= b. \end{aligned}$$

Next, we have $\varphi(g(\xi)) = \xi^h$ where $h = (0, 0)$ and $b(\xi) = |g'(\xi)|$. Since $(k_1, k_2) = (1, 1)$ and $(h_1, h_2) = (0, 0)$ we have $(\lambda_1, \lambda_2) = (1/2, 1/2)$. Therefore for this particular model-truth-prior triplet, the RLCT is $1/2$ with multiplicity 2.

We should note that, to date, there is a rather small collection of strictly singular model-truth-prior triplets where the RLCT and multiplicity are known.

C Rewriting the variational approximation gap

Recall the posterior distribution in the new coordinate ξ in (7). For q in (17), we have

$$\begin{aligned} & \text{KL}(q(w) \parallel p(w|\mathcal{D}_n)) \\ &= \text{KL}(q_0(\xi) \parallel p(\xi|\mathcal{D}_n)) \\ &= \mathbb{E}_{q_0} n K_n(g(\xi)) + \text{KL}(q_0(\xi) \parallel \varphi(g(\xi)) |g'(\xi)|) + \log \bar{Z}(n). \end{aligned}$$

Following the notation in Bhattacharya et al. [2020], we defined

$$\Psi_n(q_0) = -\mathbb{E}_{q_0} n K_n(g(\xi)) - \text{KL}(q_0(\xi) \parallel \varphi(g(\xi)) |g'(\xi)|).$$

As long as the support of $q_0(\xi)$ is contained in the support of the posterior $p(\xi|\mathcal{D}_n)$, we have $\text{KL}(q_0(\xi) \parallel p(\xi|\mathcal{D}_n)) \geq 0$, leading to the lower bound

$$\Psi_n(q_0) \leq \log \bar{Z}(n).$$

Equality is achieved if and only if $q_0(\xi) = p(\xi|\mathcal{D}_n)$.

D Lemmas and proofs

Lemma D.1. *Suppose the model-truth-prior triplet (p, p_0, φ) is such that Theorem 2.1 holds. Let $g = g_\alpha$ where α is such that M_α is an essential chart. On this essential chart, write the local RLCTs*

$$\tilde{\lambda}_j = \frac{\tilde{h}_j + 1}{2\tilde{k}_j}, j = 1, \dots, d$$

in descending order so that $\tilde{\lambda}_1$ is the RLCT of the triplet (p, p_0, φ) , i.e., $\tilde{\lambda}_1 = \lambda(p, p_0, \varphi)$. Let \mathcal{Q}_0 be as in (16). For $q_0 \in \mathcal{Q}_0$, we have

$$\Psi(q_0) = -E_1 - E_2 + E_3 + E_4$$

with the individual terms E_1, \dots, E_4 given below in the body of the proof.

Proof. With standard form and Main Formula 1 in Watanabe [2009], we have

$$\begin{aligned} nK(g(\xi)) &= n\xi^{2\tilde{k}} \\ \varphi(g(\xi))|g'(\xi)| &= b(\xi) \left| \xi^{\tilde{h}} \right| \end{aligned}$$

with $b(\xi) > 0$. We therefore have

$$\begin{aligned} \Psi(q_0) &= -n\mathbb{E}_{q_0} \left[\xi^{2\tilde{k}} \right] - \mathbb{E}_{q_0} \log q_0 + \mathbb{E}_{q_0} \left[\log \xi^{\tilde{h}} \right] + \mathbb{E}_{q_0} [\log b(\xi)] \\ &= -E_1 - E_2 + E_3 + E_4 \end{aligned}$$

where we have named each term in the sum

$$\begin{aligned} E_1 &:= n\mathbb{E}_{q_0} \left[\xi^{2\tilde{k}} \right], & E_2 &:= \mathbb{E}_{q_0} \log q_0 \\ E_3 &:= \mathbb{E}_{q_0} \left[\log \xi^{\tilde{h}} \right], & E_4 &:= \mathbb{E}_{q_0} [\log b(\xi)] \end{aligned}$$

In the following we will make use of the following elementary facts about the univariate generalized gamma density truncated to $[0, 1]$. They are stated in the same notation as in Bhattacharya et al. [2020]. The normalizing constant of q_j is given by $B(\lambda_j, k_j, \beta_j)$ where

$$B(\lambda, k, \beta) = \frac{\beta^{-\lambda} \Gamma(\lambda) \gamma(\lambda, \beta)}{2k} \quad (26)$$

and $\gamma(a, x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt$ is the (regularized) lower incomplete gamma function. The quantity $E_{q_j} \xi^{2k_j} = G(\lambda_j, \beta_j)$ where

$$G(\lambda, \beta) = \frac{\lambda \gamma(\lambda + 1, \beta)}{\beta \gamma(\lambda, \beta)}. \quad (27)$$

First we have

$$\begin{aligned} E_1 &= n \prod_{j=1}^d \mathbb{E}_{q_0^j} \xi^{2\tilde{k}_j} \\ &= n \prod_{j=1}^d \beta_j^{-\frac{\tilde{k}_j}{k_j}} \frac{\Gamma(\lambda_j + \frac{\tilde{k}_j}{k_j}) \gamma(\lambda_j + \frac{\tilde{k}_j}{k_j}, \beta_j)}{\Gamma(\lambda_j) \gamma(\lambda_j, \beta_j)}. \end{aligned}$$

Next we have

$$E_2 = \sum_{j=1}^d h_j \mathbb{E}_{q_0^j} \log \xi_j - \beta_j G(\lambda_j, \beta_j) - \log B(k_j, h_j, \beta_j).$$

For the third term we have

$$E_3 = \sum_{j=1}^d \tilde{h}_j \mathbb{E}_{q_0^j} \log \xi_j.$$

□

In the lemma below, we improve upon the lower bound provided in Theorem 3.1 in Bhattacharya et al. [2020] where the constant is given by

$$\tilde{\lambda} \left(1 - \prod_{j=m+1}^d G(\tilde{\lambda}_j, \beta_j) \right) + \sum_{j=m+1}^d [\beta_j G(\tilde{\lambda}_j, \beta_j) + \log B(\tilde{k}_j, \tilde{h}_j, \beta_j)] - \sum_{j=1}^d \log(2\tilde{k}_j) - \sum_{j=1}^d \log(\tilde{\lambda}_j),$$

where $\beta_j = 1$ for $j \geq m + 1$.

Lemma D.2. *Suppose the conditions of Lemma D.1 hold. We have, for n large,*

$$\sup_{q_0 \in \mathcal{Q}_0} \Psi(q_0) \geq -\lambda(p, p_0, \varphi) \log n + C,$$

where

$$C = \sup_{\lambda_{[-1]}, \mathbf{k}_{[-1]}, \beta_{[-1]}} C(\lambda_{[-1]}, \mathbf{k}_{[-1]}, \beta_{[-1]}).$$

Proof. Let q_0 be such that

$$\lambda_1 = \tilde{\lambda}_1, \quad k_1 = \tilde{k}_1, \quad \beta_1 = n.$$

We can use Lemma D.1 to obtain the expression for $\Psi(q_0)$. Next, using the fact that $nG(\lambda, n) \approx \lambda$ and $\log B(k, h, n) \approx -\lambda \log n$ and $b(\xi)$ is bounded below away from zero, $b(\xi) > b_0 := \inf_{\xi} b(\xi) > 0$. We get that for n large,

$$\sup_{q_0} \Psi(q_0) \geq -\lambda(p, p_0, \varphi) \log n + C(\boldsymbol{\lambda}_{[-1]}, \mathbf{k}_{[-1]}, \boldsymbol{\beta}_{[-1]}),$$

where

$$\begin{aligned} C(\boldsymbol{\lambda}_{[-1]}, \mathbf{k}_{[-1]}, \boldsymbol{\beta}_{[-1]}) &= \lambda(p, p_0, \varphi) \left(1 - \prod_{j=2}^d \beta_j^{-\frac{\tilde{k}_j}{k_j}} \frac{\Gamma(\lambda_j + \frac{\tilde{k}_j}{k_j}) \gamma(\lambda_j + \frac{\tilde{k}_j}{k_j}, \beta_j)}{\Gamma(\lambda_j) \gamma(\lambda_j, \beta_j)} \right) \\ &+ \sum_{j=2}^d \left((\tilde{h}_j - h_j) \mathbb{E}_{q_0} \log \xi_j + \beta_j G(\lambda_j, \beta_j) - \log B(k_j, h_j, \beta_j) \right) + \log b_0. \end{aligned} \quad (28)$$

□

E Experiment details

We first provide details on the model-truth-prior triplets considered in Section 7. Next we describe the architecture adopted for G_θ in the implementation of the normalizing flow. We then detail the training procedure for learning the normalizing flow and the estimation of the evaluation measures. Finally, additional experimental results are given and discussed.

E.1 Model-truth-prior triplets

In all triplets considered, the prior over the neural network weights is chosen to be the standard Gaussian.

In the **one-layer** tanh experiment, the input $x \in \mathbb{R}$ follows the uniform distribution on $[-1, 1]$, and the response variable $y \in \mathbb{R}$ is modeled as

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - f(x, w))^2\right),$$

where

$$f(x, w) = \sum_{h=1}^H b_h \tanh(a_h x)$$

is a tanh network with H hidden units and w is the collection of neural network weights $\{(a_h, b_h)\}_{h=1}^H$. We shall consider two true distributions, one in which we know the true RLCT and multiplicity, which we call **one-layer** tanh **zero-mean**, and the other where we do not, which we call simply **one-layer** tanh. For the *zero-mean* setting, we set

$$p_0(y|x) = p(y|x, 0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right).$$

In this case, it was shown in Aoyagi and Watanabe [2006] that

$$\lambda(p, p_0, \varphi) = \frac{H + i^2 + i}{4i + 2}$$

and $m = 2$ if $i^2 = H$, and $m = 1$ if $i^2 < H$ where i is the maximum integer satisfying $i^2 \leq H$. In contrast, were this a regular statistical model, we would have $\lambda(p, p_0, \varphi) = H$. For the other truth setting, we simply take a fixed draw of w_0 from the standard Gaussian. In this case the true RLCT and multiplicity are *unknown*.

In the **reduced rank regression** experiment, the input $x \in \mathbb{R}^M$ is generated from standard Gaussian and the response variable $y \in \mathbb{R}^N$ is modeled as

$$p(y|x, w) = (2\pi)^{-N/2} \exp\left\{-\frac{1}{2}\|y - BAx\|^2\right\},$$

where $\{w = (A, B) | A \in \mathbb{R}^{H \times M}, B \in \mathbb{R}^{N \times H}\}$. This model is readily seen to be a special case of a neural network with hidden units H and identity activation function. We shall set $M = H + 3$ and $N = H$. The true parameters A_0

and B_0 are given as follows. The matrix B_0 is set to be the identity matrix $I_{N \times N}$. The matrix A_0 is set to be an identity matrix with dimension H plus three additional columns of 1: $A_0 = [I_{H \times H}; J_{H \times 3}]$. The rank r for $B_0 A_0$ equals H . Under this condition, $N + H < M + r$ is trivially satisfied and we are in Case iii) of Aoyagi and Watanabe [2005] for which the RLCT was derived in Aoyagi and Watanabe [2005] to be

$$\lambda(p, p_0, \varphi) = (NH - Hr + Mr)/2, m = 1.$$

Note that were this a regular model, we would instead have $\lambda(p, p_0, \varphi) = (MH + NH)/2$. Notably the multiplicity is always either $m = 1$ or $m = 2$ for the reduced rank regression model.

In the **feedforward ReLU** experiment, the input $x \in \mathbb{R}^{13}$ is generated from the standard multivariate Gaussian and the response variable $y \in \mathbb{R}$ is modeled as Gaussian $N(f(x, w), 1)$ where $f(x, w) = w_2 \text{ReLU}(w_1 x)$ for $w_1 \in \mathbb{R}^{H \times 13}$ and $w_2 \in \mathbb{R}^{1 \times H}$. The true distribution $p_0(y|x)$ is fixed at a random draw of w_1, w_2 from the standard Gaussian. The true RLCT and multiplicity are *unknown* for this truth-prior-triplet.

E.2 Normalizing flow

The generalized gamma base distribution q_0 is initialized (and frozen) at

$$\begin{aligned} \lambda_0 &= (1, \dots, 1), \\ \mathbf{k}_0 &= (1, \dots, 1), \\ \beta_0 &= (n, d/2, \dots, d/2). \end{aligned}$$

The Gaussian base distribution is initialized (and frozen) at the standard multivariate Gaussian with mean zero and identity covariance. Only the weights θ in the invertible architecture G_θ are updated.

Next, we detail the implementation of G_θ . With r denoting a binary mask, a so-called affine coupling layer acts as follows for $u, v \in \mathbb{R}^d$,

$$u \mapsto v = (1 - r) \odot u + r \odot (u \odot \exp(s(r \odot u)) + t(r \odot u)),$$

where s and t are scaling and translation networks, respectively. We implement the translation network t as a two-hidden-layer feedforward (leaky) ReLU neural network with tanh output activation function. The scaling s is another two-hidden-layer feedforward (leaky) ReLU neural network with identity output activation function. Note the binary mask r must alternate from one affine coupling layer to the next, for otherwise there would be little expressive power in the resulting network. Note that the specific architecture of G_θ has rendered the log Jacobian term, $\log |G'_\theta(\cdot)|$, computationally tractable. Below is a printout of the network G_θ with 2 alternating coupling pairs and 4 hidden units:

```
(s): ModuleList(
  (0): Sequential(
    (0): Linear(in_features=210, out_features=4, bias=True)
    (1): LeakyReLU(negative_slope=0.01)
    (2): Linear(in_features=4, out_features=4, bias=True)
    (3): LeakyReLU(negative_slope=0.01)
    (4): Linear(in_features=4, out_features=4, bias=True)
    (5): LeakyReLU(negative_slope=0.01)
    (6): Linear(in_features=4, out_features=210, bias=True)
    (7): Tanh()
  )
  (1): Sequential(
    (0): Linear(in_features=210, out_features=4, bias=True)
    (1): LeakyReLU(negative_slope=0.01)
    (2): Linear(in_features=4, out_features=4, bias=True)
    (3): LeakyReLU(negative_slope=0.01)
    (4): Linear(in_features=4, out_features=4, bias=True)
    (5): LeakyReLU(negative_slope=0.01)
    (6): Linear(in_features=4, out_features=210, bias=True)
    (7): Tanh()
  )
  (2): Sequential(
    (0): Linear(in_features=210, out_features=4, bias=True)
    (1): LeakyReLU(negative_slope=0.01)
    (2): Linear(in_features=4, out_features=4, bias=True)
    (3): LeakyReLU(negative_slope=0.01)
    (4): Linear(in_features=4, out_features=4, bias=True)
    (5): LeakyReLU(negative_slope=0.01)
    (6): Linear(in_features=4, out_features=210, bias=True)
    (7): Tanh()
  )
  (3): Sequential(
    (0): Linear(in_features=210, out_features=4, bias=True)
    (1): LeakyReLU(negative_slope=0.01)
    (2): Linear(in_features=4, out_features=4, bias=True)
    (3): LeakyReLU(negative_slope=0.01)
    (4): Linear(in_features=4, out_features=4, bias=True)
    (5): LeakyReLU(negative_slope=0.01)
    (6): Linear(in_features=4, out_features=210, bias=True)
  )
)
```


To estimate coefficients λ_{vfe} and λ_{vge} , we generate 30 realizations of training data \mathcal{D}_n for each of 10 possible sample sizes n evenly spaced on the log scale between 3.0 and 3.7: $n \in \{1000, 1196, 1431, 1711, 2047, 2448, 2929, 3503, 4190, 5012\}$. This allows us to estimate the left-hand sides of (13) and (15). The coefficients themselves are estimated by fitting least squares, against $\log n$ for the average normalized MVFE and $1/n$ for the average VGE. For the former, we fit an intercept, while for the latter the intercept is forced to be zero.

E.5 Additional experimental results

In this section, we display the MVFE and VGE for all four experiments in Table 1. We first group by the individual base distributions, which allows for greater readability as the y-axis scale is consistent within the base distribution. We then juxtapose the “best” performing G_θ , according to MVFE, for each base distribution, which usually happens to be the architecture G_θ with 4 alternating pairs and 16 hidden units. Similarly, we also plot the least expressive G_θ which is the 2_4 configuration.

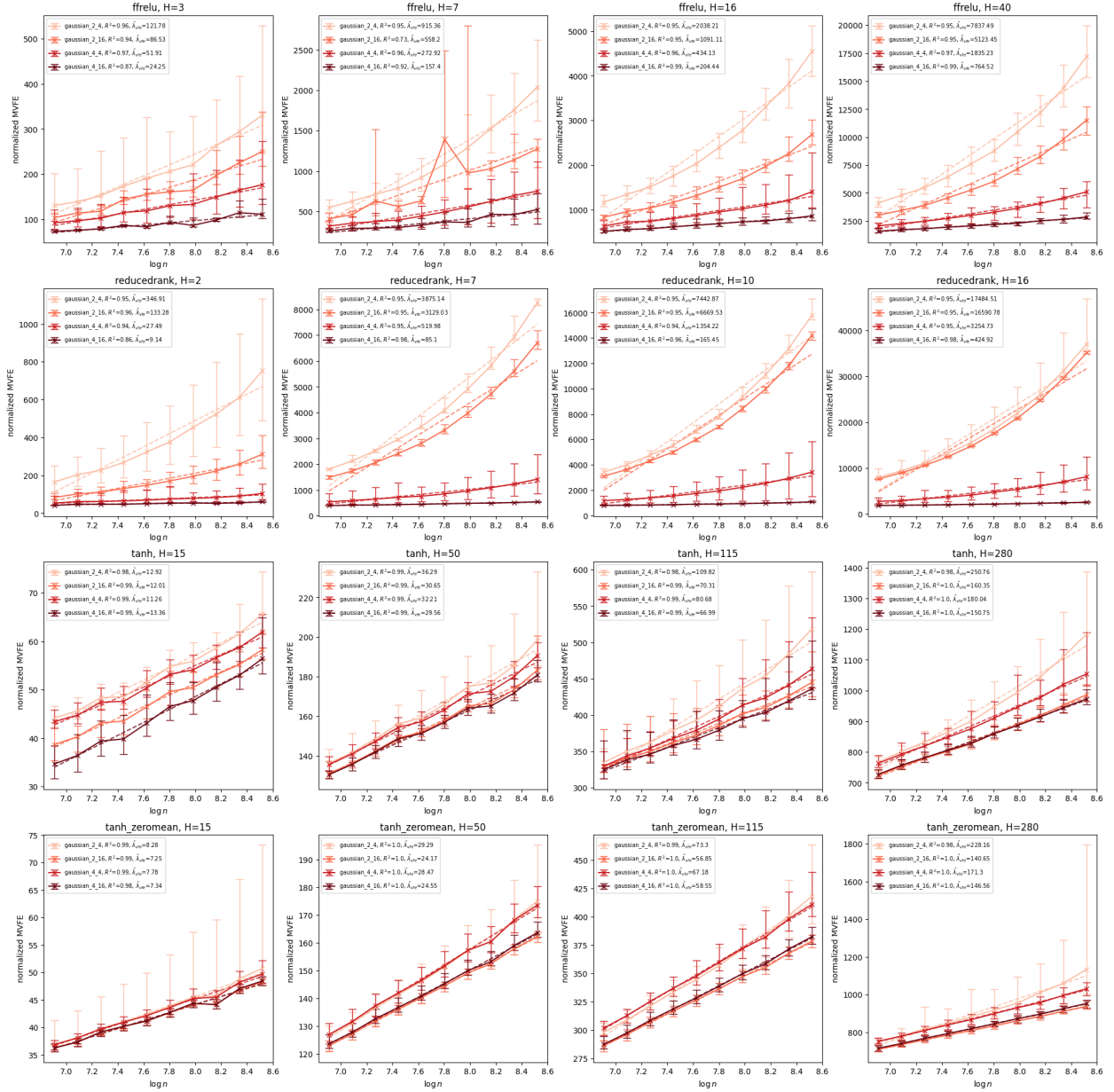


Figure 4: MVFE for Gaussian base distribution.

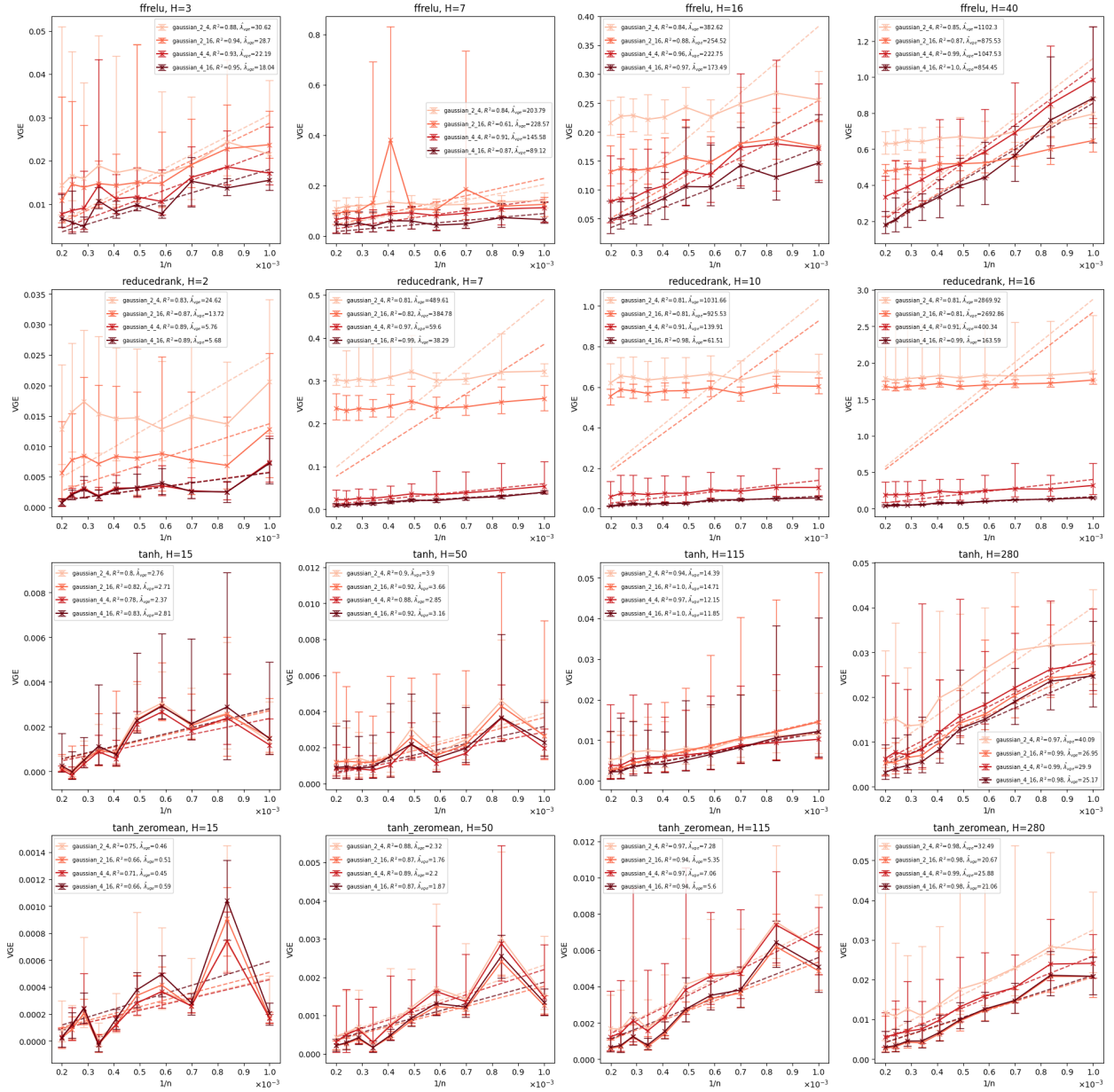


Figure 5: VGE for Gaussian base distribution.

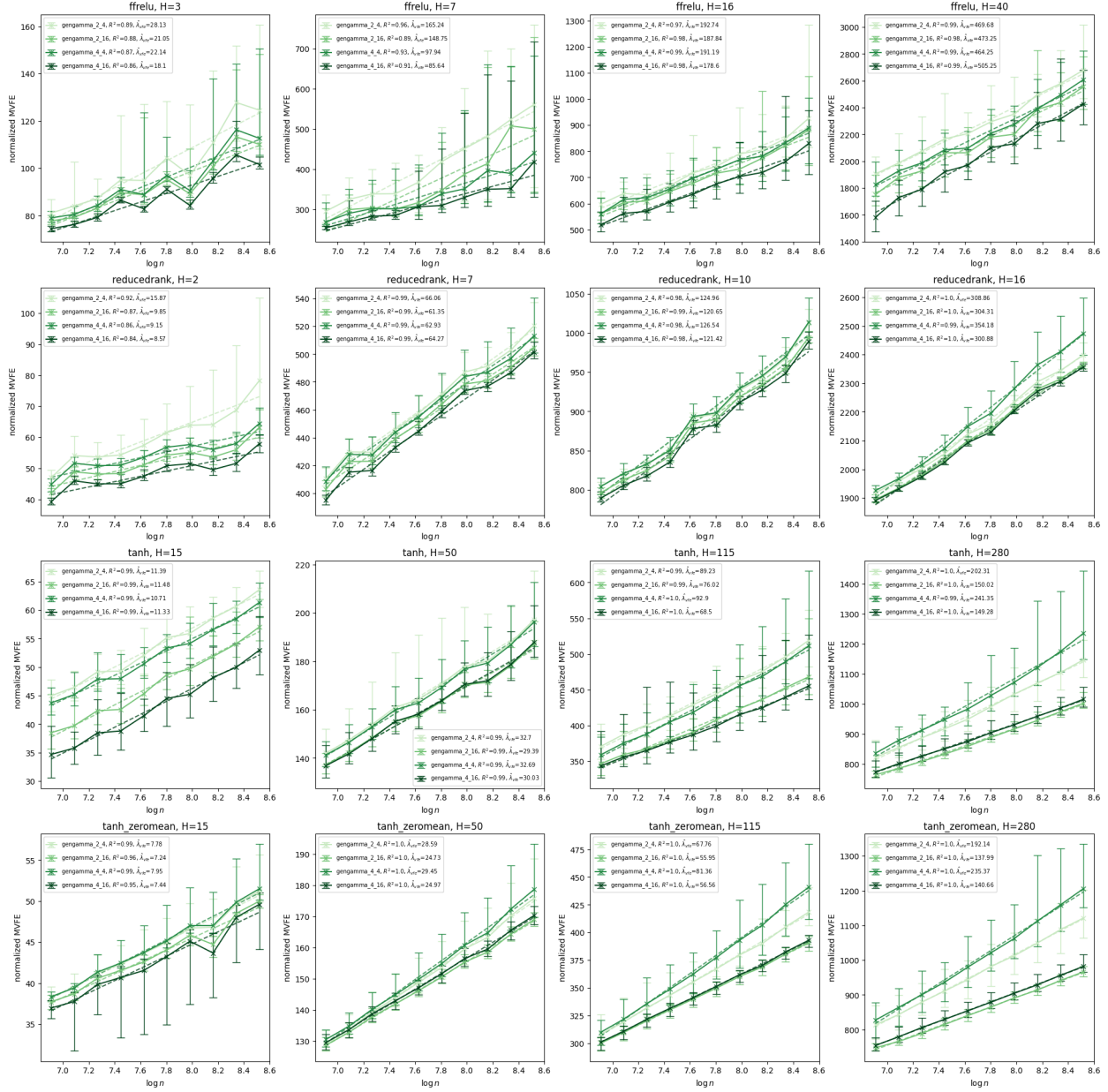


Figure 6: MVFE for generalized gamma base distribution.

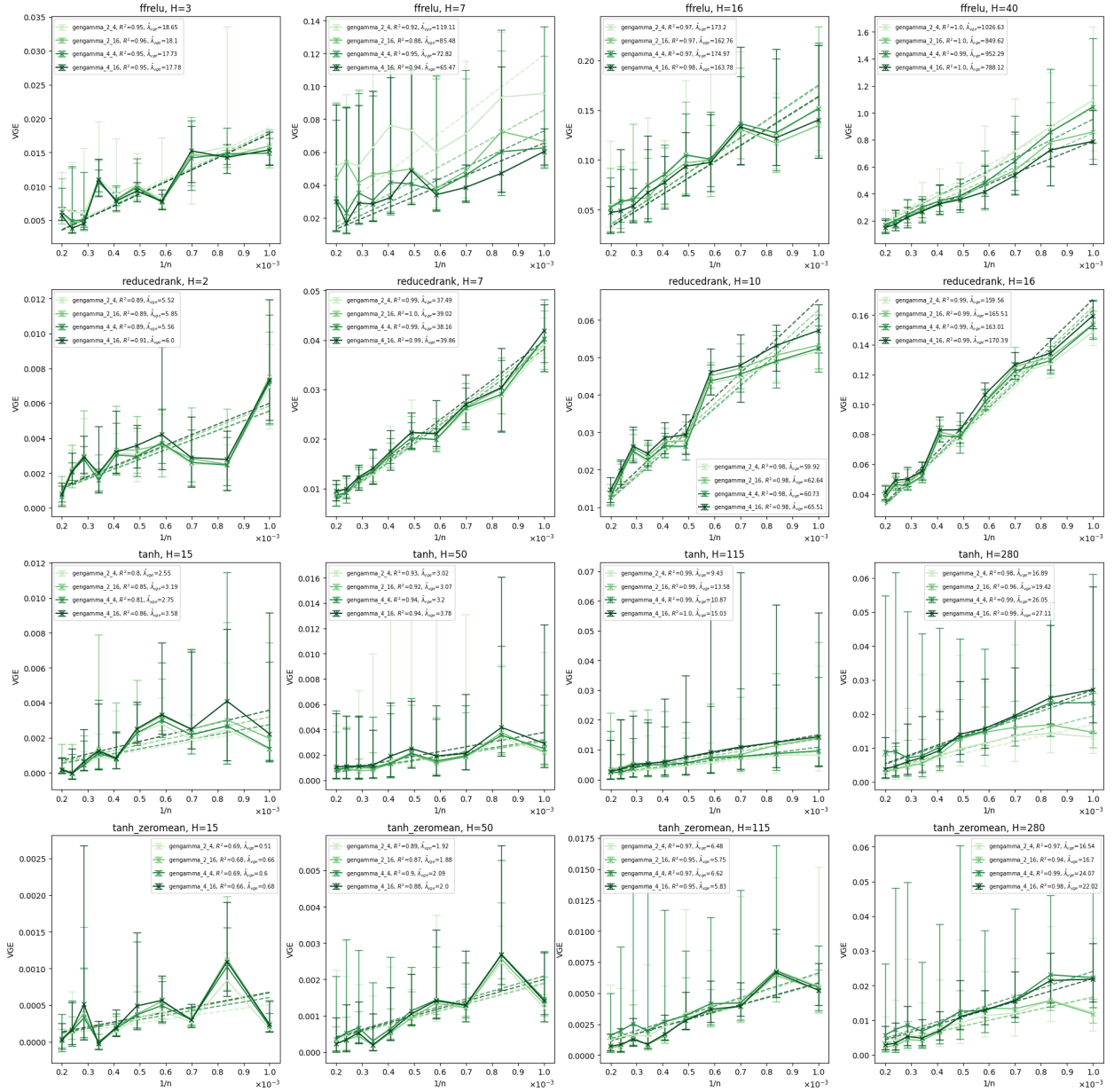
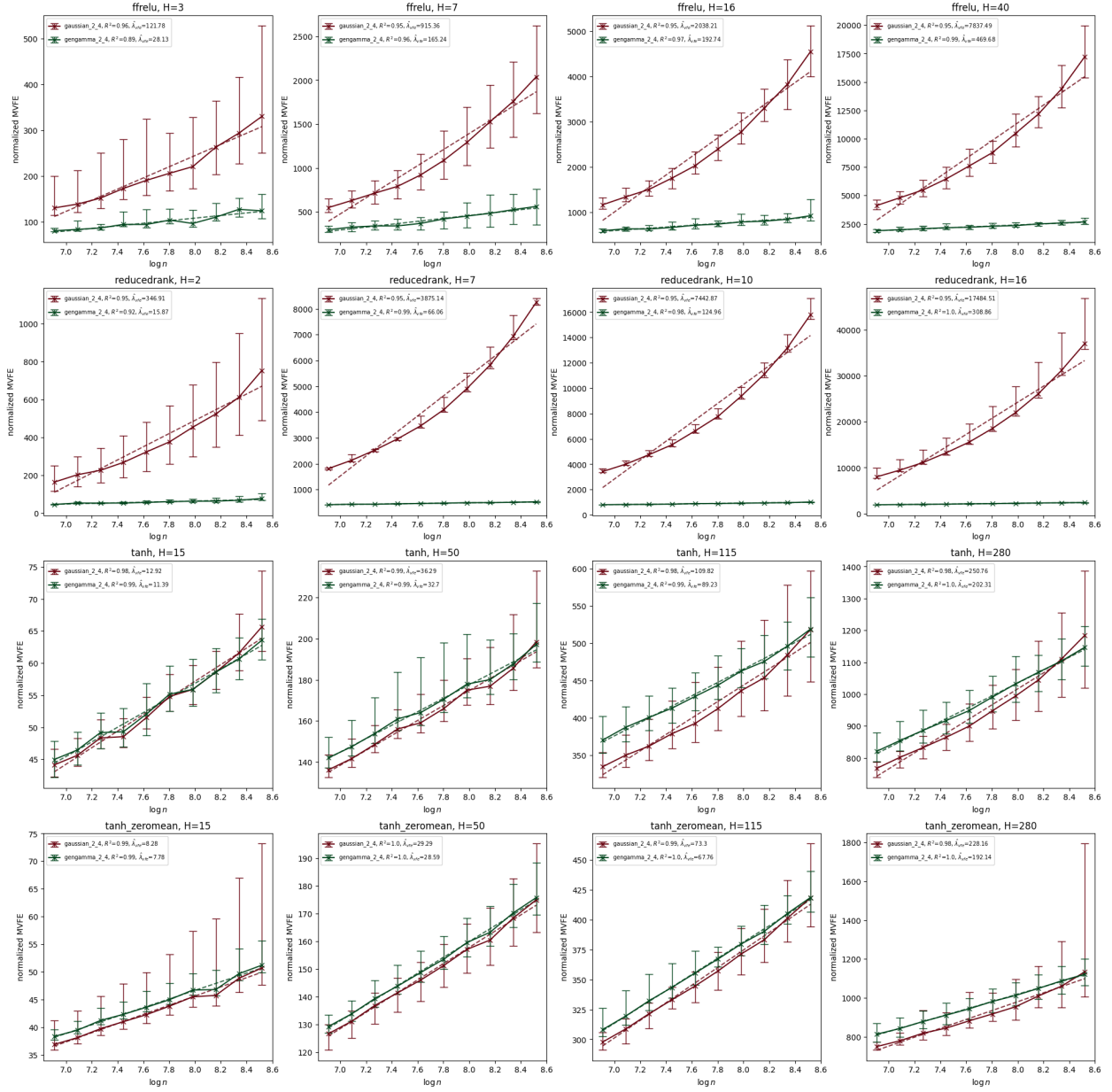
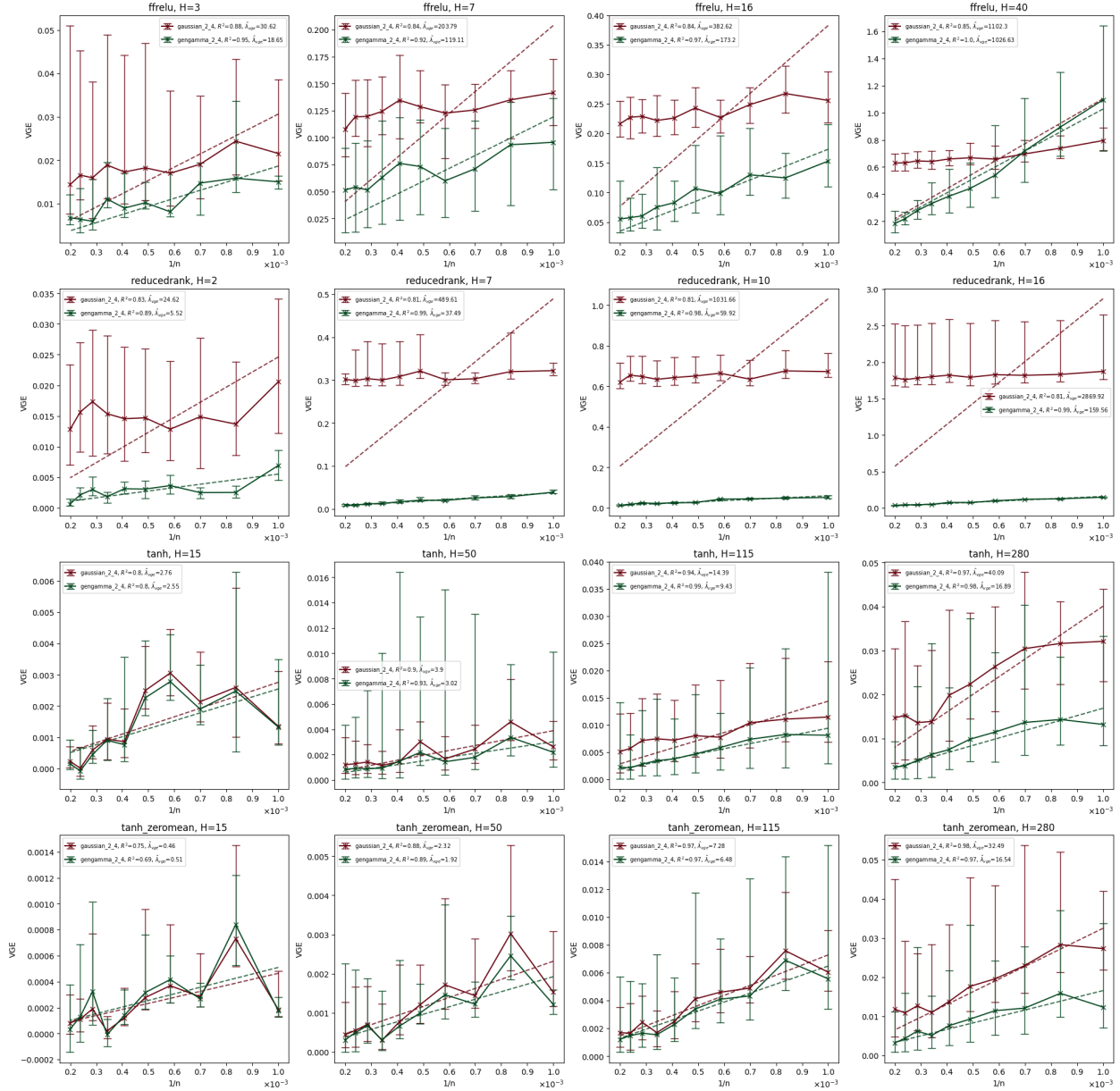


Figure 7: VGE for generalized gamma base distribution.

Figure 8: MVFE for G_θ with the least expressive 2_4 configuration.

Figure 9: VGE for G_θ with the least expressive 2_4 configuration.

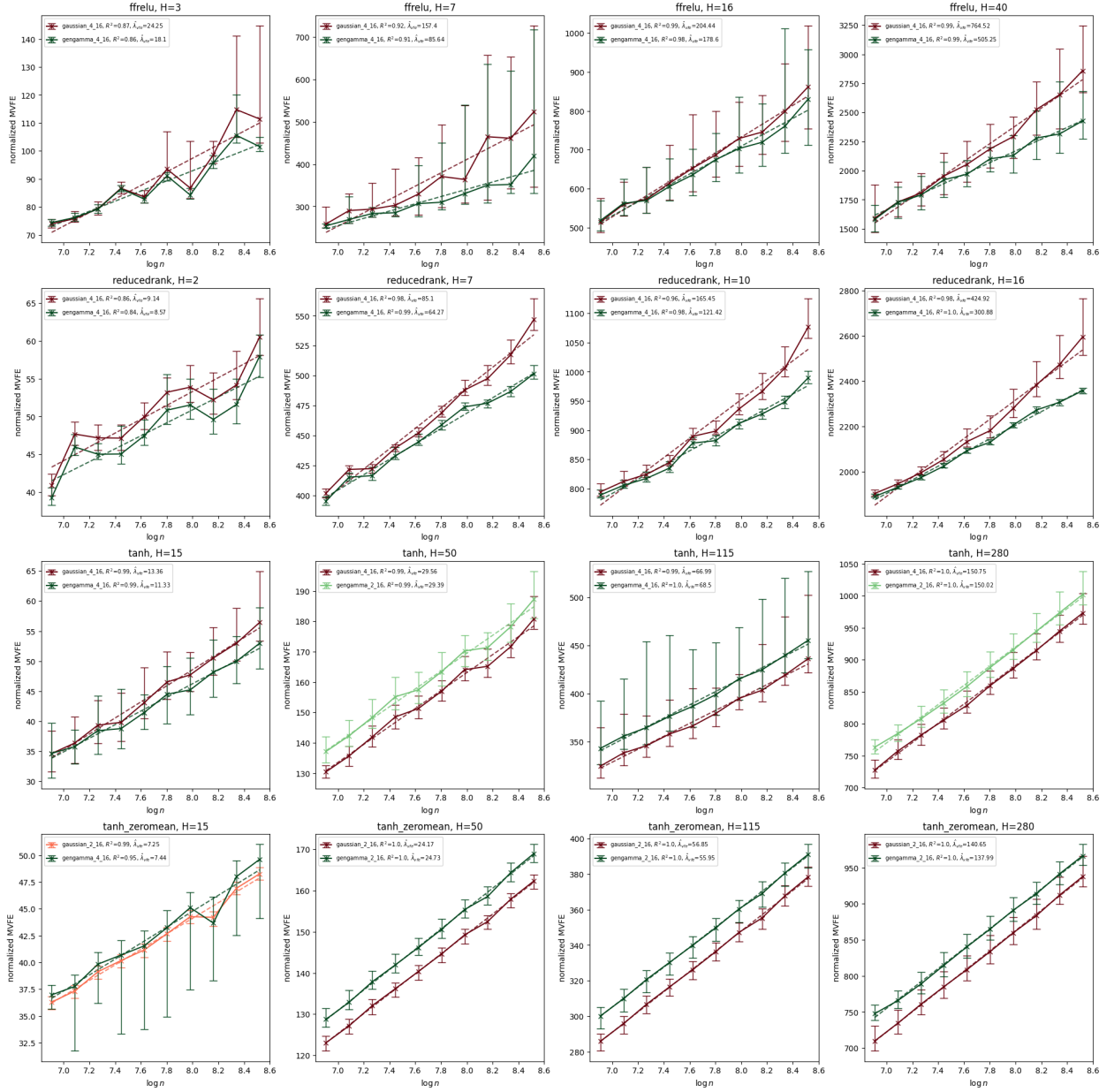


Figure 10: MVFE for G_θ with the best performing architecture for each base distribution, as judged by MVFE. This is usually the 4_16 configuration, but not always.

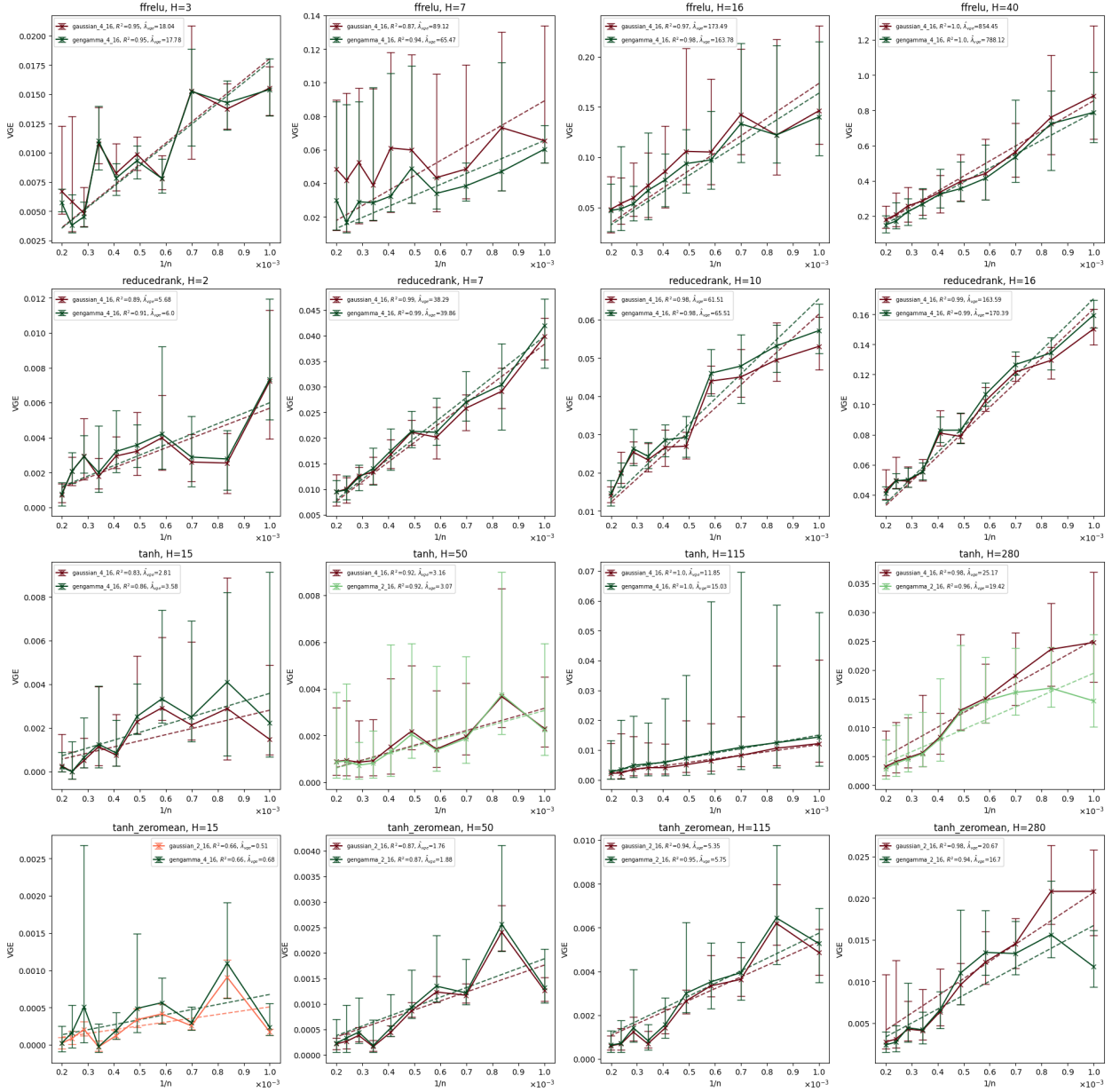


Figure 11: VGE for G_θ with the best performing architecture for each base distribution, as judged by MVFE. This is usually the 4_16 configuration, but not always.

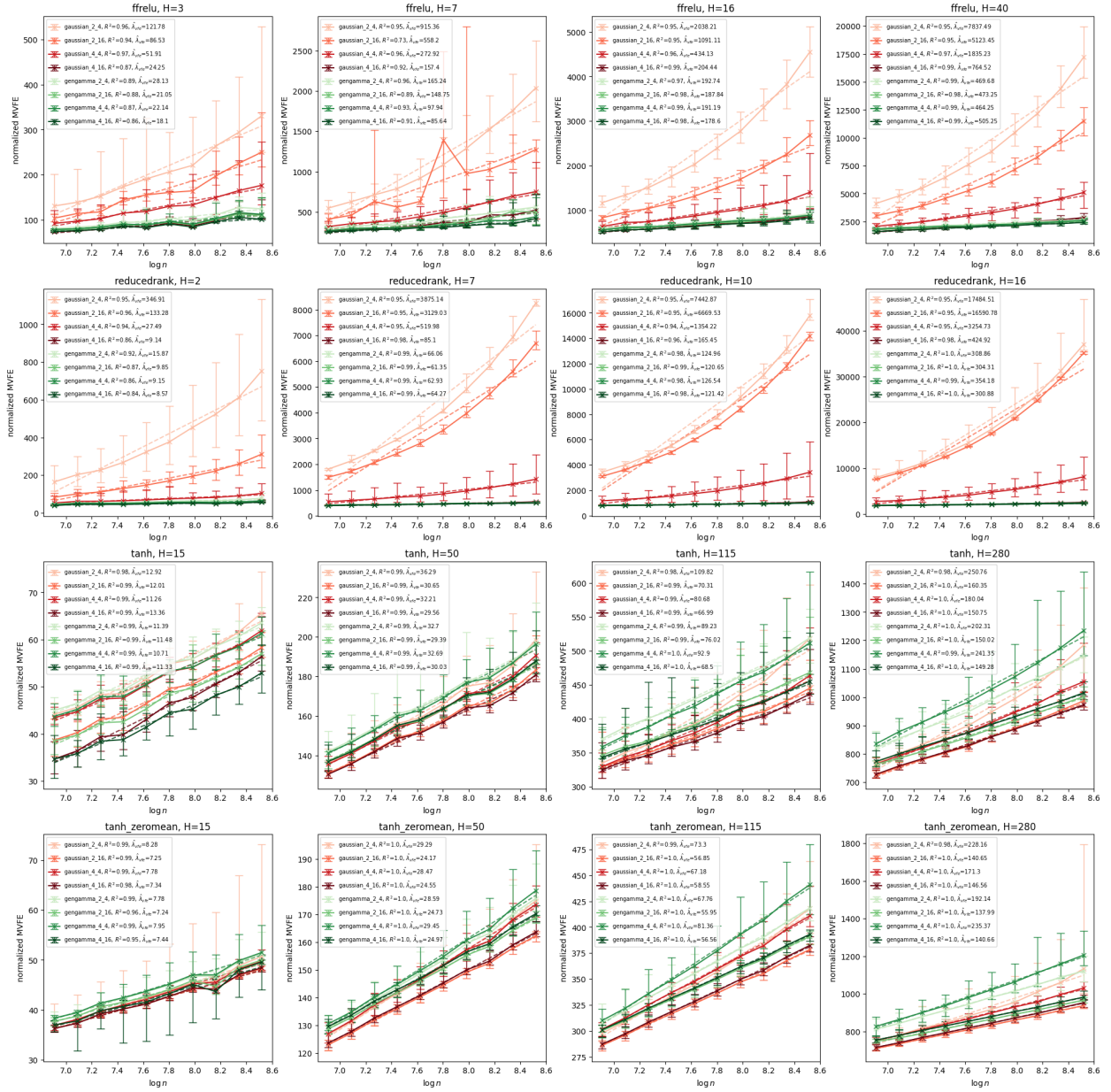


Figure 12: MVFE for all base distributions and G_θ architectures considered. Note that the first column of Figure 3 in the main text is a subset of the plots here.

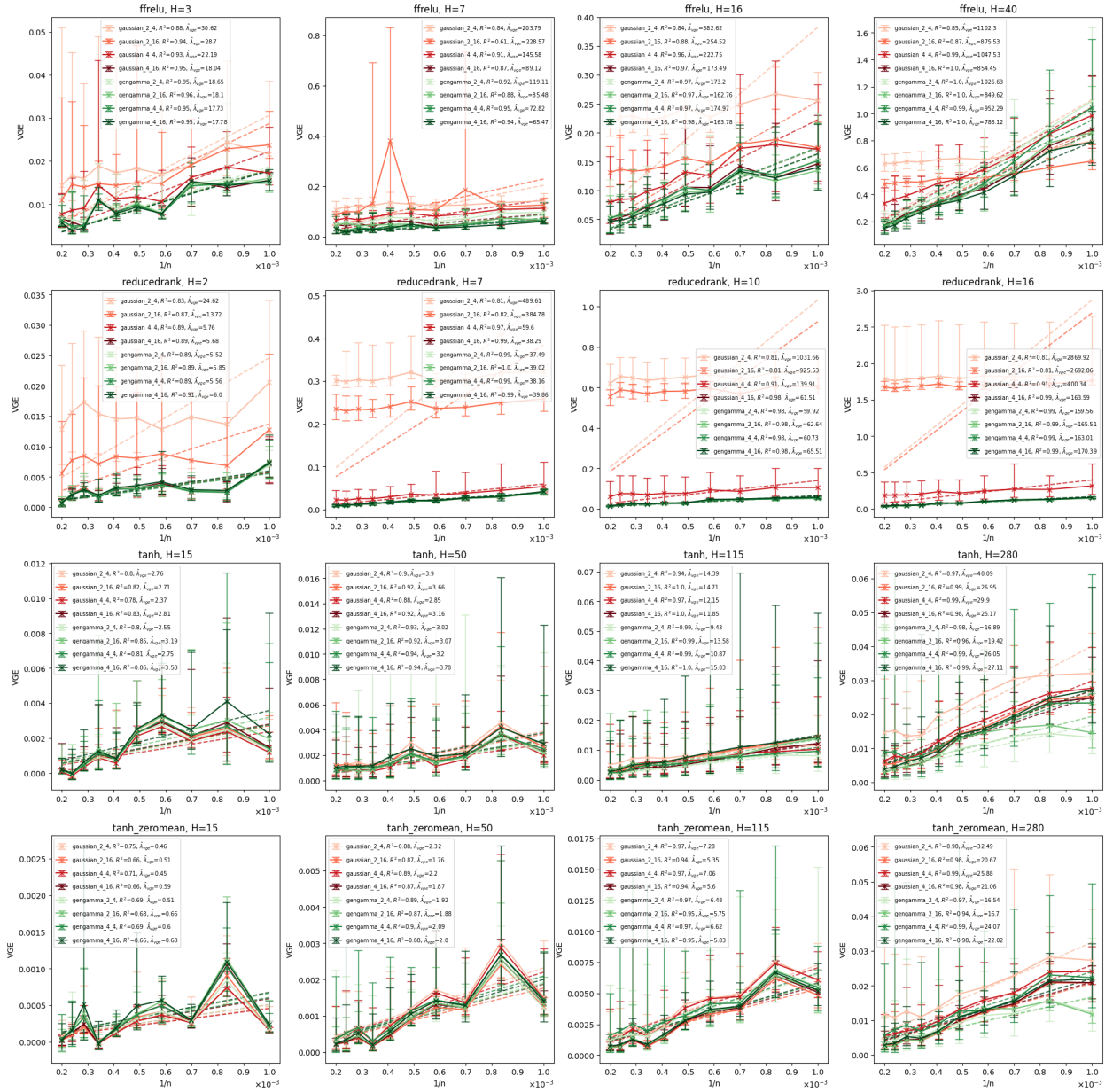


Figure 13: VGE for all base distributions and G_θ architectures considered. Note that the second column of Figure 3 in the main text is a subset of the plots here.