Wide Mean-Field Bayesian Neural Networks Ignore the Data

Beau Coker[*1] **Wessel P. Bruinsma**[*23] **David R. Burt**[*2] **Weiwei Pan**[1] **Finale Doshi-Velez**[1]
[1]Harvard University [2]University of Cambridge [3]Invenia Labs

Abstract

Bayesian neural networks (BNNs) combine the expressive power of deep learning with the advantages of Bayesian formalism. In recent years, the analysis of wide, deep BNNs has provided theoretical insight into their priors and posteriors. However, we have no analogous insight into their posteriors under approximate inference. In this work, we show that mean-field variational inference *entirely fails to model the data* when the network width is large and the activation function is odd. Specifically, for fully-connected BNNs with odd activation functions and a homoscedastic Gaussian likelihood, we show that the *optimal* mean-field variational posterior predictive (i.e., function space) distribution converges to the prior predictive distribution as the width tends to infinity. We generalize aspects of this result to other likelihoods. Our theoretical results are suggestive of underfitting behavior previously observered in BNNs. While our convergence bounds are non-asymptotic and constants in our analysis can be computed, they are currently too loose to be applicable in standard training regimes. Finally, we show that the optimal approximate posterior need not tend to the prior if the activation function is not odd, showing that our statements cannot be generalized arbitrarily.

1 INTRODUCTION

Bayesian neural networks (BNNs) provide a systematic method of capturing uncertainty in neural networks by placing priors on the weights of the network. Although it has been speculated for decades that BNNs are capable of combining the benefits of Bayesian inference and deep learning, we are only beginning to understand the theoretical properties of this model class and its associated inference techniques. One tool for understanding the behavior of modern BNNs with large architectures is to study the limiting behavior of this model as the number of hidden units in each layer, i.e., the *width* of the model, goes to infinity. In this case, the prior predictive distribution of a BNN converges in distribution to the *NNGP*, a Gaussian process (GP) with the *neural network kernel* that depends on the prior on the weights and architecture of the network (Neal, 1996; Matthews et al., 2018). Analogously, in the case of regression with a Gaussian likelihood, the associated BNN posterior converges to the NNGP posterior (Hron et al., 2020).

However, since exact inference for BNNs is intractable, approximate inference is commonly used in practical settings. While asymptotically exact sampling MCMC methods have been successfully applied to BNNs (Neal, 1996; Izmailov et al., 2021), these methods can require considerable amounts of computation and it is generally not feasible to ensure mixing. Variational inference offers a computationally appealing alternative by converting the problem of (approximate) inference into a gradient-based optimization problem.

Unfortunately, the properties of commonly used approximations of BNN posteriors, like mean-field variational inference (MFVI), have not been extensively studied. MFVI assumes complete posterior independence between the weights, but generalizing asymptotic analysis to this approximation is non-trivial. Unlike the true posterior predictive distribution, we do not know if the variational posterior predictive distribution approaches a GP as the width approaches infinity. We also do not know if documented properties of BNNs in the finite-width regime generalize to the wide limit. Empirical evidence suggests that finite BNNs trained with MFVI underestimate certain types of uncertainty (Foong et al., 2020) and underfit the data (Tomczak et al., 2021; Dusenberry et al., 2020). In the case of single hidden layer networks with ReLU activa-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. *Equal contribution.

tion, (Foong et al., 2020) showed that MFVI networks underestimate uncertainty in-between clusters of data, but their proof fundamentally cannot be extended to the case of several hidden layers. We establish the strong theoretical results for MFVI networks, under the assumption that they are sufficiently wide.

In this paper, we show that, unfortunately, a number of notable deficiencies of these approximate posteriors become more severe as width increases. For mean-field variational Bayesian neural networks of arbitrary depth with odd, Lipschitz activation functions, we prove a surprising result: the optimal variational posterior predictive distribution converges to the prior predictive distribution as the width tends to infinity. That is, asymptotically, the mean-field variational posterior predictive distribution of a wide BNN completely ignores the data, unlike the true posterior predictive distribution. Furthermore, we derive non-asymptotic, computable bounds that offer insight into the relative rates with which the number of observations, depth, and width of the network affect this convergence. The bounds we prove in their current form are generally too loose to provide numerically useful results for networks of the commonly trained widths, but they offer theoretical support for previously observed issues of underfitting in these networks. Finally, we show by a counterexample that this result does not hold for non-odd activation functions, including ReLU, but we provide an example showing that ReLU BNNs can nonetheless underfit data. Code to reproduce all of the experiments is available on GitHub.[1]

2 RELATED WORK

Wide-limits of BNNs. There are many works that analyze distributions over wide neural networks with the goal of gaining theoretical insight into neural network performance. As the width tends to infinity, Neal (1996) showed that single-layer, fully-connected BNN priors with bounded activation functions converge to GPs. Lee et al. (2017) and Matthews et al. (2018) extend this result to deeper networks with activations that satisfy a "linear envelope" condition (which includes ReLU, and is implied by Lipschitz-ness). Hron et al. (2020) extend the result by showing BNN posteriors converge to GP posteriors. All of these works can be seen as offering insights into modeling assumptions made when employing BNNs. Unfortunately, the *true* BNN posterior is computationally intractable for all but the smallest networks. In contrast, we analyze properties of approximate inference, which allows us to make statements about the BNN posterior typically used in practice.

[1] `https://github.com/dtak/wide-bnns-public`

Neural Tangent Kernel. Other works analyze wide neural networks after training the weights with gradient descent, showing that the network output approaches kernel regression with the neural tangent kernel (NTK) (Jacot et al., 2018; Lee et al., 2019). This also provides a Bayesian interpretation to ensembles of trained neural networks (He et al., 2020). The key insight in these works is that as the width increases, the weight parameters change less and less during training, permitting the network to be approximated by a first-order Taylor expansion around the initial weights. For this phenomenon to happen, these works assume the weights are unregularized during training (Chen et al., 2020). In contrast, in variational inference, one trains the variational parameters of a distribution over the weights, rather than the weights themselves. Because the variational parameters are regularized by the Kullback-Leibler (KL) divergence to the prior over the weights, the variational parameters do not stick near their initial values and thus the same first-order approximation cannot be used. We show that this regularization is too strong for wide networks, since it forces the resulting approximation of the posterior to converge to the prior as the width tends to infinity. Unlike NTK our result does not rely on the dynamics of any particular optimization algorithm, and instead characterizes the optimal posterior.

Issues with MFVI Inference in BNNs. Many works have empirically observed challenges with mean-field approximations to Bayesian neural networks. MacKay (1992) noted deficiencies of factorized Laplace approximations in single-hidden layer BNNs. However, little is known theoretically about mean-field variational inference in BNNs. Foong et al. (2020) showed that single-hidden layer networks with ReLU activations and mean-field distributions over the weights cannot have high variance between two regions with low variance. However, the authors also show that BNNs with two hidden layers can uniformly approximate *any* function-space mean and variance so long as the width is sufficiently large. Farquhar et al. (2020) suggested the universality result could be extended to other properties (e.g., higher moments) of the approximate posterior, leading them to recommend training deeper networks. This means that there *exist* mean-field variational distributions that do not exhibit the known pathologies of approximate inference in BNNs. However, despite this existence, we show that even for wide, deep networks the *optimal* mean-field variational distribution (i.e., the one that maximizes the evidence lower bound (ELBO)) converges to the prior, regardless of the data.

Trippe and Turner (2017) discuss *over-pruning*, which is the phenomenon whereby the variational posterior over many of the output-layer weights concentrates to a point mass around zero, allowing the variational posterior over any corresponding incoming weights to revert to the prior. This is undesirable behavior because the amount of over-pruning increases with the degree of over-parameterization and because over-pruning degrades performance — simpler models that do not permit pruning often perform better. As in our work, the explanation for over-pruning centers around the tension between the likelihood term and the KL divergence term in the objective function, the ELBO. To reduce the KL divergence, the optimization procedure may result in hidden units being pruned from the model (i.e., since many weights before the last layer can be set to the prior). Ultimately, we show that the KL divergence of the optimal variational posterior can only be so large, which prevents the variational posterior of wide networks from modeling anything but the prior.

Our work offers theoretical insight into earlier works on underfitting. Empirically, it has been found that re-scaling the regularization to the prior improves the performance of BNNs trained with variational inference (Osawa et al., 2019). This is closely related to observations regarding the performance of *cold posteriors*, which is the empirical phenomenon that down-weighting the importance of the KL divergence in the ELBO (and/or overcounting the data in the likelihood) yields better model performance (Wenzel et al., 2020). It is possible this practice serves to undo the over-regularization of the KL divergence that we investigate.

3 BACKGROUND

We consider the application of Bayesian neural networks in supervised learning: we have observed a dataset with N points, $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ with inputs $\mathbf{x}_n \in \mathbb{R}^{D_\text{i}}$ and outputs $\mathbf{y}_n \in Y$. Our goal is to infer a (probabilistic) mapping from $\mathbb{R}^{D_\text{i}}$ to Y that is consistent with the data and generalizes to new, unseen observations. We use a Bayesian neural network as the model for this mapping.

Bayesian Neural Networks (BNNs). Consider the feed-forward neural network of width M and depth L given by

$$\mathbf{f}(\mathbf{x}) = \tfrac{1}{\sqrt{M}}\mathbf{W}_{L+1}\phi(\mathbf{z}_L) + \mathbf{b}_{L+1}, \tag{1}$$

$$\mathbf{z}_\ell = \tfrac{1}{\sqrt{M}}\mathbf{W}_\ell\phi(\mathbf{z}_{\ell-1}) + \mathbf{b}_\ell \quad \text{for} \quad \ell = 2, \ldots, L, \tag{2}$$

$$\mathbf{z}_1 = \tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1 \tag{3}$$

$(\mathbf{W}_{L+1}, \mathbf{b}_{L+1}) \in \mathbb{R}^{D_\text{o}\times M} \times \mathbb{R}^{D_\text{o}}$, $(\mathbf{W}_\ell, \mathbf{b}_\ell) \in \mathbb{R}^{M\times M} \times \mathbb{R}^M$ for $\ell = 2, \ldots, L$, and $(\mathbf{W}_1, \mathbf{b}_1) \in \mathbb{R}^{M\times D_\text{i}} \times \mathbb{R}^M$ are the weight and bias parameters, respectively; $\phi\colon \mathbb{R} \to \mathbb{R}$ is the activation function, applied element-wise.

Let $\boldsymbol{\theta}$ represent the concatenation of all parameters. A *Bayesian* neural network places a prior distribution P over $\boldsymbol{\theta}$ and a likelihood distribution $\mathcal{L}(\boldsymbol{\theta})$ over Y conditional on $\boldsymbol{\theta}$. In this paper, we study the prior composed of independent standard Gaussian distributions over the weights: $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Often, we will be interested in the distribution induced over $\mathbf{f} = \mathbf{f}_{\boldsymbol{\theta}}$ through the randomness in $\boldsymbol{\theta}$. For a distribution over the weights, P', we will refer to the distribution induced over $\mathbf{f}_{\boldsymbol{\theta}}$ by P' as the P' *predictive distribution*. We note that this is a minor abuse of terminology, as a predictive distribution would typically be defined over subsets of Y and depends on the likelihood function. For example, in classification, the predictive refers to the distribution over the output of the network (i.e., logits).

Convergence to Gaussian Processes (GPs). As the width M tends to infinity, an application of the central limit theorem reveals that for any finite collection of inputs $\{\mathbf{x}_s\}_{s=1}^S$, the distribution over the neural network $\{\mathbf{f}(\mathbf{x}_s)\}_{s=1}^S$ induced by the prior P converges in distribution to a multivariate normal distribution (Neal, 1996; Matthews et al., 2018). In other words, as the width tends to infinity $\mathbf{f}$ converges to a multi-output Gaussian process, called the *neural network Gaussian process* (NNGP).

Variational Inference. Unfortunately, the posterior distribution of a finite-width BNN is not available in closed form. Markov chain Monte Carlo (MCMC) methods can be employed to approximately sample from the posteriors (e.g., Izmailov et al., 2021); however due to a high-dimensional and multi-modal posterior, these methods will generally not mix in a practical amount of time. Because of its advantageous computational properties on high-dimensional problems, variational inference is an appealing alternative (Blundell et al., 2015). Variational inference proposes a tractable family of distributions $\mathcal{Q}$ and finds an approximation of the true posterior $Q \in \mathcal{Q}$. This approximation is found by minimizing the KL divergence between Q and the true posterior, which is equivalent to maximizing a lower bound on the marginal likelihood called the evidence lower bound (ELBO):

$$\text{ELBO}(Q) = \mathbb{E}_{\theta\sim Q}[\log \mathcal{L}(\boldsymbol{\theta})] - \text{KL}(Q, P), \tag{4}$$

The first term in the ELBO is the expected log likelihood, which measures how well the model fits the data, and the second term is a regularization term, which measures how close Q is to the prior P.

A common choice for the family of variational distributions $\mathcal{Q}$ is the set of factorized (independent) Gaussian distributions. Under $Q \in \mathcal{Q}$, we write $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_Q, \mathrm{diag}(\boldsymbol{\sigma}_Q^2))$. Since both the prior and variational distribution are Gaussian, the KL divergence can be calculated in closed-form:

$$\mathrm{KL}(Q, P) = \tfrac{1}{2}(\|\boldsymbol{\mu}_Q\|_2^2 + \|r(\boldsymbol{\sigma}_Q^2)\|_1), \tag{5}$$

where $r\colon (0,\infty) \to [0,\infty)$, $r(a) = a - 1 - \log(a)$ is applied element-wise. Notice that Equation (5) acts like ℓ^2-regularization of the mean parameters, which will play an important role in the proof of Theorem 2. For this variational family $\mathcal{Q}$, under weak regularity conditions, it can be shown that an optimal solution $Q^* \in \arg\max_{Q\in\mathcal{Q}} \mathrm{ELBO}(Q)$ always exists (see Appendix B). Note, however, that an optimal solution is certainly not unique, because permutations of neurons have the same expected log-likelihood and KL divergence to the prior.

While mean-field variational inference scales gracefully from a computational perspective, its success ultimately relies on the variational family being sufficiently large so that the maximizer of the ELBO qualitatively resembles the posterior. In the next section, we prove that this fails badly for certain BNN models.

4 THE VARIATIONAL POSTERIOR PREDICTIVE REVERTS TO THE PRIOR PREDICTIVE

In this section, we analyze the convergence of optimal mean-field Gaussian variational posterior predictive distributions for Gaussian and other likelihoods. We give a sketch of the proof strategy. Additionally, we discuss the quantitative effect of depth and the number of observations on our results.

4.1 Gaussian Likelihood

We begin by stating a simplified version of our main result for a homoscedastic Gaussian likelihood: under fairly broad conditions, the variational BNN posterior predictive converges to the prior predictive.

We assume in our statements that the prior is $\mathcal{N}(\mathbf{0}, \mathbf{I})$; we additionally assume the network has no bias after the final hidden layer; an analogous result holds in the case with a final bias.

Theorem 1 (*Convergence in distribution to the prior, simplified*)**.** Assume a Gaussian likelihood and an odd, Lipschitz activation function. Then, for any fixed dataset, as the width tends to infinity, any finite-dimensional distribution of any optimal mean-field variational posterior predictive distribution of a BNN of any depth converges to the corresponding finite-dimensional distribution of the NNGP prior predictive distribution.

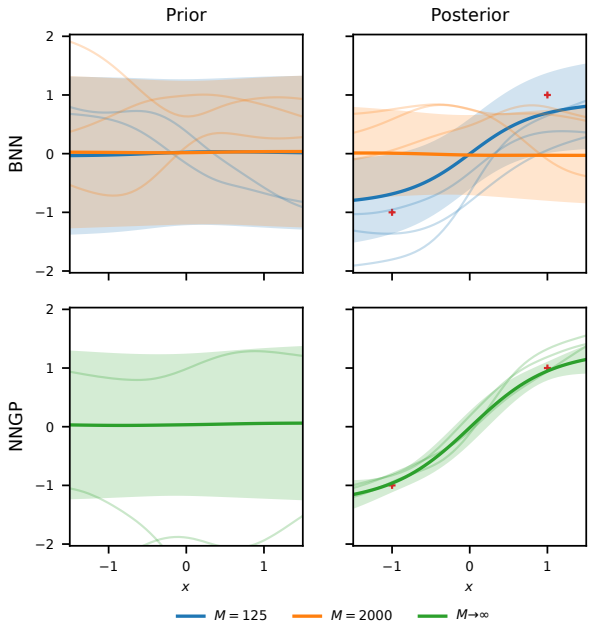

Figure 1: Prior and posterior predictive distributions for single-layer mean-field variational BNNs of different widths compared to the NNGP, to which the true posterior of the BNN converges. For a large width, the mean-field variational BNN ignores the data, unlike the NNGP. The shaded regions constitute ±1 standard deviation around the means (solid lines). All estimates are based on 1,000 function samples (a few of which are drawn faintly).

Figure 1 illustrates our result on a small dataset. In contrast to the *true* BNN posterior predictive, which converges to the NNGP posterior in the limit as the width approaches infinity, the *variational* BNN posterior predictive converges to the NNGP prior, completely ignoring the data.

A more general version of the theorem, which incorporates the final layer bias and allows for odd functions with a constant offset (e.g., a sigmoid activation), can be found in Appendix G. The output bias serves only to shift the network by a constant and can sometimes be optimized in closed-form (e.g., in the Gaussian likelihood case it accounts for the overall mean of the observations, $\bar{y}$). Theorem 1 and its generalization apply to several commonly used activation functions, notably tanh, sigmoid and linear.

While a Gaussian likelihood is necessary for our proof of convergence of the entire variational posterior predictive distribution to the prior predictive distribution,

we also prove convergence of the first two moments of the variational posterior predictive to the corresponding prior predictive moments for a variety of other likelihoods (logistic, Student's t). Additionally, we derive computable bounds on the first two moments of the variational posterior predictive distributions that show that for large, finite widths they must resemble the corresponding prior moments. In contrast, will see in Section 5 that the oddness assumption in Theorem 1 is necessary for any of these results.

4.2 General Likelihoods

Theorem 1 follows from a more general result that holds for a large class of likelihoods. In particular, for a range of likelihoods including Gaussian, Student's t, and logistic, we show convergence of the first two moments of the posterior predictive to the corresponding prior predictive moments. The convergence statement has two parts. First, we provide a non-asymptotic bound on the difference between the first two moments of the prior and approximate posterior predictive distributions (Theorem 2) and goes to 0 like $O(\frac{1}{\sqrt{M}})$. This aspect is independent of the likelihood and the upper bounds depend on $\mathrm{KL}(Q, P)$. Second, we provide an upper bound on $\mathrm{KL}(Q, P)$ that depends on the dataset and likelihood, but importantly, is independent of the width of the network (Lemma 3).

Theorem 2 (*Bounds on the mean and variance, simplified*)**.** Under the same conditions as Theorem 1 (except for the likelihood assumption), there exist universal constants $c_1, c_2, c_3, c_4 > 0$ such that

$$\begin{aligned}
&\|\mathbb{E}_Q[\mathbf{f}(\mathbf{x})] - \mathbb{E}_P[\mathbf{f}(\mathbf{x})]\|_2 \\
&\quad \leq c_1 c_2^{L-1} \frac{1 + \frac{1}{\sqrt{D_i}}\|\mathbf{x}\|_2}{\sqrt{M}} \mathrm{KL}(Q, P)(\mathrm{KL}(Q, P)^{\frac{L-1}{2}} \vee 1), \\
&\|\mathbb{E}_Q[\mathbf{f}^2(\mathbf{x})] - \mathbb{E}_P[\mathbf{f}^2(\mathbf{x})]\|_\infty \\
&\quad \leq c_3 c_4^{L-1} \frac{1 + \frac{1}{D_i}\|\mathbf{x}\|_2^2}{\sqrt{M}} \mathrm{KL}(Q, P)^{\frac{1}{2}}(\mathrm{KL}(Q, P)^{L+\frac{1}{2}} \vee 1)
\end{aligned}$$

where $a \vee b = \max(a, b)$.

In the special case when $L = 1$, our bound on the mean has the simpler form

$$\|\mathbb{E}_Q[\mathbf{f}(\mathbf{x})] - \mathbb{E}_P[\mathbf{f}(\mathbf{x})]\|_2 \leq \frac{2}{3}\left(\frac{1 + \frac{1}{D_i}\|\mathbf{x}\|_2^2}{M}\right)^{\frac{1}{2}} \mathrm{KL}(Q, P). \quad (6)$$

While a similar result to Equation (6) can be derived as a special case of Theorem 2, we derive this result specifically for the case $L = 1$ to improve the constant factors; see Appendix I.

Given the bounds in Theorem 2, we can immediately obtain convergence of the variational predictive mean and variance to the prior as $M \to \infty$ by bounding $\mathrm{KL}(Q^*, P)$ by a constant.

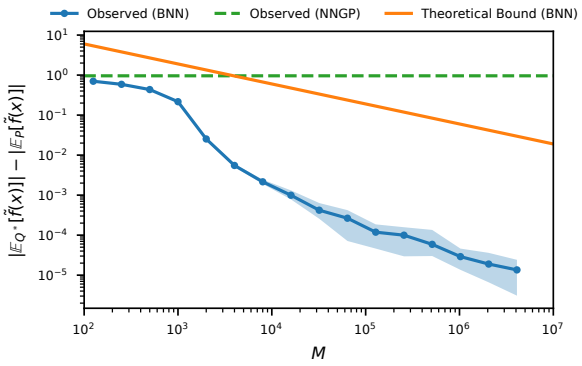

Figure 2: Maximum observed distance of the optimal posterior predictive mean to the prior predictive mean over a grid of points in $[-1, 1]$ compared to the theoretical $O(M^{-1/2})$ upper bound given by Theorem 2. For each M we train 10 single-layer networks on the same two observations shown in Figure 1. The shaded region shows the range of estimates over the 10 random initializations. We also show the analogous distance for the NNGP.

Lemma 3 (*Bounds on the KL, simplified*)**.** For Gaussian, Student's t, and logistic likelihood functions, and for an optimal mean-field variational posterior Q^*, $\mathrm{KL}(Q^*, P)$ is bounded by a constant that does not depend on the network width M.

Figure 2 illustrates the upper bound given by Equation (6) and Lemma 3 for the optimal posterior, Q^*. Empirically, the observed distance of the optimal posterior predictive mean to the prior predictive mean is well below the upper bound, which may be due to our bound of $\mathrm{KL}(Q^*, P)$. See Step 2 of Section 4.4 for further discussion of this bound. For example, above a width of 10^3, we observe the distance to the prior predictive within approximately 10^{-2}, which is well below scale of the y observations (-1 and $+1$) and the corresponding distance for the NNGP.

Figure 3 confirms that convergence to the prior leads to a poor fit of the data. We see that across datasets, the RMSE between the posterior mean and the test data increases with the network width (right panel). For comparison, we show the RMSE between the posterior and the prior mean (left panel), which decreases as expected. The datasets "concrete" and "slump" are from the UCI Machine Learning Repository and the rest are synthetic. The "2 points" dataset is the same as in Figures 1 and Figure 2. See Appendix L for details and an analogous plot of the posterior variance.

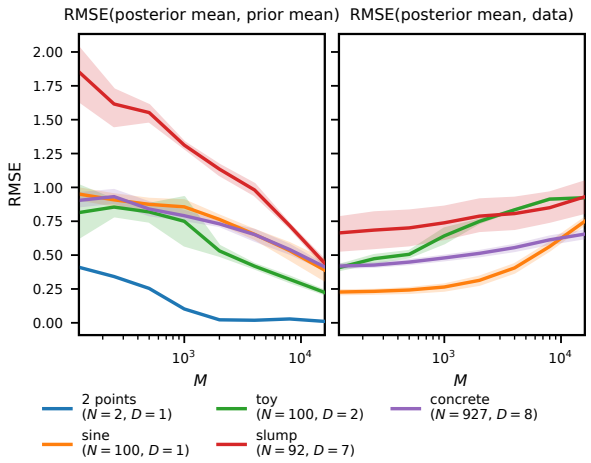

Figure 3: Root mean squared error (RMSE) of the posterior mean to the prior mean (i.e., $\mathbb{E}_P[f(x)] = 0$), and the data, y, for a few real and synthetic datasets. We use a tanh activation function. The shaded regions are 95% confidence intervals that reflect 5 train/test splits. The datasets "concrete" and "slump" are from the UCI Machine Learning Repository, while "2 points" is the same dataset from the previous figures. The posterior approaches the prior as the width (M) increases, as expected by Theorem 1, resulting in a poor fit of the data.

4.3 Influence of Depth and Number of Observations

In practice, the upper bounds given by Theorem 2 can be large, limiting their immediate use to practitioners. Furthermore, we see that greater depth increases the bound, since the dependence on $\mathrm{KL}(Q, P)$ grows with L and $c_2, c_4 > 1$. It is therefore important to investigate whether a faster rate of convergence can be achieved.

The case of linear networks (i.e., $\phi(z) = z$) provides for a relevant discussion. We show in Appendix J that

$$\|\mathbb{E}_Q[\mathbf{f}(\mathbf{x})] - \mathbb{E}_Q[\mathbf{f}(\mathbf{0})]\|_2 = \Theta(M^{-\frac{L}{2}} \mathrm{KL}(Q, P)^{\frac{L+1}{2}}). \quad (7)$$

Thus, Theorem 2 correctly captures the dependence on the KL divergence, but not the dependence on the width M, which is much faster for the linear case: $M^{-\frac{L}{2}}$ versus $M^{-\frac{1}{2}}$. This raises the question of whether the dependence on M can be improved in case of non-linear activations. Unfortunately, the answer in general is no. Appendix K shows an example where the dependence is $M^{-\frac{1}{2}}$. Although Theorem 2 cannot be generally improved for a generic Q, Q^* maximizes the ELBO, which introduces additional structure. In particular, in the Gaussian case, we know that $\mathrm{KL}(Q^*, P)$ tends to 0 with M (see Step 3 in Section 4.4), which could potentially be used to derive faster rates.

It is also important to consider the dependence of Theorem 2 on the number of observations, N, which influences the bound through $\mathrm{KL}(Q^*, P)$. If $\mathbb{E}[\|\mathbf{y}\|_2^2] = O(N)$, we show in Appendix F that $\mathrm{KL}(Q^*, P) \leq CN$ for a constant $C > 0$. Therefore, the first two moments of the optimal variational posterior predictive approach their respective values under the prior if $\lim_{N,M\to\infty} \frac{N^{L+1}}{M} = 0$. Hence, for our results to be non-vacuous for deep networks, M needs to be larger than for shallow networks.

4.4 Proof Sketch

The proof of Theorem 1 proceeds in three steps:

Step 1. Establish Theorem 2, which bounds the posterior predictive mean and variance at any $\mathbf{x}$ in terms of $\mathrm{KL}(Q, P)$.

Step 2. Establish Lemma 3. Combined with Theorem 2, it follows that, in the limit $M \to \infty$, the first and second moments of the approximate posterior predictive and the prior predictive agree.

Step 3. For a Gaussian likelihood, observe that the ELBO depends only on the first and second moments of the variational posterior predictive distribution at each datapoint and $\mathrm{KL}(Q, P)$. Since (i) $\mathrm{ELBO}(Q) \geq \mathrm{ELBO}(P)$ and (ii) the first and second variational predictive moment converge to the prior predictive moments, it follows that $\mathrm{KL}(Q, P) \to 0$.

The complete proof of step 1 can be found in Appendix D for the first moment and Appendix E for the second moment. A more complete version of step 2 that can be made quantitative is given in Appendix F. A version of step 3 incorporating the final bias can be found in Appendix G. Below we expand on each step to give insight into how it is achieved.

Step 1: Bounding the Moments. Here we prove the result for convergence of the mean for a network with $L = 1$ and sub-optimal constants. The variance argument follows a generally similar — though more involved — argument, and the $L > 1$ case is achieved by inductively applying a variant of the argument used in the $L = 1$ case.

We have

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})]\|_2 \overset{\text{(i)}}{=} \frac{1}{\sqrt{M}} \|\mathbb{E}[\mathbf{W}_2]\mathbb{E}[\phi(\frac{1}{\sqrt{D_i}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)]\|_2 \quad (8)$$

$$\overset{\text{(ii)}}{\leq} \frac{1}{\sqrt{M}} \|\mathbb{E}[\mathbf{W}_2]\|_F \|\mathbb{E}[\phi(\frac{1}{\sqrt{D_i}}\mathbf{W}_1 x + \mathbf{b}_1)]\|_2 \quad (9)$$

where in (i) we use independence and in (ii) we use that $\|\cdot\|_2 \leq \|\cdot\|_F$.

Define $\mathbf{W}' = \mathbf{W}_1 - \mathbb{E}[\mathbf{W}_1]$ and $\mathbf{b}' = \mathbf{b}_1 - \mathbb{E}[\mathbf{b}_1]$. Note that $(\mathbf{W}', \mathbf{b}') \overset{\mathrm{d}}{=} (-\mathbf{W}, -\mathbf{b})$ as these random variables are mean-centered and jointly Gaussian, hence symmetric about 0. Then,

$$\mathbb{E}[\phi(\mathbf{W}'\mathbf{x} + \mathbf{b}')] \overset{\text{(i)}}{=} \mathbb{E}[\phi(-\mathbf{W}'\mathbf{x} - \mathbf{b}')] \tag{10}$$
$$\overset{\text{(ii)}}{=} -\mathbb{E}[\phi(\mathbf{W}'\mathbf{x} + \mathbf{b}')], \tag{11}$$

where (i) follows from the equality in distribution and (ii) by oddness of ϕ. From this, we conclude $\mathbb{E}[\phi(\mathbf{W}'\mathbf{x} + \mathbf{b}')] = 0$. We then make the following calculation:

$$\|\mathbb{E}[\phi(\tfrac{1}{\sqrt{D_\mathrm{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)]\|_2 \tag{12}$$
$$= \|\mathbb{E}[\phi(\tfrac{1}{\sqrt{D_\mathrm{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) - \phi(\tfrac{1}{\sqrt{D_\mathrm{i}}}\mathbf{W}'\mathbf{x} + \mathbf{b}')]\|_2 \tag{13}$$
$$\overset{\text{(i)}}{\leq} \mathbb{E}\|\phi(\tfrac{1}{\sqrt{D_\mathrm{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)] - \phi(\tfrac{1}{\sqrt{D_\mathrm{i}}}\mathbf{W}'\mathbf{x} + \mathbf{b}')\|_2 \tag{14}$$
$$\overset{\text{(ii)}}{\leq} \mathbb{E}\|\tfrac{1}{\sqrt{D_\mathrm{i}}}(\mathbf{W}_1 - \mathbf{W}')\mathbf{x} + (\mathbf{b}_1 - \mathbf{b}')\|_2 \tag{15}$$
$$\overset{\text{(iii)}}{\leq} \|\mathbb{E}[\mathbf{W}_1]\|_\mathrm{F} \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2 + \|\mathbb{E}[\mathbf{b}_1]\|_2 \tag{16}$$

where (i) uses convexity of norm and Jensen's inequality, (ii) uses that ϕ is 1-Lipschitz, and (iii) combines the triangle inequality and $\|\bullet\|_2 \leq \|\bullet\|_\mathrm{F}$.

Combining Equation (9) and Equation (16) gives

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})]\|_2$$
$$\leq \tfrac{1}{\sqrt{M}}\|\mathbb{E}[\mathbf{W}_2]\|_\mathrm{F}\Big(\|\mathbb{E}[\mathbf{W}_1]\|_\mathrm{F}\tfrac{\|\mathbf{x}\|_2}{\sqrt{D_\mathrm{i}}} + \|\mathbb{E}[\mathbf{b}]\|_2\Big). \tag{17}$$

We now note that the Frobenius norm $\|\mathbb{E}[\mathbf{W}_2]\|_\mathrm{F}$ is the ℓ^2-norm of the mean parameters of weights in the second layer, and similar conditions apply to $\|\mathbb{E}[\mathbf{W}_1]\|_\mathrm{F}, \|\mathbb{E}[\mathbf{b}]\|_2$. Recalling Equation (5),

$$\|\mathbb{E}[\mathbf{W}_1]\|_\mathrm{F}, \|\mathbb{E}[\mathbf{W}_2]\|_\mathrm{F}, \|\mathbb{E}[\mathbf{b}]\|_2 \leq \sqrt{2\,\mathrm{KL}(Q, P)},$$

so

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})]\|_2 \leq \tfrac{1}{\sqrt{M}} 2(1 + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)\,\mathrm{KL}(Q, P). \tag{18}$$

This is of the same form as the bound in Theorem 2.

Step 2: Bounding $\mathrm{KL}(Q^*, P)$. In order for Theorem 2 to be useful, we need to understand how large $\mathrm{KL}(Q^*, P)$ could be. We make the following three assumptions when doing this:

(i) The likelihood factorizes over data points, i.e. $\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^N \log p(\mathbf{y}_n | \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))$, for some function p;

(ii) there exists a C such that $\log p(\mathbf{y}_n | \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)) \leq C$;

(iii) for any fixed $\mathbf{y}_n$, $\log p(\mathbf{y}_n | \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))$ can be lower bounded by a quadratic function in $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)$.

By the optimality of Q^*, we have

$$0 \leq \mathrm{ELBO}(Q^*) - \mathrm{ELBO}(P) \tag{19}$$
$$= \mathbb{E}_{\boldsymbol{\theta}\sim Q^*}[\log \mathcal{L}(\boldsymbol{\theta})] - \mathrm{KL}(Q^*, P) - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\log \mathcal{L}(\boldsymbol{\theta})]. \tag{20}$$

Rearranging and using the assumptions on $\log \mathcal{L}(\boldsymbol{\theta})$,

$$\mathrm{KL}(Q^*, P) \leq \mathbb{E}_{\boldsymbol{\theta}\sim Q^*}[\log \mathcal{L}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\log \mathcal{L}(\boldsymbol{\theta})] \tag{21}$$
$$\leq CN - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\textstyle\sum_{n=1}^N \log p(\mathbf{y}_n | \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))] \tag{22}$$
$$\leq CN - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\textstyle\sum_{n=1}^N h_n(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))] \tag{23}$$

where h_n is quadratic. Since h_n is quadratic, $\mathbb{E}_P[h_n(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))]$ is a linear combination of the first and second moments of $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)$. As we know the moments of $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)$ converge to those of the corresponding NNGP (Matthews et al., 2018), and since any convergent sequence is bounded, this gives an upper bound on $\mathrm{KL}(Q^*, P)$ that is independent of width.

Step 3: Convergence in Distribution. For Gaussian likelihoods, we can go one step further and prove Theorem 1 using the optimality of Q^* yet again. In particular, by the same argument as in the previous paragraph, we have

$$\mathrm{KL}(Q^*, P) \leq \mathbb{E}_{\boldsymbol{\theta}\sim Q^*}[\log \mathcal{L}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\log \mathcal{L}(\boldsymbol{\theta})]. \tag{24}$$

For simplicity, assume $\mathbf{y}_n = y_n \in \mathbb{R}$ and a homoscedastic likelihood with variance parameter σ^2 is used. Then, using

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^N (y_n - f(\mathbf{x}_n))^2 \tag{25}$$

in combination with $|(a-b)^2 - (a-c)^2| \leq 2|a||b-c| + |a^2 - b^2|$, we find that

$$\mathrm{KL}(Q^*, P) \leq \sum_{n=1}^N \Big[2|y_n||\mathbb{E}_{Q^*}[f(\mathbf{x}_n)] - \mathbb{E}_P[f(\mathbf{x}_n)]| + |\mathbb{E}_{Q^*}[f(\mathbf{x}_n)^2] - \mathbb{E}_P[f(\mathbf{x}_n)^2]|\Big]. \tag{26}$$

By Theorem 2, we conclude $\lim_{M\to\infty} \mathrm{KL}(Q^*, P) = 0$. Since the KL divergence between any finite dimensional distribution of the predictive of Q^* and P is upper bounded by this KL divergence, we conclude that a similar statement holds for finite-dimensional distributions. Finally, convergence in this sense implies weak convergence, so convergence of finite dimensional distributions of the posterior predictive of Q^* to the NNGP follows.

5 NON-ODD ACTIVATIONS

In Section 4, our theorems assume odd activation functions. The following theorem shows that this assumption is necessary.

Theorem 4 (*non-odd counterexample, simplified*)**.** Given any non-odd, 1-Lipschitz activation function ϕ (e.g., ReLU), we can construct a homoscedastic Gaussian likelihood and a dataset where the optimal mean-field variational mean is bounded away from the prior mean as the width tends to infinity[2].

Figure 4 illustrates this counterexample dataset along with the resulting mean-field posterior predictive distributions of networks with ReLU and erf activations (left panel). We also train on the same two observations as in Figures 1 and 2 (right panel). In the erf activation case, the posterior predictive converges to the prior predictive on both datasets, as expected by Theorem 1, and in the ReLU activation case the posterior predictive does not converge to the prior predictive on the counterexample dataset, as expected by Theorem 4 (recall that we mean without the output bias, which can generally differ from the prior). Interestingly, in the ReLU case, on the dataset that is not constructed as a counterexample, the approximate posterior closely resembles the prior. However, an examination of additional datasets in Figure 5 reveals the story is generally less clear. For some datasets, the wider networks are closer to the prior than the narrower networks, while for other datasets the opposite is true. See Section 6 for further discussion.

Our proof strategy starts by finding a sequence of variational distributions (indexed by M) with a mean function that does not tend to a constant. We then show that there exists a dataset for which the sequence of ELBOs defined by this sequence of variational distribution converges a number that exceeds the ELBOs of any sequence of variational distributions that have a mean function that does tend to a constant.

The key observation to the counterexample is that in the odd activation case, the expected value of the final layer post-activations, $\mathbb{E}_P[\phi(\mathbf{z}_L)]$, is zero, whereas in the non-odd activation case this expectation will generally depend on $\mathbf{x}$.

To construct the counterexample, we define Q_M as the mean-field variational distribution that is equivalent to the prior except in the last layer, where $\mathbf{w}_{L+1} \sim \mathcal{N}(\frac{1}{\sqrt{M}}\mathbf{1}, \mathbf{I})$. Notice that under Q_M the predictive mean is equal to $\mathbb{E}_P[\phi(\mathbf{z}_L)]$. Q_M will serve as a candidate set of distributions with non-constant means and "good" ELBOs. We select the Y values in our dataset to fall very near the mean predictor for Q_M, or more precisely, to coincide with the mean predictor as $M \to \infty$. This will ensure Q_M has small error. On the other hand, we can ensure that the error of any predictor that gives a constant prediction is large. The left panel of Figure 4 confirms that the posterior predictive under the odd activation (erf) converges to the prior, whereas the posterior predictive under the non-odd activation (ReLU) is able to model the data. These networks are very wide ($M \approx 4 \times 10^6$).

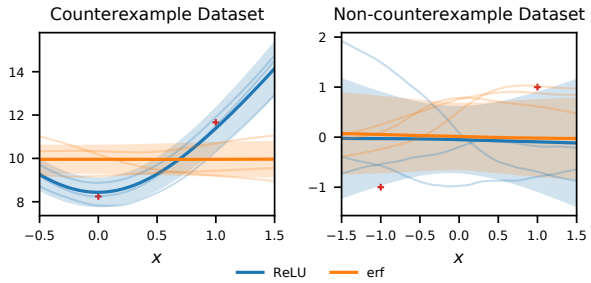

Figure 4: Mean-field posterior predictive distributions for very wide networks with odd and non-odd activations (erf and ReLU, respectively) trained on one of two datasets — the counterexample we construct (left panel) and the same dataset as in Figures 1 and 2, which does not meet the conditions of the counterexample (right panel). We observe convergence of the posterior predictive to the prior predictive in all cases except for the ReLU network trained on the counterexample dataset.

6 DISCUSSION

What Do Our Results Mean for Bayesian Methods in Over-Parameterized Models? The successes of modern deep learning have given strong evidence to the claim that over-parameterized models can lead to better empirical performance. As the flexibility of the model relative to the amount of data increases, it has been suggested that Bayesian methods have more to offer in terms of promoting generalization (Wilson and Izmailov, 2020). However, in cases when inaccurate inference is combined with very large models, our results prove a crippling and previously unknown limitation to this approach. This highlights the need for accurate inference methods in Bayesian neural networks, as well as robust techniques for monitoring and diagnosing inference quality.

Does Using ReLU Solve All of the Problems with MFVI? Another question raised by our results concerns the use of odd activation functions, which are necessary for our results to hold. Can the issues we raised be avoided by simply using a non-odd activation

[2]The theorem has additional technical conditions, but applies to all non-odd activation functions used in practice.

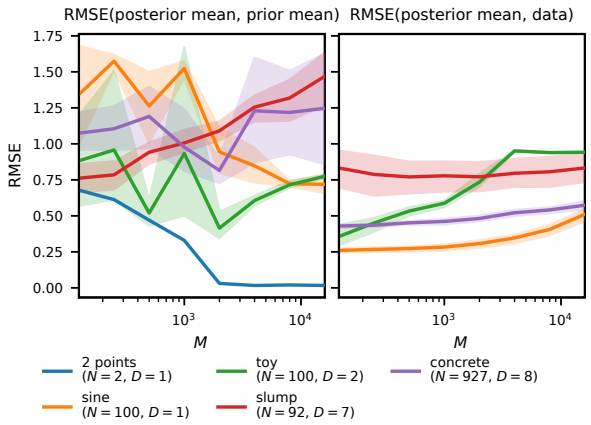

Figure 5: Analogous to Figure 3 but using a ReLU activation. Wider networks tend not to fit the data as well but it is unclear if this is due to convergence to the prior.

such as ReLU? While our counterexample shows there exists a dataset for which the approximate posterior mean under a non-odd activation does not converge to the prior, it is still possible the approximate posterior will converge to the prior on other datasets. Moreover, even without exact convergence to the prior, the approximate posterior could still be a poor model of the data. For the dataset in Figure 4 that does not meet the conditions of the counterexample, this is exactly what we see: a poor model of the data and a close resemblance of the approximate posterior to the prior. Figures 3 and 5 investigate these two attributes — distance to the data and distance to the prior — across a variety of datasets. In the tanh case, we see convergence to the prior as expected. However, in the ReLU case the behavior is unclear. For some datasets it is possible the approximate posterior is converging to the prior, whereas for other datasets there is little indication of this. Yet, we emphasize that in all cases (datasets and activations), we see an increasingly poor fit of the data as the network width increases. Our work frames the characterization of the optimal mean-field posterior under ReLU activations as an important area of future work.

Should MFVI be Abandoned Entirely in BNNs? Our results raise an important question for practitioners — should mean-field posteriors be thrown out, since asymptotically the optimal one converges to something degenerate, or should practitioners merely be careful about the relative scaling of the width, depth, and dataset size? By providing non-asymptotic upper bounds on how much the predictive mean and variance can differ from the prior, our results provide a regime where mean-field variational inference is guaranteed to fail. For example, for a given depth we can provide a width above which the optimal posterior predictive mean and variance are within a given threshold of their values under the prior. Yet, we emphasize our results are upper bounds, with constants that could likely be further optimized. There is possibly a smaller width that would give the same behavior, and this is what we observed in our experiments. These results suggest serious shortcomings of mean-field variational inference, and we would recommend practitioners take great care in applying MFVI, even with networks with narrower widths where our bounds do not provably show the approximate posterior will revert to the prior.

7 OPEN PROBLEMS

We believe our analysis leads to several interesting generalizations, which we leave as open problems. The first question we pose is whether Theorem 1 can be generalized to all likelihoods where Theorem 2 applies:

Conjecture 5. For any likelihood satisfying the assumptions needed for Lemma 3, the optimal variational posterior predictive (excluding the final bias) converges in distribution to the corresponding NNGP.

A potential avenue for proving Conjecture 5 would be to establish a central limit theorem for any one-dimensional predictive distribution under Q, using that $\mathrm{KL}(Q, P)$ is bounded. While the post-activations in the final hidden layer are not exchangeable, they are in some sense close to an exchangeable sequence. Given a central limit theorem Conjecture 5 can be established following the same argument sketched in Section 4.4, step 3. Establishing Conjecture 5 would make the consequences of the behavior we analyze to classification much clearer.

Another fascinating open question is the precise non-asymptotic dependence of Theorem 2 on the depth of the network. Our current bounds suggest that deeper networks may need to be wider before the optimal MFVI posterior converges to the prior. However, it is difficult to determine how much of this effect is due to the analysis becoming more complicated, leading to sub-optimal constants. We therefore pose the following, somewhat imprecise open problem,

Open Problem 6. Can the dependence of Theorem 2 on L be improved? In particular, should we expect the optimal MFVI posterior in deeper networks to converge more or less quickly to the prior as width increases?

We believe both of these questions are interesting theoretical questions with concrete ramifications for practitioners that may be challenging, but seem to be approachable problems for future research.

Acknowledgements The authors would like to thank Andrew Y.K. Foong for useful discussion and comments. Additionally, the authors would like to thank Richard E. Turner for helping to facilitate this collaboration. DRB acknowledges funding from the Qualcomm Innovation Fellowship. Wessel P. Bruinsma was supported by the Engineering and Physical Research Council (studentship number 10436152).

References

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. (2020). A generalized neural tangent kernel analysis for two-layer neural networks. In *Advances In Neural Information Processing Systems*.

Dusenberry, M., Jerfel, G., Wen, Y., Ma, Y., Snoek, J., Heller, K., Lakshminarayanan, B., and Tran, D. (2020). Efficient and scalable Bayesian neural nets with rank-1 factors. In *International conference on machine learning (ICML)*, pages 2782–2792. PMLR.

Farquhar, S., Smith, L., and Gal, Y. (2020). Liberty or depth: Deep Bayesian neural nets do not need complex weight posterior approximations. In *Advances In Neural Information Processing Systems*.

Foong, A. Y. K., Burt, D. R., Li, Y., and Turner, R. E. (2020). On the expressiveness of approximate inference in Bayesian neural networks. In *Neural Information Processing Systems (NeurIPS)*.

He, B., Lakshminarayanan, B., and Teh, Y. W. (2020). Bayesian deep ensembles via the neural tangent kernel.

Hron, J., Bahri, Y., Novak, R., Pennington, J., and Sohl-Dickstein, J. (2020). Exact posterior distributions of wide Bayesian neural networks. In *Workshop on Uncertainty and Robustness in Deep Learning (ICML)*.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. (2021). What are Bayesian neural network posteriors really like? *arXiv preprint arXiv:2104.14421*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2017). Deep neural networks as Gaussian processes. *arXiv preprint arXiv:1711.00165*.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

Matthews, A., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*.

Mykie (https://math.stackexchange.com/users/832/mykie) (2012). Density of polynomials in weighted $L^2(\mathbb{R}, w)$-space. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/126471 (version: 2012-03-31).

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Springer Verlag.

Osawa, K., Swaroop, S., Jain, A., Eschenhagen, R., Turner, R. E., Yokota, R., and Khan, M. E. (2019). Practical deep learning with Bayesian principles. *arXiv preprint arXiv:1906.02506*.

themaker (https://math.stackexchange.com/users/114509/themaker) (2020). Gaussian with zero mean dense in L^2. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/3749793 (version: 2020-07-08).

Tomczak, M. B., Swaroop, S., Foong, A. Y. K., and Turner, R. E. (2021). Collapsed variational bounds for bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 35.

Trippe, B. L. and Turner, R. E. (2017). Overpruning in variational Bayesian neural networks. In *Advances in Approximate Bayesian Inference (NeurIPS)*.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wenzel, F., Roth, K., Veeling, B. S., Swiatkowski, J., Mandt, L. T. S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the Bayes

posterior in deep neural networks really? In *International Conference on Machine Learning (ICML)*.

Wilson, A. G. and Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708.

Yeh, I.-C. (1998). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Modeling of strength of high performance concrete using artificial neural networks*, 28(12):1797–1808.

Yeh, I.-C. (2007). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29(6):474–480.

Supplementary Material: Wide Mean-Field Bayesian Neural Networks Ignore the Data

Table of Contents

Symbols

$\mathbf{W}, \mathbf{V}, \mathbf{U}$ Matrices are bold, capital letter

$\mathbf{w}, \mathbf{v}, \mathbf{u}$ Vectors are bold, lower case letter

M Number of neurons per hidden layer (width)

L Number of hidden layers (depth)

N Number of observations

K Shorthand for $\sqrt{\mathrm{KL}(Q, P)}$

θ All parameters in neural network

$\mathbf{x}$ An arbitrary input

$\mathbf{y}$ An arbitrary output

$z_{\ell,m}$ Preactivation for neuron m in hidden layer ℓ

$\mathbf{z_h}$ Vector of preactivations at layer h

$D_\mathbf{i}$ Dimensionality of input, i.e. $\mathbf{x} \in \mathbb{R}^{D_\mathrm{i}}$

$D_\mathbf{o}$ Dimensionality of output, i.e. $\mathbf{y} \in \mathbb{R}^{D_\mathrm{o}}$

ϕ Activation function in neural network (non-linearity)

$\phi_\mathbf{e}$ Even part of activation function, i.e. $\phi_\mathrm{e}(a) = \frac{\phi(a)+\phi(-a)}{2}$

$\phi_\mathbf{o}$ Odd part of activation function, i.e. $\phi_\mathrm{o}(a) = \frac{\phi(a)-\phi(-a)}{2}$

P Prior distribution, usually $\mathcal{N}(\mathbf{0}, \mathbf{I})$

Q A variational posterior distribution, $\mathcal{N}(\boldsymbol{\mu}_Q, \mathrm{diag}(\boldsymbol{\sigma}_Q^2))$

$\mathbb{E}$ Expectation, optionally with subscript to clarify the measure to be integrated over

$\mathbb{V}$ Variance, optionally with subscript to clarify the measure to be integrated over

$\mathbb{V}_\mathbf{d}$ Diagonal of $\mathbb{V}$

KL Kullback-Leibler divergence

$\mathbf{f}_\theta$ $\mathbf{f}_\theta : \mathbb{R}^{D_\mathrm{i}} \to \mathbb{R}^{D_\mathrm{o}}$ represents the output of the network with parameters θ

$\widetilde{\mathbf{f}}_\theta$ $\widetilde{\mathbf{f}}_\theta : \mathbb{R}^{D_\mathrm{i}} \to \mathbb{R}^{D_\mathrm{o}}$ represents the output of the network with parameters θ, excluding the contribution from the final bias

$\vee$ $a \vee b = \max(a, b)$

$\|\cdot\|_\mathbf{F}$ Frobenius Norm of a matrix, equal to the ℓ^2 norm of the singular values, also equal to the sum of squared entries

$\|\cdot\|_\mathbf{2}$ Spectral Norm of a matrix, equal to the ℓ^∞ norm of the singular values, also the matrix norm induced by the ℓ^2 norm on vectors

$\lesssim$ $f(x) \lesssim g(x) \iff$ there exists an irrelevant proportionality constant C such that $f(x) \leq Cg(x)$.

A Map of the Appendix and Sketches of the Results

Our main results show that for fully-connected networks with odd, Lipschitz continuous activation functions,

- For Gaussian likelihoods, the variational posterior converges in distribution to the prior as the width of the network tends to infinity.
- For a wide class of other likelihoods, including the Student's t likelihood, the first two moments of one-dimensional marginals of the variational posterior converge to the prior.

The key idea in both cases is to upper bound the difference between the first two moments of one-dimensional marginals of any mean-field posterior and the corresponding moments of the prior in terms of its KL divergence to the prior. Crucially, we show that we can derive a bound of this form, *that goes to* 0 *as* M *goes to* ∞. While we often make asymptotic statements about the width as these have the simplest form, all of the bounds are non-asymptotic and can be explicitly computed for finite widths.

In the following sketch, we ignore the final bias as it must be handled separately and leads to notational clutter and slightly more unwieldy statements. Hence, the following statements will all be true for a network without a final output bias, and some minor modifications of them is true for a normal fully-connected neural network.

Bounding the mean in terms of $\mathrm{KL}(Q, P)$ We use $\widetilde{\mathbf{f}}_\theta$ to denote the network output excluding the final output bias and even part of the activation, i.e., $\widetilde{\mathbf{f}}_\theta = \mathbf{f}_\theta - \mathbf{b}_{L+1} - \alpha \mathbf{W}_{L+1}\mathbf{1}$. Then if we consider the mean, and use the independence structure of Q,

$$\|\mathbb{E}_Q[\widetilde{\mathbf{f}}_\theta(\mathbf{x})]\|_2 = \tfrac{1}{\sqrt{M}}\|\mathbb{E}_Q[\mathbf{W}_{L+1}]\mathbb{E}[\phi_o(\mathbf{z}_L(\mathbf{x}))]\|_2 \leq \tfrac{1}{\sqrt{M}}\|\mathbb{E}_Q[\mathbf{W}_{L+1}]\|_2\|\mathbb{E}_Q[\phi_o(\mathbf{z}_L(\mathbf{x}))]\|_2. \tag{27}$$

The first term, $\|\mathbb{E}_Q[\mathbf{W}_{L+1}]\|_2$, can directly be upper bounded in terms of $\mathrm{KL}(Q, P)$. The second term is more difficult. The simplest thing to do would be to push the norm inside the expectation using Jensen's inequality and use that ϕ is assumed to be Lipschitz continuous. However, this prevents us from taking advantage of any cancellation due to the oddness of ϕ, which we will see is essential to the proof (cf. Theorem 41), and the resulting bound need not tend to 0 with M.

We instead setup a recursion to show that for each ℓ, $\|\mathbb{E}[\phi(\mathbf{z}_\ell(\mathbf{x}))]\|_2$ is upper bounded in terms of an expression that is independent of M. This will be the main work done in Appendix D.1, and crucially relies on the oddness and Lipschitz continuity of ϕ.

Bounding the variance in terms of $\mathrm{KL}(Q, P)$ We work with the un-centered second moment, as in combination with a bound on the mean this implies a bound on the marginal variance. The proof will proceed by splitting the variance into two terms. The first term, which we term the diagonal, arises from the product of the variance of the weights with the second moment of each activation. We show that under the optimal variational posterior, this term is close to the same term under the prior. The second term, which we term the off-diagonal arises from the product of the mean of the weights with the second moment of each activation. Under the prior, this term vanishes.

Convergence for Gaussian likelihoods Having established that the mean and variance of the variational posterior both converge to the prior, the proof diverges based on the likelihood. In the case of a homoscedastic, Gaussian likelihood, the analysis becomes particularly nice. By noting that the evidence lower bound essentially has three terms, one depending on the mean at each data point, one depending on the variance at each data point and one depending on the KL divergence between the variational posterior and the prior, we can upper bound the KL divergence between *any variational posterior with an ELBO at least as good as the prior* and the prior, by something that will tend to 0 with width. Combining this with standard inequalities between divergences on probability measures and the results of Matthews et al. (2018) suffices to show weak convergence of finite marginals of the the optimal posterior to the NNGP prior.

Convergence for more general likelihoods In the case of more general likelihoods, the evidence lower bound may depend on quantities besides the first and second moment of outputs of the variational posterior, so the above argument breaks down. However, so long as we can derive upper bounds on the KL divergence between

any optimal posterior and the prior that are independent of the width of the network, we can use our earlier results to conclude that the predictive mean and variance of any variational posterior with a better ELBO than the prior converges to the corresponding values under the prior. In order to upper bound the KL divergence, we assume that the log likelihood is bounded above and has a quadratic lower bound.

Counterexample for non-odd activations Appendix H examines whether it was essential for the results that the activation function is odd, or whether these results may be extended to other common activation functions such as ReLU. We answer this question in the negative, by exhibiting a dataset and Gaussian likelihood such that the mean of the optimal posterior does not converge to the prior mean. The key distinction in this case is that, where in the odd case we had, $\|\mathbb{E}_P[\phi_L(\mathbf{z}(\mathbf{x}))]\|_2 = 0$, in the non-odd case we have,

$$\|\mathbb{E}_P[\phi(\mathbf{z}_L(\mathbf{x}))]\|_2 = \Theta(\sqrt{M}). \tag{28}$$

This means that our earlier proof technique cannot possibly work. We use this observation to construct a counterexample.

B Existence of an Optimal Mean-Field Solution

Proposition 7. Let $\mathcal{Q}$ be the family of factorized Gaussian distributions. Assume the following:

(i) The activation $\phi\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous.

(ii) The likelihood factorizes over data points: $\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))$ for some function p.

(iii) The likelihood is continuous: for all $\mathbf{y}$, the function $\mathbf{f} \mapsto p(\mathbf{y} \,|\, \mathbf{f})$ is continuous.

(iv) The likelihood is upper bounded: there exists a $C \in \mathbb{R}$ such that, for all $\mathbf{y}$ and $\mathbf{f}$, $\log p(\mathbf{y} \,|\, \mathbf{f}) \leq C$.

(v) The likelihood admits a quadratic lower bound: for all $\mathbf{y}$, the function $\mathbf{f} \mapsto \log p(\mathbf{y} \,|\, \mathbf{f})$ can be lower bounded by a quadratic function in $\mathbf{f}$.

Then $\arg\max_{Q \in \mathcal{Q}} \mathrm{ELBO}(Q)$ is non-empty: an optimal mean-field solution $Q \in \arg\max_{Q \in \mathcal{Q}} \mathrm{ELBO}(Q)$ exists.

Proof. Let I denote the total number of parameters in the network. Then the variational optimization problem can be phrased as

$$\sup_{Q \in \mathcal{Q}} \mathrm{ELBO}(Q) \quad \text{where} \quad \mathrm{ELBO}\colon \mathcal{Q} \to \mathbb{R}, \quad \mathcal{Q} = \{\mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) : \boldsymbol{\mu} \in \mathbb{R}^I,\ \boldsymbol{\sigma}^2 \in (0,\infty)^I\}. \tag{29}$$

Call a sequence $(\mathcal{N}(\boldsymbol{\mu}_k, \mathrm{diag}(\boldsymbol{\sigma}_k^2)))_{k\geq 1} \subseteq \mathcal{Q}$ *parameter convergent* [to $\mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) \in \mathcal{Q}$] if $(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_i^2)_{k\geq 1} \subseteq \mathbb{R}^I \times (0,\infty)^I$ is convergent [to $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathbb{R}^I \times (0,\infty)^I$]. Using the definition of the supremum, extract a sequence $(Q_k)_{k\geq 1} \subseteq \mathcal{Q}$ such that $\mathrm{ELBO}(Q_k) \to \sup_{Q\in\mathcal{Q}} \mathrm{ELBO}(Q)$. The argument now consists of two parts. First, we show that there exists a subsequence $(Q_{n_k})_{k\geq 1} \subseteq (Q_k)_{k\geq 1}$ which is parameter convergent to some limit $Q^* \in \mathcal{Q}$ (compactness). Second, we show that ELBO is upper semi-continuous with respect to parameter convergence (continuity). Assuming the two parts,

$$\sup_{Q\in\mathcal{Q}} \mathrm{ELBO}(Q) = \lim_{k\to\infty} \mathrm{ELBO}(Q_k) = \limsup_{k\to\infty} \mathrm{ELBO}(Q_{n_k}) \overset{\text{(i)}}{\leq} \mathrm{ELBO}(Q^*), \tag{30}$$

where we use in (i) that ELBO is upper semi-continuous with respect to parameter convergence. Therefore, $Q^* \in \arg\max_{Q\in\mathcal{Q}} \mathrm{ELBO}(Q)$, which concludes the proof.

Compactness. We show that there exists a subsequence $(Q_{n_k})_{k\geq 1} \subseteq (Q_k)_{k\geq 1}$ which is parameter convergent to some limit $Q^* \in \mathcal{Q}$. Let $P = \mathcal{N}(\mathbf{0}, \mathbf{I}) \in \mathcal{Q}$ be the prior. Assume that $\mathrm{ELBO}(P) < \sup_{Q\in\mathcal{Q}} \mathrm{ELBO}(Q)$; for if equality holds, an optimal mean-field solution certainly exists. Since $\mathrm{ELBO}(Q_k) \to \sup_{Q\in\mathcal{Q}} \mathrm{ELBO}(Q)$, it follows that, for large enough k, $\mathrm{ELBO}(P) < \mathrm{ELBO}(Q_k)$. Therefore, by step 2 from Section 4.4, it follows that there exists a $C > 0$ such that, for large enough k, $\mathrm{KL}(Q_k, P) < C$. Consequently, denoting $Q_k = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$, by Appendix C and the observation that $r(a) < c$ for $c \geq 0$ implies that $a \in [R^{-1}, R]$ for some $R \in [1,\infty)$, it follows that there exists an $R \in [1,\infty)$ such that, for large enough k and all $i \in [I]$, $|\mu_{k,i}| \leq R$ and $R^{-1} \leq |\sigma_{k,i}^2| \leq R$. Hence, by Bolzano–Weierstrass, there exists a subsequence $(Q_{n_k})_{k\geq 1} \subseteq (Q_k)_{k\geq 1}$ which is parameter convergent to some limit $Q^* \in \mathcal{Q}$.

Continuity. Let $(Q_k)_{k\geq 1} \subseteq \mathcal{Q}$ be parameter convergent to some $Q \in \mathcal{Q}$. To conclude the proof, we show that $\limsup_{k\to\infty} \mathrm{ELBO}(Q_k) \leq \mathrm{ELBO}(Q)$. Decompose the ELBO as follows:

$$\mathrm{ELBO}(Q_k) = \textstyle\sum_{n=1}^{N} \mathbb{E}_{Q_k}[\log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))] - \mathrm{KL}(Q_k, P). \tag{31}$$

From Appendix C, $Q \mapsto -\mathrm{KL}(Q, P)$ is clearly continuous with respect to parameter convergence, so it is also upper semi-continuous with respect to parameter convergence. Hence, it remains to show that

$$\limsup_{k\to\infty} \mathbb{E}_{Q_k}[\log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))] \leq \mathbb{E}_{Q}[\log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))]. \tag{32}$$

To show this, we use the reparametrization trick: rewrite $\mathbb{E}_{Q_k}[\log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))] = \mathbb{E}[\log p(\mathbf{y}_n \,|\, \mathbf{f}^{Q_k}(\mathbf{x}_n))]$ where

$$\mathbf{f}^{Q_k}(\mathbf{x}) = \tfrac{1}{\sqrt{D_\mathrm{o}}}(\mathbf{S}_{L+1}^{Q_k} \circ \boldsymbol{\mathcal{E}}_{L+1} + \mathbf{M}_{L+1}^{Q_k})\phi(\mathbf{z}_L^{Q_k}) + (\mathbf{s}_{L+1}^{Q_k} \circ \boldsymbol{\varepsilon}_{L+1} + \mathbf{m}_{L+1}^{Q_k}), \tag{33}$$

$$\mathbf{z}_\ell^{Q_k} = \tfrac{1}{\sqrt{M}}(\mathbf{S}_\ell^{Q_k} \circ \boldsymbol{\mathcal{E}}_\ell + \mathbf{M}_\ell^{Q_k})\phi(\mathbf{z}_{\ell-1}^{Q_k}) + (\mathbf{s}_\ell^{Q_k} \circ \boldsymbol{\varepsilon}_\ell + \mathbf{m}_\ell^{Q_k}), \qquad \ell = L, \ldots, 2, \tag{34}$$

$$\mathbf{z}_1^{Q_k} = \tfrac{1}{\sqrt{D_\mathrm{i}}}(\mathbf{S}_1^{Q_k} \circ \boldsymbol{\mathcal{E}}_1 + \mathbf{M}_1^{Q_k})\mathbf{x} + (\mathbf{s}_1^{Q_k} \circ \boldsymbol{\varepsilon}_1 + \mathbf{m}_1^{Q_k}). \tag{35}$$

where $(\boldsymbol{\mathcal{E}}_\ell)_{\ell=1}^{L+1}$ are matrices of i.i.d. standard Gaussian random variables, $(\boldsymbol{\varepsilon}_\ell)_{\ell=1}^{L+1}$ are vectors of i.i.d. standard Gaussian random variables, $(\mathbf{M}_\ell^{Q_k})_{\ell=1}^{L+1}$ are matrices consisting of the means of each weight in each layer under Q_k, $(\mathbf{m}_\ell^{Q_k})_{\ell=1}^{L+1}$ are vectors consisting of the means of each bias in each layer under Q_k, $(\mathbf{S}_\ell^{Q_k})_{\ell=1}^{L+1}$ are matrices consisting of the standard deviations of each weight in each layer under Q_k, and $(\mathbf{s}_\ell^{Q_k})_{\ell=1}^{L+1}$ are vectors consisting of the standard deviations of each bias in each layer under Q_k. Since ϕ is Lipschitz, it is continuous, so clearly $\mathbf{f}^{Q_k}(\mathbf{x}) \to \mathbf{f}^{Q}(\mathbf{x})$. Let C be the upper bound on the likelihood. Then

$$\begin{aligned}
\limsup_{k\to\infty} \mathbb{E}_{Q_k}[\log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))] &= C + \limsup_{k\to\infty} \mathbb{E}[-C + \log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))] \\
&= C - \liminf_{k\to\infty} \mathbb{E}[C - \log p(\mathbf{y}_n \,|\, \mathbf{f}^{Q_k}(\mathbf{x}_n))] \\
&\overset{\text{(i)}}{\leq} C - \mathbb{E}[\liminf_{k\to\infty} (C - \log p(\mathbf{y}_n \,|\, \mathbf{f}^{Q_k}(\mathbf{x}_n)))] \\
&\overset{\text{(ii)}}{=} C - \mathbb{E}[C - \log p(\mathbf{y}_n \,|\, \mathbf{f}^{Q}(\mathbf{x}_n))] \\
&= \mathbb{E}[\log p(\mathbf{y}_n \,|\, \mathbf{f}^{Q}(\mathbf{x}_n))] \\
&= \mathbb{E}_{Q}[\log p(\mathbf{y}_n \,|\, \mathbf{f}(\mathbf{x}_n))],
\end{aligned}$$

where in (i) we use Fatou's lemma in combination with that $C - \log p(\mathbf{y}_n \,|\, \mathbf{f}^{Q_k}(\mathbf{x}_n)) \geq 0$ by definition of C and in (ii) we use (ii.a) continuity of $\mathbf{f} \mapsto p(\mathbf{y}_n \,|\, \mathbf{f})$ and (ii.b) $\mathbf{f}^{Q_k}(\mathbf{x}_n) \to \mathbf{f}^{Q}(\mathbf{x}_n)$. □

C Bounds on Parameters in Terms of the Kullback–Leibler Divergence

Fundamentally, if $\mathrm{KL}(Q, P)$ is small, we know that Q and P are 'close' in some sense. We want to translate this notion of 'close' to a notion directly related to the moments of the predictive distributions implied by the networks. In order to do this, we desire statements about how close the parameters of Q and P are, according to some norm. In this section, we show how to upper bound various norms of the parameters of Q in terms of $\mathrm{KL}(Q, P)$. These bounds will be a key ingredient in proofs of Theorems 11 and 23.

Lemma 8 (Kullback-Leibler Divergence between diagonal multivariate Gaussian distributions)**.** If $P = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $Q = \mathcal{N}(\boldsymbol{\mu}_Q, \mathrm{diag}(\boldsymbol{\sigma}_Q^2))$ It holds that

$$\mathrm{KL}(Q, P) = \tfrac{1}{2}(\|\boldsymbol{\mu}_Q\|_2^2 + \|r(\boldsymbol{\sigma}_Q^2)\|_1) \tag{36}$$

where $r\colon (0, \infty) \to [0, \infty)$, $r(a) = a - 1 - \log(a)$ is applied element-wise.

We note that r is a convex function, with a minimum at $r(1) = 0$. We now turn to proving bounds on the parameters of Q, which amounts to various methods of rearranging Lemma 8.

Lemma 9 (Bounds on parameters in term of KL divergence)**.** The following inequalities are true:

(i) $\|\boldsymbol{\mu}_Q\|_2^2 \leq 2\,\mathrm{KL}(Q,P)$,

(ii) $\|\boldsymbol{\sigma}_Q - \mathbf{1}\|_2^2 \leq 2\,\mathrm{KL}(Q,P)$,

(iii) $\sigma_{\max} \leq 1 + \sqrt{2\,\mathrm{KL}(Q,P)} \leq 2\sqrt{2\,\mathrm{KL}(Q,P) \vee 1}$, and

(iv) $\|\boldsymbol{\sigma}_Q^2 - \mathbf{1}\|_2^2 \leq (\sigma_{\max}+1)^2\|\boldsymbol{\sigma}_Q - 1\|_2^2 \leq (2+\sqrt{2\,\mathrm{KL}(Q,P)})^2(2\,\mathrm{KL}(Q,P))$.

Proof. We prove each bound in turn.

(i): This follows directly from Lemma 8 and non-negativity of norms.

(ii): To prove (ii), we recall the identity $\log(a') \leq a'-1$ for all $a'>0$. Define $a' = \sqrt{a}$. Then $\log(a) = 2(\sqrt{a}-1)$. Rearranging, $\log(a) \leq 2\sqrt{a}-2$ for all $a>0$. Then compute

$$r(a) = a - 1 - \log(a) \geq a - 1 - (2\sqrt{a}-2) = a - 2\sqrt{a} + 1 = (\sqrt{a}-1)^2. \tag{37}$$

(iii): For (iii), estimate

$$\sigma_{\max} - 1 \leq |\sigma_{\max} - 1| \leq \|\boldsymbol{\sigma}_Q - \mathbf{1}\|_2 \leq \sqrt{2\,\mathrm{KL}(Q,P)}, \tag{38}$$

where the last inequality uses (ii).

(iv): For (iv), factoring the difference of two squares,

$$\begin{aligned}
\|\boldsymbol{\sigma}_Q^2 - \mathbf{1}\|_2^2 &= \textstyle\sum_{i=1}^I (\sigma_i^2 - 1)^2 && (39)\\
&= \textstyle\sum_{i=1}^I (\sigma_i + 1)^2(\sigma_i - 1)^2 && (40)\\
&\leq \textstyle\sum_{i=1}^I (\sigma_{\max} + 1)^2(\sigma_i - 1)^2 && (41)\\
&= (\sigma_{\max}+1)^2 \textstyle\sum_{i=1}^I (\sigma_i - 1)^2 && (42)\\
&= (\sigma_{\max}+1)^2 \|\boldsymbol{\sigma}_Q - \mathbf{1}\|_2^2 && (43)\\
&\leq (2+\sqrt{2\,\mathrm{KL}(Q,P)})^2(2\,\mathrm{KL}(Q,P)) && (44)
\end{aligned}$$

where in the final inequality we have used (ii) and (iii). □

Proposition 10. Let $\mathbf{W}$ be an arbitrary weight matrix and let $\mathbf{b}$ be an arbitrary bias vector. Then

$$\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty + \|\mathbb{V}_{\mathrm{d}}[\mathbf{b}]\|_\infty + \|\mathbb{E}[\mathbf{b}^2]\|_\infty \leq (\sqrt{2} + \sqrt{2\,\mathrm{KL}(Q,P)})^2. \tag{45}$$

Proof. Denote $K = \sqrt{2\,\mathrm{KL}(Q,P)}$. By Lemma 8, we have the constraint

$$\|r(\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})])\|_1 + \|\mathbb{E}[\mathbf{b}]\|_2^2 + \|r(\mathbb{V}_{\mathrm{d}}[\mathbf{b}])\|_1 \leq K^2. \tag{46}$$

We argue that this constraint implies that $\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty + \|\mathbb{E}[\mathbf{b}^2]\|_\infty \leq (\sqrt{2}+K)^2$. To argue this, consider optimising $\|\boldsymbol{\sigma}_1^2\|_\infty + \|\boldsymbol{\mu}_2^2\|_\infty + \|\boldsymbol{\sigma}_2^2\|_\infty$ over $(\boldsymbol{\sigma}_1^2, \boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$ such that $\|r(\boldsymbol{\sigma}_1^2)\|_1 + \|\boldsymbol{\mu}_2\|_2^2 + \|r(\boldsymbol{\sigma}_2^2)\|_1 \leq K^2$. Without loss of generality, assume that $\boldsymbol{\sigma}_1^2 \succeq 1$ and $\boldsymbol{\sigma}_2^2 \succeq 1$. Then, without loss of generality, by the observation that the objective comprises ∞-norms, for all $i \geq 2$, assume that $\sigma_{1,i}^2 = \sigma_{2,i}^2 = 1$ and $\mu_i = 0$. Since, on $[1,\infty)$, $r'(a) = 1 - \frac{1}{a} < 1$ and r' is strictly increasing, it is clear that, at the maximum, $\sigma_{1,1}^2 = \sigma_{2,1}^2 = \sigma^2$ and $\mu_1 = 0$. From Lemma 9 and the constraint, we have that $2(\sigma-1)^2 \leq 2r(\sigma^2) \leq K^2$, so $\sigma \leq 1 + \frac{1}{\sqrt{2}}K$. Therefore, at the maximum, $\|\boldsymbol{\sigma}_1^2\|_\infty + \|\boldsymbol{\mu}_2^2\|_\infty + \|\boldsymbol{\sigma}_2^2\|_\infty = 2\sigma^2 \leq 2(1+\frac{1}{\sqrt{2}}K)^2 = (\sqrt{2}+K)^2$. □

D Proof of Convergence of the Mean of the Variational Posterior for Odd Activation Functions

The main result in this section we prove will be the following,

Theorem 11 (Convergence of mean prediction)**.** Let Q be a mean-field variational posterior and $P = \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote the prior over a neural network with L hidden layers and M neurons per hidden layer. Suppose $\phi_{\mathrm{e}} = \alpha$ for some $\alpha \in \mathbb{R}$ and $\phi \colon \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz. Let $\mathbf{x} \in \mathbb{R}^{D_{\mathrm{i}}}$ and let $\widetilde{\mathbf{f}}_\theta(\mathbf{x}) = \mathbf{f}_\theta(\mathbf{x}) - \frac{\alpha}{\sqrt{M}} \mathbf{W}_{L+1} \mathbf{1} - \mathbf{b}_{L+1}$ denote the network output excluding final bias and even part of the final activation. Then there exist universal constants $c_1 \leq 4$ and $c_2 \leq 6$ such that

$$\|\mathbb{E}_Q[\widetilde{\mathbf{f}}_\theta(\mathbf{x})]\|_2 \leq c_1 c_2^{L-1} L \frac{|\alpha| + 1 + \|\mathbf{x}\|_2/\sqrt{D_{\mathrm{i}}}}{\sqrt{M}} \operatorname{KL}(Q, P)((2 \operatorname{KL}(Q, P))^{\frac{L-1}{2}} \vee 1). \tag{47}$$

For the proof we assume the Lipschitz constant of the activation function is 1, but we note that any Lipschitz function can be scaled to have a Lipschitz constant of 1.

Corollary 12. With the same notation and assumptions as in Theorem 11, for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{D_{\mathrm{i}}}$ we have

$$\|\mathbb{E}_Q[\mathbf{f}_\theta(\mathbf{x})] - \mathbb{E}_Q[\mathbf{f}_\theta(\mathbf{x}')]\|_2 \leq c_1 c_2^{L-1} L \frac{2|\alpha| + 2 + (\|\mathbf{x}\|_2 + \|\mathbf{x}'\|_2)/\sqrt{D_{\mathrm{i}}}}{\sqrt{M}} \operatorname{KL}(Q, P)((2 \operatorname{KL}(Q, P))^{\frac{L-1}{2}} \vee 1) \tag{48}$$

Corollary 12 follows from Theorem 11 by noting that $\mathbb{E}_Q[\mathbf{f}_\theta(\mathbf{x})] - \mathbb{E}_Q[\mathbf{f}_\theta(\mathbf{x}')] = \mathbb{E}_Q[\widetilde{\mathbf{f}}_\theta(\mathbf{x})] - \mathbb{E}_Q[\widetilde{\mathbf{f}}_\theta(\mathbf{x}')]$, then applying triangle inequality.

The crucial technical result for the proof of Theorem 11 will be the following lemma, which upper bounds the norm of the expected value of the final layer of hidden units. We defer the proof of this lemma, which essentially inducts on the number of hidden layers, to Appendix D.1.

Lemma 13. Suppose Q is mean-field Gaussian, $P = \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\phi_e = \alpha$, with $\alpha \in \mathbb{R}$ and ϕ is 1-Lipschitz. Then,

$$\|\mathbb{E}[\phi_{\mathrm{o}}(\mathbf{z}_L(\mathbf{x}))]\|_2 \leq 2L(2 + |\alpha| + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(2 + 2c)^{L-1} \sqrt{2 \operatorname{KL}(Q, P)} (\sqrt{2 \operatorname{KL}(Q, P)} \vee 1)^{L-1}, \tag{49}$$

where $c > 0$ is a universal constant.

We now turn to the proof of Theorem 11, which is relatively direct once Lemma 13 has been established.

Proof of Theorem 11. Let every expectation be under Q. Define $K = \sqrt{2 \operatorname{KL}(Q, P)}$. By a slight abuse of notation, let $\mathbf{z}_L = \mathbf{z}_L(\mathbf{x})$.

To begin with, note that

$$\|\mathbb{E}[\widetilde{\mathbf{f}}_\theta(\mathbf{x})]\|_2 = \tfrac{1}{\sqrt{M}} \|\mathbb{E}[\mathbf{W}_{L+1}] \mathbb{E}[(\phi_{\mathrm{o}}(\mathbf{z}_L)]\|_2. \tag{50}$$

Using Lemma 9, $\|\mathbb{E}[\mathbf{W}_{L+1}]\|_{\mathrm{F}} \leq K$. We then apply Lemma 13,

$$\begin{aligned} \|\mathbb{E}[\widetilde{\mathbf{f}}_\theta(\mathbf{x})]\|_2 &\leq \frac{K}{\sqrt{M}} 2L \left(2 + |\alpha| + \frac{1}{\sqrt{D_{\mathrm{i}}}} \|\mathbf{x}\|_2\right) (2 + 2c)^{L-1} K (K \vee 1)^{L-1} && (51) \\ &= \frac{2L}{\sqrt{M}} \left(2 + |\alpha| + \frac{1}{\sqrt{D_{\mathrm{i}}}} \|\mathbf{x}\|_2\right) (2 + 2c)^{L-1} \operatorname{KL}(Q, P)((2 \operatorname{KL}(Q, P))^{\frac{L-1}{2}} \vee 1). && (52) \end{aligned}$$

□

D.1 Main Recursion: Proof of Lemma 13

The main purpose of this section will be the proof of Lemma 13. We begin by proving several results that build up to bounds on the norm of the expected value and the expected value of the norm of one layer in terms of the previous layer (Lemmas 18 and 19). Once we have established these bounds, the proof of Lemma 13 follows by recursive application of these bounds. See Figure 6 for a diagram of the dependencies between the results.

For a matrix $\mathbf{W}$, recall that $\|\mathbf{W}\|_2 \leq \|\mathbf{W}\|_{\mathrm{F}}$. Call a random variable *a symmetric around its mean* if $a - \mathbb{E}[a] \overset{\mathrm{d}}{=} -(a - \mathbb{E}[a])$

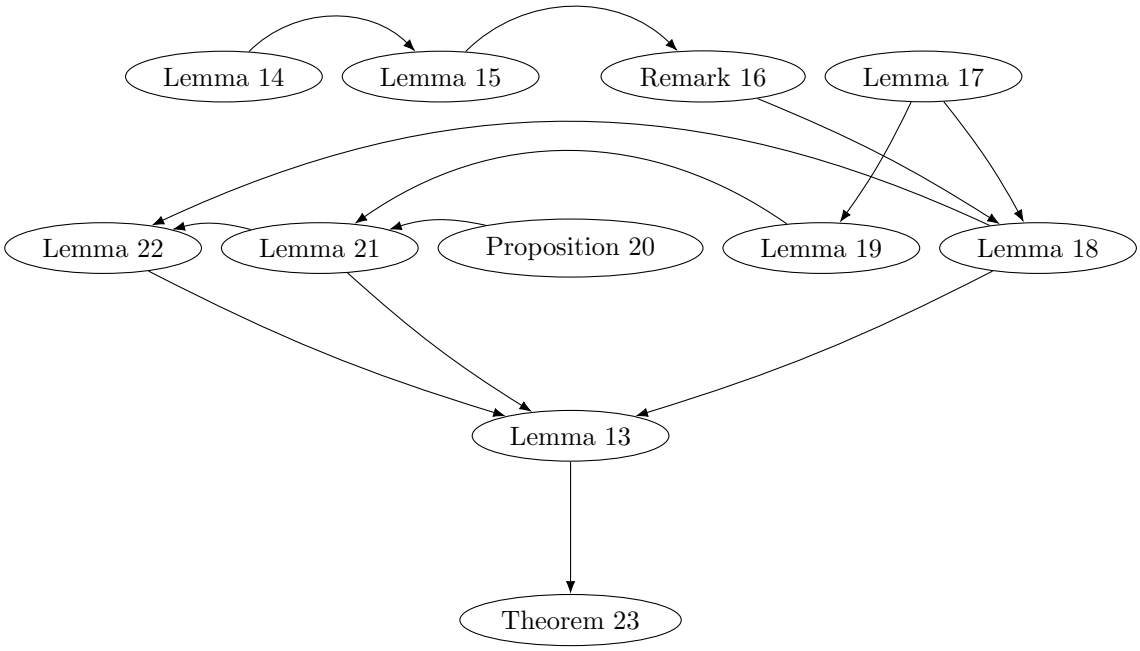

Figure 6: Dependency structure of the results in Appendix D.

Lemma 14. Let $\phi_1, \phi_2 \colon \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and odd. Then

$$\begin{aligned}&\|\phi_1(\mathbf{W}\phi_2(\mathbf{z}) + \mathbf{b}) - \phi_1(\mathbf{W}'\phi_2(\mathbf{z}') + \mathbf{b}')\|_2 \\ &\quad \leq \|\mathbf{z}\|_2\|\mathbf{W} - \mathbf{W}'\|_{\mathrm{F}} + \|\mathbf{W}'\|_2\|\mathbf{z} - \mathbf{z}'\|_2 + \|\mathbf{b} - \mathbf{b}'\|_2. \end{aligned} \tag{53}$$

Proof. When applied element-wise, ϕ_1 and ϕ_2 are also 1-Lipschitz as functions $(\mathbb{R}^n, \|\bullet\|_2) \to (\mathbb{R}^n, \|\bullet\|_2)$. The result then follows from an application of 1-Lipschitzness and the triangle inequality:

$$\begin{aligned}&\|\phi_1(\mathbf{W}\phi_2(\mathbf{z}) + \mathbf{b}) - \phi_1(\mathbf{W}'\phi_2(\mathbf{z}') + \mathbf{b}')\|_2 \\ &\quad \leq \|\mathbf{W}\phi_2(\mathbf{z}) + \mathbf{b} - (\mathbf{W}'\phi_2(\mathbf{z}') + \mathbf{b}')\|_2 && (54)\\ &\quad = \|(\mathbf{W} - \mathbf{W}')\phi_2(\mathbf{z}) + \mathbf{W}'(\phi_2(\mathbf{z}) - \phi_2(\mathbf{z}')) + (\mathbf{b} - \mathbf{b}')\|_2 && (55)\\ &\quad \leq \|\mathbf{W} - \mathbf{W}'\|_2\|\phi_2(\mathbf{z})\|_2 + \|\mathbf{W}'\|_2\|\phi_2(\mathbf{z}) - \phi_2(\mathbf{z}')\|_2 + \|\mathbf{b} - \mathbf{b}'\|_2 && (56)\\ &\quad \leq \|\mathbf{W} - \mathbf{W}'\|_{\mathrm{F}}\|\mathbf{z}\|_2 + \|\mathbf{W}'\|_2\|\mathbf{z} - \mathbf{z}'\|_2 + \|\mathbf{b} - \mathbf{b}'\|_2 && (57)\end{aligned}$$

where in the last inequality we use that since ϕ_2 is odd $\phi_2(\mathbf{0}) = \mathbf{0}$ so $\|\phi_2(\mathbf{z})\|_2 = \|\phi_2(\mathbf{z}) - \phi_2(\mathbf{0})\|_2 \leq \|\mathbf{z}\|_2$. □

Lemma 15. Let $\phi_1, \phi_2 \colon \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and odd. Let the triple $(\mathbf{W}, \mathbf{z}, \mathbf{b})$ be (possibly dependent) random variables such that

$$(\mathbf{W} - \mathbb{E}[\mathbf{W}], \mathbf{z}, \mathbf{b} - \mathbb{E}[\mathbf{b}]) \overset{\mathrm{d}}{=} (-(\mathbf{W} - \mathbb{E}[\mathbf{W}]), \mathbf{z}, -(\mathbf{b} - \mathbb{E}[\mathbf{b}])) \tag{58}$$

Then

$$\|\mathbb{E}[\phi_1(\mathbf{W}\phi_2(\mathbf{z}) + \mathbf{b})]\|_2 \leq \|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}\mathbb{E}[\|\mathbf{z}\|_2] + \mathbb{E}[\|\mathbf{W} - \mathbb{E}[\mathbf{W}]\|_2]\|\mathbb{E}[\mathbf{z}]\|_2 + \|\mathbb{E}[\mathbf{b}]\|_2 \tag{59}$$

Proof. Consider $\mathbf{W}' = \mathbf{W} - \mathbb{E}[\mathbf{W}]$, $\mathbf{z}' = \mathbf{z} - \mathbb{E}[\mathbf{z}]$, and $\mathbf{b}' = \mathbf{b} - \mathbb{E}[\mathbf{b}]$. By assumption, $(\mathbf{W}', \mathbf{z}', \mathbf{b}') \overset{\mathrm{d}}{=} (-\mathbf{W}', \mathbf{z}', -\mathbf{b}')$. Therefore, using that ϕ_1 is odd

$$\mathbb{E}[\phi_1(\mathbf{W}'\phi_2(\mathbf{z}') + \mathbf{b}')] = \mathbb{E}[\phi_1(-\mathbf{W}'\phi_2(\mathbf{z}') - \mathbf{b}')] = -\mathbb{E}[\phi_1(\mathbf{W}'\phi_2(\mathbf{z}') + \mathbf{b}')], \tag{60}$$

which means that $\mathbb{E}[\phi(\mathbf{W}'\phi(\mathbf{z}')+\mathbf{b}')]=\mathbf{0}$. We now apply Lemma 14:

$$\begin{aligned}
&\|\mathbb{E}[\phi_1(\mathbf{W}\phi_2(\mathbf{z})+\mathbf{b})]\|_2 \\
&\quad= \|\mathbb{E}[\phi_1(\mathbf{W}\phi_2(\mathbf{z})+\mathbf{b})-\phi_1(\mathbf{W}'\phi_2(\mathbf{z}')+\mathbf{b}')]\|_2 && (61)\\
&\quad\le \mathbb{E}[\|\phi_1(\mathbf{W}\phi_2(\mathbf{z})+\mathbf{b})-\phi_1(\mathbf{W}'\phi_2(\mathbf{z}')+\mathbf{b}')\|_2] && (62)\\
&\quad\le \mathbb{E}[\|\mathbf{W}-\mathbf{W}'\|_{\mathrm{F}}\|\mathbf{z}\|_2]+\mathbb{E}[\|\mathbf{W}'\|_2\|\mathbf{z}-\mathbf{z}'\|_2]+\mathbb{E}[\|\mathbf{b}-\mathbf{b}'\|_2] && (63)\\
&\quad\le \|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}\mathbb{E}[\|\mathbf{z}\|_2]+\mathbb{E}[\|\mathbf{W}-\mathbb{E}[\mathbf{W}]\|_2]\|\mathbb{E}[\mathbf{z}]\|_2+\|\mathbb{E}[\mathbf{b}]\|_2, && (64)
\end{aligned}$$

which proves the result. □

Remark 16. The symmetry condition for $(\mathbf{W},\mathbf{z},\mathbf{b})$ is satisfied if $\mathbf{W}$, $\mathbf{z}$, and $\mathbf{b}$ are independent and $\mathbf{W}$ and $\mathbf{b}$, not $\mathbf{z}$, are symmetric around their means. For such $(\mathbf{W},\mathbf{z},\mathbf{b})$, the symmetry condition is also satisfied for the triple $(\mathbf{W},\mathbf{z},\alpha\mathbf{W}\mathbf{1}+\mathbf{b})$ where $\mathbf{1}$ is the vector of all ones and $\alpha\in\mathbb{R}$. In that case, a similar conclusion holds:

$$\begin{aligned}
&\|\mathbb{E}[\phi_1(\mathbf{W}(\phi_2+\alpha)(\mathbf{z})+\mathbf{b})]\|_2 && (65)\\
&\quad= \|\mathbb{E}[\phi_1(\mathbf{W}\phi_2(\mathbf{z})+\alpha\mathbf{W}\mathbf{1}+\mathbf{b})]\|_2 && (66)\\
&\quad\le \|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}\mathbb{E}[\|\mathbf{z}\|_2]+\mathbb{E}[\|\mathbf{W}-\mathbb{E}[\mathbf{W}]\|_2]\|\mathbb{E}[\mathbf{z}]\|_2+\|\mathbb{E}[\alpha\mathbf{W}\mathbf{1}+\mathbf{b}]\|_2 && (67)\\
&\quad\le \|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}\mathbb{E}[\|\mathbf{z}\|_2]+\mathbb{E}[\|\mathbf{W}-\mathbb{E}[\mathbf{W}]\|_2]\|\mathbb{E}[\mathbf{z}]\|_2+|\alpha|\|\mathbb{E}[\mathbf{W}]\|_2\|\mathbf{1}\|_2+\|\mathbb{E}[\mathbf{b}]\|_2 && (68)\\
&\quad= (\mathbb{E}[\|\mathbf{z}\|_2+|\alpha|\sqrt{M})\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}+\mathbb{E}[\|\mathbf{W}-\mathbb{E}[\mathbf{W}]\|_2]\|\mathbb{E}[\mathbf{z}]\|_2+\|\mathbb{E}[\mathbf{b}]\|_2. && (69)
\end{aligned}$$

Lemma 17. Let $\mathbf{A}\in\mathbb{R}^{I\times J}$ be a zero-mean random matrix with independent Gaussian entries with variances bounded by σ^2. Then

$$\mathbb{E}[\|\mathbf{A}\|_2]\le 2\sigma\sqrt{I\vee J}. \tag{70}$$

Moreover, if the variances of the entries of $\mathbf{A}$ are all equal to one, then

$$\mathbb{P}(\|\mathbf{A}\|_2\ge 2\sqrt{I\vee J}+\delta)\le 2\exp(-\tfrac{1}{2}\delta^2). \tag{71}$$

Proof. We slightly generalise Exercise 5.14 from Wainwright (2019). To begin with, rewrite the operator norm as

$$\|\mathbf{A}\|_2=\sup_{(\mathbf{u},\mathbf{v})\in\mathbb{S}^{I-1}\times\mathbb{S}^{J-1}}\langle\mathbf{u},\mathbf{A}\mathbf{v}\rangle. \tag{72}$$

Define the zero-mean Gaussian process $Z_{\mathbf{u},\mathbf{v}}=\langle\mathbf{u},\mathbf{A}\mathbf{v}\rangle$ indexed on $\mathbb{S}^{I-1}\times\mathbb{S}^{J-1}$ and define $\mathbf{S}$ by $S_{i,j}=\mathbb{V}[A_{i,j}]\le\sigma^2$. Note that, by independence of the entries of $\mathbf{A}$,

$$\begin{aligned}
\mathbb{E}[(Z_{\mathbf{u},\mathbf{v}}-Z_{\mathbf{w},\mathbf{x}})^2] &= \mathbb{E}[Z^2_{\mathbf{u},\mathbf{v}}-2Z_{\mathbf{u},\mathbf{v}}Z_{\mathbf{w},\mathbf{x}}+Z^2_{\mathbf{w},\mathbf{x}}] && (73)\\
&= \textstyle\sum_{i=1,j=1}^{I,J}S_{ij}u_i^2v_j^2-2\sum_{i=1,j=1}^{I,J}S_{ij}u_iv_jw_ix_j+\sum_{i=1,j=1}^{I,J}S_{ij}w_i^2x_j^2 && (74)\\
&= \textstyle\sum_{i=1,j=1}^{I,J}S_{ij}(u_iv_j-w_ix_j)^2 && (75)\\
&\le \sigma^2\textstyle\sum_{i=1,j=1}^{I,J}(u_iv_j-w_ix_j)^2 && (76)\\
&= \sigma^2\|\mathbf{u}\mathbf{v}^{\mathsf{T}}-\mathbf{w}\mathbf{x}^{\mathsf{T}}\|_{\mathrm{F}}^2. && (77)
\end{aligned}$$

Also consider the zero-mean Gaussian process $Y_{\mathbf{u},\mathbf{v}}=\sigma\langle\mathbf{u},\boldsymbol{\varepsilon}_1\rangle+\sigma\langle\mathbf{v},\boldsymbol{\varepsilon}_2\rangle$ again indexed on $\mathbb{S}^{I-1}\times\mathbb{S}^{J-1}$, where $\boldsymbol{\varepsilon}_1\in\mathbb{R}^I$ and $\boldsymbol{\varepsilon}_2\in\mathbb{R}^J$ are standard Gaussian vectors. Then

$$\begin{aligned}
\mathbb{E}[(Y_{\mathbf{u},\mathbf{v}}-Y_{\mathbf{w},\mathbf{x}})^2] &= \sigma^2\mathbb{E}[(\langle\mathbf{u}-\mathbf{w},\boldsymbol{\varepsilon}_1\rangle+\langle\mathbf{v}-\mathbf{x},\boldsymbol{\varepsilon}_2\rangle)^2] && (78)\\
&= \sigma^2\mathbb{E}[\langle\mathbf{u}-\mathbf{w},\boldsymbol{\varepsilon}_1\rangle^2+\langle\mathbf{v}-\mathbf{x},\boldsymbol{\varepsilon}_2\rangle^2] && (79)\\
&= \sigma^2(\|\mathbf{u}-\mathbf{w}\|_2^2+\|\mathbf{v}-\mathbf{x}\|_2^2). && (80)
\end{aligned}$$

Using that $\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=\|\mathbf{x}\|_2=\|\mathbf{w}\|_2=1$, careful algebra (see, e.g., page 164 from Wainwright, 2019) shows that $\|\mathbf{u}\mathbf{v}^{\mathsf{T}}-\mathbf{w}\mathbf{x}^{\mathsf{T}}\|_{\mathrm{F}}^2\le\|\mathbf{u}-\mathbf{w}\|_2^2+\|\mathbf{v}-\mathbf{x}\|_2^2$:

$$\begin{aligned}
(u_iv_j-w_ix_j)^2 &= (u_iv_j-w_iv_j+w_iv_j-w_ix_j)^2 && (81)\\
&= (u_i-w_i)^2v_j^2+w_i^2(v_j-x_j)^2+2(u_iv_j-w_iv_j)(w_iv_j-w_ix_j) && (82)\\
&= (u_i-w_i)^2v_j^2+w_i^2(v_j-x_j)^2+2(u_iw_i-w_i^2)(v_j^2-v_jx_j). && (83)
\end{aligned}$$

Therefore, summing over $i \in [I]$ and $j \in [J]$ and using that $\|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1$,

$$\|\mathbf{u}\mathbf{v}^\mathsf{T} - \mathbf{w}\mathbf{x}^\mathsf{T}\|_\mathrm{F}^2 = \|\mathbf{u} - \mathbf{w}\|_2^2 + \|\mathbf{v} - \mathbf{x}\|_2^2 + 2(\langle \mathbf{u}, \mathbf{w}\rangle - 1)(1 - \langle \mathbf{v}, \mathbf{x}\rangle) \leq \|\mathbf{u} - \mathbf{w}\|_2^2 + \|\mathbf{v} - \mathbf{x}\|_2^2 \tag{84}$$

where the inequality follows from additionally using that $\|\mathbf{u}\|_2 = \|\mathbf{x}\|_2 = 1$. Using this result, we find

$$\mathbb{E}[(Z_{\mathbf{u},\mathbf{v}} - Z_{\mathbf{w},\mathbf{x}})^2] \leq \mathbb{E}[(Y_{\mathbf{u},\mathbf{v}} - Y_{\mathbf{w},\mathbf{x}})^2]. \tag{85}$$

Hence, denoting $Z^* = \sup_{(\mathbf{u},\mathbf{v}) \in \mathbb{S}^{I-1} \times \mathbb{S}^{J-1}} Z_{\mathbf{u},\mathbf{v}}$ and $Y^* = \sup_{(\mathbf{u},\mathbf{v}) \in \mathbb{S}^{I-1} \times \mathbb{S}^{J-1}} Y_{\mathbf{u},\mathbf{v}}$, by the Sudakov–Fernique comparison theorem (Theorem 5.27, Wainwright, 2019), we conclude that

$$\mathbb{E}[\|\mathbf{A}\|_2] = \mathbb{E}[Z^*] \leq \mathbb{E}[Y^*] \overset{\text{(i)}}{=} \sigma\, \mathbb{E}[\|\boldsymbol{\varepsilon}_1\|_2 + \|\boldsymbol{\varepsilon}_2\|_2] \overset{\text{(ii)}}{\leq} \sigma(\sqrt{I} + \sqrt{J}) \tag{86}$$

where (i) follows from that the suprema are achieved by $\boldsymbol{\varepsilon}_1/\|\boldsymbol{\varepsilon}_1\|_2$ and $\boldsymbol{\varepsilon}_2/\|\boldsymbol{\varepsilon}_2\|_2$ for $\mathbf{u}$ and $\mathbf{v}$, respectively, and (ii) follows from Jensen's inequality applied to the square root.

For the second statement, assume that all entries of $\mathbf{A}$ have variance one, so $\sigma^2 = 1$. We apply concentration of Gaussian suprema (e.g., Exercise 5.10 by Wainwright, 2019) to $Z_{\mathbf{u},\mathbf{v}}$:

$$\mathbb{P}(|Z^* - \mathbb{E}[Z^*]| \geq \delta) \leq 2\exp(-\tfrac{1}{2v}\delta^2) \tag{87}$$

where

$$v = \sup_{(\mathbf{u},\mathbf{v}) \in \mathbb{S}^{I-1} \times \mathbb{S}^{J-1}} \mathbb{V}[Z_{\mathbf{u},\mathbf{v}}] = \sup_{(\mathbf{u},\mathbf{v}) \in \mathbb{S}^{I-1} \times \mathbb{S}^{J-1}} \|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 = 1. \tag{88}$$

Therefore, using the bound on $\mathbb{E}[Z^*]$ from the first statement,

$$\mathbb{P}(\|\mathbf{A}\|_2 \geq 2\sqrt{I \vee J} + \delta) \leq 2\exp(-\tfrac{1}{2}\delta^2), \tag{89}$$

which concludes the proof. □

Having established the preliminaries, we can now bound the norm of the expectation of one layer in terms of the previous layer.

Lemma 18. Let $\phi_1, \phi_2 \colon \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and odd, and let $\alpha \in \mathbb{R}$. Let $(\mathbf{W}, \mathbf{z}, \mathbf{b})$ be independent random variables with $\mathbf{W}$ and $\mathbf{b}$, not $\mathbf{z}$, symmetric around their means. Moreover, assume that $\mathbf{W}$ has independent sub-Gaussian entries with sub-Gaussian parameters bounded by σ. Then there exists a universal constant $c > 0$ such that

$$\begin{aligned}&\|\mathbb{E}[\phi_1(\tfrac{1}{\sqrt{M}}\mathbf{W}(\phi_2 + \alpha)(\mathbf{z}) + \mathbf{b})]\|_2 \\ &\quad \leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}\|_2] + |\alpha|)\|\mathbb{E}[\mathbf{W}]\|_\mathrm{F} + c\sigma\|\mathbb{E}[\mathbf{z}]\|_2 + \|\mathbb{E}[\mathbf{b}]\|_2.\end{aligned} \tag{90}$$

with c the same constant as in Lemma 17.

Proof. By Lemma 17, there exists a universal constant $c > 0$ such that

$$\mathbb{E}[\|\mathbf{W} - \mathbb{E}[\mathbf{W}]\|_2] \leq c\sigma\sqrt{M} \implies \mathbb{E}[\|\tfrac{1}{\sqrt{M}}\mathbf{W} - \mathbb{E}[\tfrac{1}{\sqrt{M}}\mathbf{W}]\|_2] \leq c\sigma. \tag{91}$$

The result then follows from Remark 16 with $\frac{1}{\sqrt{M}}\mathbf{W}$ instead of $\mathbf{W}$. □

We will also need upper bounds on the expected value of the norm of a hidden layer in terms of the previous layer.

Lemma 19. Assume the conditions of Lemma 18. Then

$$\begin{aligned}&\mathbb{E}[\|\tfrac{1}{\sqrt{M}}\mathbf{W}(\phi_2 + \alpha)(\mathbf{z}) + \mathbf{b}\|_2] \\ &\quad \leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}\|_2] + |\alpha|)(\|\mathbb{E}[\mathbf{W}]\|_2 + c\sigma\sqrt{M}) + \|\mathbb{E}[\mathbf{b}]\|_2 + c\sigma\sqrt{M}.\end{aligned} \tag{92}$$

Proof. Use the triangle inequality and recall that $\|\phi_2(\mathbf{z})\|_2 \leq \|\mathbf{z}\|_2$ (proof of Lemma 14):

$$\begin{aligned}
&\mathbb{E}[\|\tfrac{1}{\sqrt{M}}\mathbf{W}\phi_2(\mathbf{z}) + \tfrac{\alpha}{\sqrt{M}}\mathbf{W}\mathbf{1} + \mathbf{b}\|_2] \\
&\quad\leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}\|_2] + |\alpha|)\mathbb{E}[\|\mathbf{W}\|_2] + \mathbb{E}[\|\mathbf{b}\|_2] &(93)\\
&\quad\leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}\|_2] + |\alpha|)(\|\mathbb{E}[\mathbf{W}]\|_2 + c\sigma\sqrt{M}) + \|\mathbb{E}[\mathbf{b}]\|_2 + c\sigma\sqrt{M} &(94)
\end{aligned}$$

where in the second inequality we use the triangle inequality in combination with Lemma 17. □

We now turn to the proof of Lemma 13, which relies of an argument with the following form.

Proposition 20. Let $a, b \in \mathbb{R}$ and $\{c_\ell\}_{\ell=1}^L$, with $c_\ell \in \mathbb{R}$. Suppose, $c_\ell \leq ac_{\ell-1} + b$ with $a, b \geq 0$ for $h > 1$. Then for $2 \leq \ell \leq L$

$$c_\ell \leq a^{\ell-1}c_1 + (1+a)^{\ell-2}b. \tag{95}$$

Proof. The proof is a standard induction. In the base case $h = 2$, we have to prove,

$$c_2 \leq ac_1 + b, \tag{96}$$

which holds because this is simply our assumption on the c_ℓ with $\ell = 2$. For the inductive step, we now assume that $c_\ell \leq a^{\ell-1}c_1 + (1+a)^{\ell-2}b$. Under this assumption, we have,

$$c_{\ell+1} \leq ac_\ell + b \leq a^\ell c_1 + a(1+a)^{\ell-2}b + b = a^\ell c_1 + (1+a)^{\ell-1}b. \tag{97}$$

□

Lemma 21. Define $K = \sqrt{2\,\mathrm{KL}(Q,P)}$. Then for $1 \leq \ell \leq L$

$$\tfrac{1}{\sqrt{M}}\mathbb{E}_Q[\|\mathbf{z}_\ell\|_2] \leq (2 + |\alpha| + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(2+2c)^\ell (K \vee 1)^\ell, \tag{98}$$

where c is the same absolute constant from Lemma 15.

Proof. By Lemma 9, all mean parameters are bounded by K and $\sigma_{\max} \leq 2(K \vee 1)$. Applying, Lemma 19 using these estimates, ,

$$\begin{aligned}
&\tfrac{1}{\sqrt{M}}\mathbb{E}_Q[\|\mathbf{z}_\ell\|_2] \leq (2c(K\vee 1) + \tfrac{1}{\sqrt{M}}K)\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}_{\ell-1}\|_2] + (1+|\alpha|)(2c(K\vee 1) + \tfrac{1}{\sqrt{M}}K), &(99)\\
&\tfrac{1}{\sqrt{M}}\mathbb{E}_Q[\|\mathbf{z}_1\|_2] \leq (1 + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(2c(K\vee 1) + \tfrac{1}{\sqrt{M}}K). &(100)
\end{aligned}$$

Bound $2c(K \vee 1) + \frac{1}{\sqrt{M}}K \leq (1+2c)(K \vee 1)$ and apply Proposition 20:

$$\begin{aligned}
\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}_\ell\|_2] &\leq (1 + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(1+2c)^\ell (K\vee 1)^\ell + (1+|\alpha|)(2+2c)^{\ell-1}(K\vee 1)^{\ell-1}, &(101)\\
&\leq (2 + |\alpha| + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(2+2c)^\ell (K\vee 1)^\ell. &\square
\end{aligned}$$

Lemma 22. Define $K = \sqrt{2\,\mathrm{KL}(Q,P)}$. Then for $1 \leq \ell \leq L$

$$\|\mathbb{E}_Q[\mathbf{z}_\ell]\|_2 \leq 2\ell(2 + |\alpha| + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(2+2c)^{\ell-1}K(K\vee 1)^{\ell-1}, \tag{102}$$

where c is the same absolute constant from Lemma 15

Proof. The proof proceeds by induction on ℓ.

Base Case For the first layer, $\ell = 1$, by linearity of expectation, triangle inequality and the definition of the spectral norm,

$$\|\mathbb{E}[\mathbf{z}_1]\|_2 \leq \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbb{E}[\mathbf{W}_1]\|_2\|\mathbf{x}\|_2 + \|\mathbb{E}[\mathbf{b}_1]\|_2 \leq \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbb{E}[\mathbf{W}_1]\|_{\mathrm{F}}\|\mathbf{x}\|_2 + \|\mathbb{E}[\mathbf{b}_1]\|_2. \tag{103}$$

We then apply Lemma 9 (i.), to conclude $\|\mathbb{E}[\mathbf{W}_1]\|, \|\mathbb{E}[\mathbf{b}_1]\|_2 \leq K$,

$$\|\mathbb{E}[\mathbf{z}_1]\|_2 \leq (1 + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)K \leq 2(2 + |\alpha| + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)K. \tag{104}$$

Inductive step We take the inductive hypothesis,

$$\|\mathbb{E}_Q[\mathbf{z}_\ell]\|_2 \leq 2\ell(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{\ell-1}K(K \vee 1)^{\ell-1}. \tag{105}$$

By an application of Lemma 18,

$$\|\mathbb{E}[\mathbf{z}_{\ell+1}]\|_2 \leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}_\ell\|_2] + |\alpha|)\|\mathbb{E}[\mathbf{W}_{\ell+1}]\|_\mathrm{F} + c\sigma_{\max}\|\mathbb{E}[\mathbf{z}_\ell]\|_2 + \|\mathbb{E}[\mathbf{b}_{\ell+1}]\|_2 \tag{106}$$

Applying Lemma 9 (i., iii.),

$$\|\mathbb{E}[\mathbf{z}_{\ell+1}]\|_2 \leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}_\ell\|_2] + |\alpha| + 1)K + 2c(K \vee 1)\|\mathbb{E}[\mathbf{z}_\ell]\|_2 \tag{107}$$

Using the inductive hypothesis,

$$2c(K \vee 1)\|\mathbb{E}_Q[\mathbf{z}_\ell]\|_2] \leq 4c\ell(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{\ell-1}K(K \vee 1)^{\ell} \tag{108}$$

$$\leq 2\ell(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{\ell}K(K \vee 1)^{\ell}. \tag{109}$$

We then make use of Lemma 21,

$$K(1 + |\alpha| + \tfrac{1}{\sqrt{M}}\mathbb{E}_Q[\|\mathbf{z}_\ell\|_2]) \leq K\left(1 + |\alpha| + (2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{\ell}(K \vee 1)^{\ell}\right) \tag{110}$$

$$\leq 2(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{\ell}K(K \vee 1)^{\ell} \tag{111}$$

Hence,

$$\|\mathbb{E}[\mathbf{z}_{\ell+1}]\|_2 \leq (2\ell+2)(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{\ell}K(K \vee 1)^{\ell}. \qquad \square$$

We are now ready to prove Lemma 13.

Proof of Lemma 13. The proof is essentially identical to the inductive step in Lemma 22.

To begin with, we apply Lemma 18 to the deep architecture. For the last hidden layer,

$$\|\mathbb{E}[\phi_\mathrm{o}(\mathbf{z}_L)]\|_2 \leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}_{L-1}\|_2] + |\alpha|)\|\mathbb{E}[\mathbf{W}_L]\|_\mathrm{F} + c\sigma_{\max}\|\mathbb{E}[\mathbf{z}_{L-1}]\|_2 + \|\mathbb{E}[\mathbf{b}_L]\|_2. \tag{112}$$

We apply Lemma 9 (i., iii.), yielding,

$$\|\mathbb{E}[\phi_\mathrm{o}(\mathbf{z}_L)]\|_2 \leq (\tfrac{1}{\sqrt{M}}\mathbb{E}[\|\mathbf{z}_{L-1}\|_2] + |\alpha|)K + 2c(K \vee 1)\|\mathbb{E}[\mathbf{z}_{L-1}]\|_2 + K. \tag{113}$$

We then make use of Lemma 21,

$$K(1 + |\alpha| + \tfrac{1}{\sqrt{M}}\mathbb{E}_Q[\|\mathbf{z}_{L-1}\|_2]) \leq K\left(1 + |\alpha| + (2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{L-1}(K \vee 1)^{L-1}\right) \tag{114}$$

$$\leq 2(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{L-1}K(K \vee 1)^{L-1}. \tag{115}$$

From Lemma 22,

$$2c(K \vee 1)\|\mathbb{E}_Q[\mathbf{z}_{L-1}]\|_2] \leq 4c(L-1)(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{L-2}K(K \vee 1)^{L-1} \tag{116}$$

$$\leq 2(L-1)(2 + |\alpha| + \tfrac{1}{\sqrt{D_\mathrm{i}}}\|\mathbf{x}\|_2)(2+2c)^{L-1}K(K \vee 1)^{L-1}. \tag{117}$$

Combining Equations (113), (115) and (117) gives the result. $\square$

E Proof of Convergence of the Marginal Variance for Deep Networks

We now turn to the problem of bounding the marginal variance of the predictive distribution for Q. We work with the uncentered second moment, as in combination with the previous section this implies the marginal variance of P and Q agree. The main result of this section is as follows:

Theorem 23 (Convergence of second moment prediction)**.** Let Q be a mean-field variational posterior and $P = \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote the prior over a neural network with L hidden layers and M neurons per hidden layer. Suppose $\phi_{\mathrm{e}} = \alpha$ for some $\alpha \in \mathbb{R}$ and $\phi\colon \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz. Let $\mathbf{x} \in \mathbb{R}^{D_{\mathrm{i}}}$ and let $\widetilde{\mathbf{f}}_\theta(\mathbf{x}) = \mathbf{f}_\theta(\mathbf{x}) - \frac{\alpha}{\sqrt{M}}\mathbf{W}_{L+1}\mathbf{1} - \mathbf{b}_{L+1}$ denote the network output excluding final bias and even part of the final activation. Then

$$\|\mathbb{E}_Q[\tilde{\mathbf{f}}^2_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}_P[\tilde{\mathbf{f}}^2_{\boldsymbol{\theta}}(\mathbf{x})]\|_\infty \leq c_1 L^{1/2} \rho^L_{\alpha,M} \frac{\alpha^2 + 1 + \frac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2}{\sqrt{M}} \sqrt{\mathrm{KL}(Q,P)} (2\,\mathrm{KL}(Q,P) \vee 1)^{L+\frac{1}{2}}, \tag{118}$$

where $c_1 = 16 + 25\sqrt{2} \in (51, 52)$ and where $\rho_{\alpha,M} \in (17, 97)$ is defined in Lemma 35.

The proof of Theorem 23 will proceed by splitting the variance into two terms. The first term, which we name the *diagonal*, arises from the product of the variance of the weights with the second moment of each activation. We show that under the optimal variational posterior, this term is close to the same term under the prior. Precisely, we have the following lemma:

Lemma 24 (Diagonal Variance terms)**.** Define $\boldsymbol{\Sigma}_Q = \mathbb{E}_Q[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^{\mathsf{T}}] - \mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^{\mathsf{T}}$, which is a diagonal matrix with the variances of $\mathbf{w}_{L+1}$ as entries. Let

$$D^i_P(\mathbf{x}) = \tfrac{1}{M}\operatorname{tr}(\mathbb{E}_P[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^{\mathsf{T}}]\mathbb{E}_P[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}]) \tag{119}$$

$$D^i_Q(\mathbf{x}) = \tfrac{1}{M}\operatorname{tr}(\boldsymbol{\Sigma}_Q\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}]) \tag{120}$$

Then,

$$|D^i_Q(\mathbf{x}) - D^i_P(\mathbf{x})| \leq \left(16 + \sqrt{2}\right) L^{1/2} \rho^L_{\alpha,M} \frac{\alpha^2 + 1 \vee \frac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2}{\sqrt{M}} \sqrt{\mathrm{KL}(Q,P)} (2\,\mathrm{KL}(Q,P) \vee 1)^{L+\frac{1}{2}} \tag{121}$$

where $\rho_{\alpha,M} \in (17, 97)$ is defined in Lemma 35.

The proof of Lemma 24 will be the main topic of Appendix E.2.

The second term, which we term the *off-diagonal* arises from the product of the mean of the weights with the second moment of each activation. Under the prior, this term vanishes. Hence, we show that this term is small for the optimal approximate posterior. This leads to the following lemma,

Lemma 25 (Off-Diagonal Variance terms)**.** Define $O^i_Q(\mathbf{x}) = \frac{1}{M}\mathbb{E}_Q[\mathbf{w}_{L+1}]^{\mathsf{T}}\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}]\mathbb{E}_Q[\mathbf{w}_{L+1}]$. Then,

$$|O^i_Q(\mathbf{x})| \leq 48\gamma^L_\alpha \frac{\alpha^2 + 1 \vee \frac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2}{\sqrt{M}} \mathrm{KL}(Q,P)(2\,\mathrm{KL}(Q,P) \vee 1)^L, \tag{122}$$

where $\gamma_\alpha = 9 + \sqrt{83} \in (18, 19)$ if $\alpha = 0$ and $\gamma_\alpha = 55$ if $\alpha \neq 0$.

The proof of Lemma 25 will be the main topic of Appendix E.3.

These two lemmas taken together allow us to prove Theorem 23. The entire structure of the proof is depicted in Figure 7.

Proof of Theorem 23. We will repeatedly use cyclic property and linearity of trace. Hence, we briefly recall that for conformable matrix $\mathbf{A}, \mathbf{B}$, we have $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ and $\mathbb{E}[\operatorname{tr}(\mathbf{AB})] = \operatorname{tr}(\mathbb{E}[\mathbf{AB}])$. We also observe that the trace of a 1×1 matrix is simply the corresponding scalar.

It suffices to consider an arbitrary output component. Let $\widetilde{f}(\mathbf{x}) = \widetilde{\mathbf{f}}(\mathbf{x})_i$ denote an arbitrary output component i and let $\mathbf{w}_{L+1}$ denote row i of $\mathbf{W}_{L+1}$, so that $\widetilde{f}(\mathbf{x}) = \frac{1}{M}\mathbf{w}^T_{L+1}\phi_{\mathrm{o}}(\mathbf{z}_L)$.

$$M\widetilde{f}(\mathbf{x})^2 = \mathbf{w}_{L+1}^{\mathsf{T}}\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}\mathbf{w}_{L+1} \tag{123}$$

$$= \operatorname{tr}\left(\mathbf{w}_{L+1}^{\mathsf{T}}\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}\mathbf{w}_{L+1}\right) \tag{124}$$

$$= \operatorname{tr}\left(\mathbf{w}_{L+1}\mathbf{w}_{L+1}^{\mathsf{T}}\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}\right). \tag{125}$$

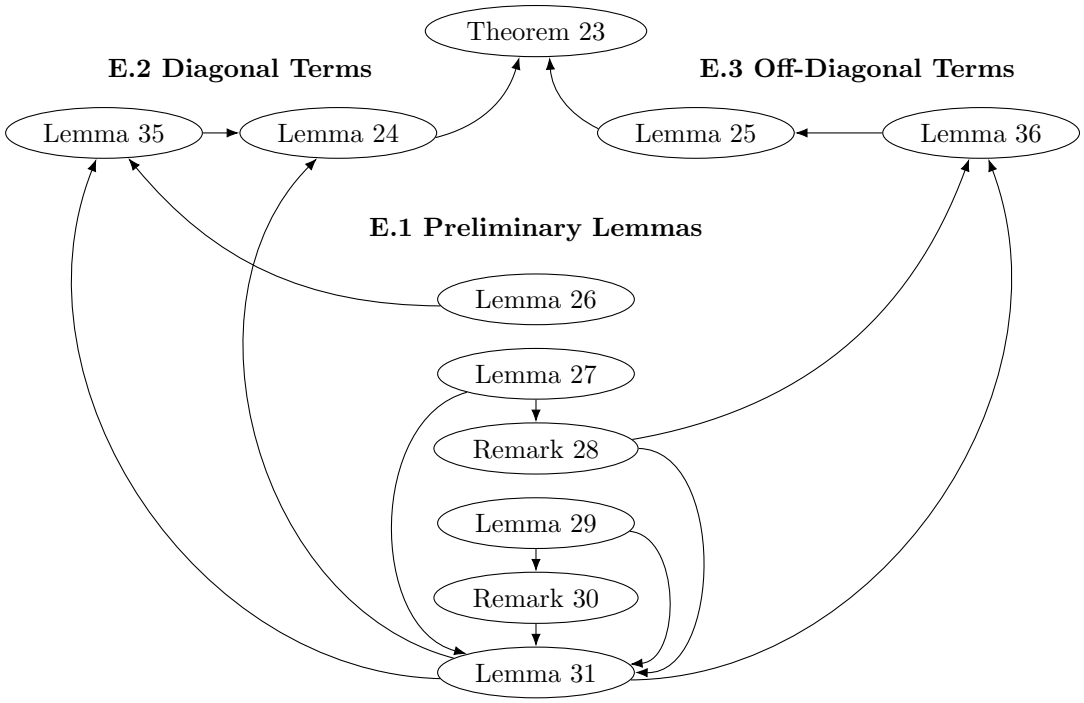

Figure 7: Dependency structure of the results leading to Theorem 23. Lemmas 32, 33, and 34 are proved in this section but only used in Appendix G.

Recall $\mathbf{\Sigma}_Q = \mathbb{E}_Q[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^\mathsf{T}] - \mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}$, which is a diagonal matrix (by indpendence) with the variances of $\mathbf{w}_{L+1}$ as entries. Then, adding and subtracting $\mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}$, and using that $\mathbf{w}_{L+1}$ is independent of $\mathbf{z}_L$ by the independence assumption

$$\mathbb{E}_Q[\widetilde{f}(\mathbf{x})^2] = \tfrac{1}{M}\operatorname{tr}((\mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T} + \mathbf{\Sigma}_Q)\mathbb{E}_Q[\phi_\mathrm{o}(\mathbf{z}_L)\phi_\mathrm{o}(\mathbf{z}_L)^\mathsf{T}]) \tag{126}$$

$$= \tfrac{1}{M}\operatorname{tr}(\mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbb{E}_Q[\phi_\mathrm{o}(\mathbf{z}_L)\phi_\mathrm{o}(\mathbf{z}_L)^\mathsf{T}]) + \tfrac{1}{M}\operatorname{tr}(\mathbf{\Sigma}_Q\mathbb{E}_Q[\phi_\mathrm{o}(\mathbf{z}_L)\phi_\mathrm{o}(\mathbf{z}_L)^\mathsf{T}]) \tag{127}$$

$$= \tfrac{1}{M}\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbb{E}_Q[\phi_\mathrm{o}(\mathbf{z}_L)\phi_\mathrm{o}(\mathbf{z}_L)^\mathsf{T}]\mathbb{E}_Q[\mathbf{w}_{L+1}] + D_Q^i \tag{128}$$

$$= O_Q^i(\mathbf{x}) + D_Q^i(\mathbf{x}). \tag{129}$$

Also, $\mathbb{E}_P[\widetilde{f}(\mathbf{x})^2] = D_P^i(\mathbf{x})$ since $\mathbb{E}_P[\mathbf{w}_{L+1}] = 0$. So,

$$|\mathbb{E}_Q[\widetilde{f}(\mathbf{x})^2] - \mathbb{E}_P[\widetilde{f}(\mathbf{x})^2]| = |O_Q^i(\mathbf{x}) + D_Q^i(\mathbf{x}) - D_P^i(\mathbf{x})| \tag{130}$$

$$\leq |O_Q^i(\mathbf{x})| + |D_Q^i(\mathbf{x}) - D_P^i(\mathbf{x})|. \tag{131}$$

The final result is obtained by plugging in the expressions from Lemmas 24 and 25, noting that $\gamma_\alpha \leq \gamma_{\alpha,M}$. □

E.1 Preliminary Lemmas

In order to prove both Lemma 24 and Lemma 25 we first establish several preliminary results. The first lemma will be useful in bounding the diagonal terms in Lemma 35. We construct the bound $\eta_{I\vee J}$ to be convenient when collecting terms in Lemma 35.

Lemma 26. Let $\mathbf{A} \in \mathbb{R}^{I\times J}$ have independent standard Gaussian entries. Then

$$\mathbb{E}[\|\mathbf{A}\|_2^2] \leq \eta_{I\vee J}(I \vee J), \tag{132}$$

where $\eta_{I\vee J}$ takes the value $\frac{1}{9}(37 + 6\sqrt{2\pi}) \in (5, 6)$ if $I \vee J \geq 36$ and $4(2 + \sqrt{2\pi}) \in (18, 19)$ otherwise.

Proof. We integrate the tail bound from Lemma 17:

$$\mathbb{P}(\|\mathbf{A}\|_2 \geq 2\sqrt{I \vee J} + \delta) \leq 2\exp(-\tfrac{1}{2}\delta^2). \tag{133}$$

To integrate the tail bound, use the layer cake trick:

$$\mathbb{E}[\|\mathbf{A}\|_2^2] = \int_0^\infty P(\|\mathbf{A}\|_2^2 > u)\,\mathrm{d}u \tag{134}$$

$$\leq 4(I \vee J) + \int_0^\infty P(\|\mathbf{A}\|_2^2 > 4(I \vee J) + u)\,\mathrm{d}u. \tag{135}$$

Consider the change of variables defined by

$$4(I \vee J) + u = (2\sqrt{I \vee J} + v)^2. \tag{136}$$

Then $\mathrm{d}u = 2(2\sqrt{I \vee J} + v)\,\mathrm{d}v$, so

$$\mathbb{E}[\|\mathbf{A}\|_2^2] \leq 4(I \vee J) + 2\int_0^\infty P(\|\mathbf{A}\|_2 > 2\sqrt{I \vee J} + v)(2\sqrt{I \vee J} + v)\,\mathrm{d}v \tag{137}$$

$$\leq 4(I \vee J) + 4\int_0^\infty e^{-\frac{1}{2}v^2}(2\sqrt{I \vee J} + v)\,\mathrm{d}v \tag{138}$$

$$= 4(I \vee J) + 4\sqrt{2\pi}\sqrt{I \vee J} + 4. \tag{139}$$

The claimed bound then follows by checking cases. □

Lemma 27. Let $\phi_1, \phi_2 \colon \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and odd. Let the triple $(\mathbf{W}, \mathbf{z}_2, \mathbf{b}) \in \mathbb{R}^{M_1 \times M_2} \times \mathbb{R}^{M_2} \times \mathbb{R}^{M_1}$ be (possibly dependent) random variables such that, for all Rademacher vectors $\boldsymbol{\varepsilon} \in \{-1, 1\}^{M_1}$,

$$(\mathbf{W} - \mathbb{E}[\mathbf{W}], \mathbf{z}_2, \mathbf{b} - \mathbb{E}[\mathbf{b}]) \overset{\mathrm{d}}{=} (\mathrm{diag}(\boldsymbol{\varepsilon})(\mathbf{W} - \mathbb{E}[\mathbf{W}]), \mathbf{z}_2, \boldsymbol{\varepsilon} \circ (\mathbf{b} - \mathbb{E}[\mathbf{b}])). \tag{140}$$

Consider $\mathbf{z}_1 = \frac{1}{\sqrt{M_2}}\mathbf{W}\phi_2(\mathbf{z}_2) + \mathbf{b}$. Then

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq 4\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty + \|\mathbb{E}[\mathbf{W}]\|_2^2\,\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty + \|\mathbb{E}[\mathbf{b}]\|_2^2) \tag{141}$$

$$\leq 4\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty + K^2\,(\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty \vee 1)) \tag{142}$$

provided that $M_1 \geq 3$. If $M_1 < 3$, then the inequality holds with the slightly worse constant 6.

Proof. Set

$$\mathbf{z}_1' = \mathbf{z}_1 - \tfrac{1}{\sqrt{M_2}}\mathbb{E}[\mathbf{W}]\phi_2(\mathbf{z}_2) - \mathbb{E}[\mathbf{b}] = \tfrac{1}{\sqrt{M_2}}(\mathbf{W} - \mathbb{E}[\mathbf{W}])\phi_2(\mathbf{z}_2) + (\mathbf{b} - \mathbb{E}[\mathbf{b}]). \tag{143}$$

Consider

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq \|\mathbb{E}[(\phi_1(\mathbf{z}_1) - \phi_1(\mathbf{z}_1'))\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 + \|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \tag{144}$$

$$\overset{\text{(i)}}{\leq} \mathbb{E}[\|(\phi_1(\mathbf{z}_1) - \phi_1(\mathbf{z}_1'))\phi_1(\mathbf{z}_1)^\mathsf{T}\|_2] + \|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \tag{145}$$

$$\overset{\text{(ii)}}{=} \mathbb{E}[\|\phi_1(\mathbf{z}_1) - \phi_1(\mathbf{z}_1')\|_2\|\phi_1(\mathbf{z}_1)\|_2] + \|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \tag{146}$$

$$\overset{\text{(iii)}}{\leq} \mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_1'\|_2\|\mathbf{z}_1\|_2] + \|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \tag{147}$$

$$\overset{\text{(iv)}}{\leq} (\mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_1'\|_2^2]\,\mathbb{E}[\|\mathbf{z}_1\|_2^2])^{1/2} + \|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \tag{148}$$

where in (i) we use Jensen's Inequality, in (ii) we compute the 2-norm, in (iii) we use 1-Lipschitzness and oddness of ϕ_1, and in (iv) we use Cauchy–Schwarz. In a similar way, we can simplify the last term from above:

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq (\mathbb{E}[\|\mathbf{z}_1'\|_2^2]\,\mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_1'\|_2^2])^{1/2} + \|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1')^\mathsf{T}]\|_2. \tag{149}$$

Therefore,

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq \mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_1'\|_2^2]^{1/2}(\mathbb{E}[\|\mathbf{z}_1\|_2^2]^{1/2} + \mathbb{E}[\|\mathbf{z}_1'\|_2^2]^{1/2}) + \|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1')^\mathsf{T}]\|_2. \tag{150}$$

Unfortunately, the square roots cannot be pushed inside: Jensen's inequality is the other way around. Instead, we bound each of the three terms involving $\mathbf{z}'$ in Equation (150) separately, starting with the first term. Applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ to Equation (143), we have

$$\mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_1'\|_2^2] \leq 2\|\mathbb{E}[\mathbf{W}]\|_2^2 \tfrac{1}{M_2}\mathbb{E}[\|\phi_2(\mathbf{z}_2)\|_2^2] + 2\|\mathbb{E}[\mathbf{b}]\|_2^2 \tag{151}$$

$$\leq 2\|\mathbb{E}[\mathbf{W}]\|_2^2 \tfrac{1}{M_2}\mathbb{E}[\|\mathbf{z}_2\|_2^2] + 2\|\mathbb{E}[\mathbf{b}]\|_2^2 \tag{152}$$

$$\leq 2\|\mathbb{E}[\mathbf{W}]\|_2^2 \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty + 2\|\mathbb{E}[\mathbf{b}]\|_2^2. \tag{153}$$

Above, the second inequality follows from the oddness and 1-Lipschitzness of ϕ_2 and that $\|\mathbf{1}\|_2^2 = M_2$ while the third inequality follows from converting the 2-norm to an ∞-norm:

$$\mathbb{E}[\|\mathbf{z}_2\|_2^2] \leq \mathbb{E}\left[\sum_{i=1}^{M_2} z_{2,i}^2\right] = \sum_{i=1}^{M_2} \mathbb{E}[z_{2,i}^2] \leq M_2 \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty. \tag{154}$$

Similarly,

$$\mathbb{E}[\|\mathbf{z}_1'\|_2^2] \leq 3\mathbb{E}[\|\mathbf{z}_1\|_2^2] + 3\|\mathbb{E}[\mathbf{W}]\|_2^2 \, \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty + 3\|\mathbb{E}[\mathbf{b}]\|_2^2. \tag{155}$$

For convenience, define the expression

$$c = (\|\mathbb{E}[\mathbf{W}]\|_2^2 \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty + \|\mathbb{E}[\mathbf{b}]\|_2^2)^{1/2}. \tag{156}$$

Using the subadditivity of the square root, we can write the previous two expressions as

$$\mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_1'\|_2^2]^{1/2} \leq \sqrt{2}c \tag{157}$$

and

$$\mathbb{E}[\|\mathbf{z}_1'\|_2^2]^{1/2} \leq \sqrt{3}\mathbb{E}[\|\mathbf{z}_1\|_2^2]^{1/2} + \sqrt{3}c, \tag{158}$$

respectively. To bound the last term in Equation (150), let $\mathbf{w}_m$ denote the m^{th} row of $\mathbf{W}$. Also let $\mathbf{w}_m' = \mathbf{w}_m - \mathbb{E}[\mathbf{w}_m]$ and $\mathbf{b}' = \mathbf{b} - \mathbb{E}[\mathbf{b}]$. Let $m, m' \in [M_1]$, $m \neq m'$. By the assumed symmetry condition,

$$((\mathbf{w}_m, \mathbf{w}_{m'}), \mathbf{z}_2, (b_m, b_{m'})) \overset{\text{d}}{=} ((-\mathbf{w}_m, \mathbf{w}_{m'}), \mathbf{z}_2, (-b_m, b_{m'})). \tag{159}$$

Therefore, using oddness,

$$-\phi_1(z_{1,m}')\phi_1(z_{1,m'}') = -\phi_1(\langle \tfrac{1}{M_2}\mathbf{w}_m', \phi_2(\mathbf{z}_2)\rangle + b_m')\phi_1(\langle \tfrac{1}{M_2}\mathbf{w}_{m'}', \phi_2(\mathbf{z}_2)\rangle + b_{m'}') \tag{160}$$

$$= \phi_1(\langle -\tfrac{1}{M_2}\mathbf{w}_m', \phi_2(\mathbf{z}_2)\rangle - b_m')\phi_1(\langle \tfrac{1}{M_2}\mathbf{w}_{m'}', \phi_2(\mathbf{z}_2)\rangle + b_{m'}') \tag{161}$$

$$\overset{\text{d}}{=} \phi_1(\langle \tfrac{1}{M_2}\mathbf{w}_m', \phi_2(\mathbf{z}_2)\rangle + b_m')\phi_1(\langle \tfrac{1}{M_2}\mathbf{w}_{m'}', \phi_2(\mathbf{z}_2)\rangle + b_{m'}') \tag{162}$$

$$= \phi_1(z_{1,m}')\phi_1(z_{1,m'}'), \tag{163}$$

which means that $\mathbb{E}[\phi_1(z_{1,m}')\phi_1(z_{1,m'}')] = 0$. Consequently, the matrix $\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1')^\mathsf{T}]$ is diagonal, so

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1')\phi_1(\mathbf{z}_1')^\mathsf{T}]\|_2 = \max_{m \in [M_1]} \mathbb{E}[\phi_1^2(z_{1,m}')] = \|\mathbb{E}[\phi_1^2(\mathbf{z}_1')]\|_\infty. \tag{164}$$

To bound $\|\mathbb{E}[\phi_1^2(\mathbf{z}_1')]\|_\infty$, consider that

$$\phi_1^2(\mathbf{z}_1') \preceq (\mathbf{z}_1')^2 \tag{165}$$

$$\preceq 3\mathbf{z}_1^2 + 3\tfrac{1}{M_2}(\mathbb{E}[\mathbf{W}]\phi_2(\mathbf{z}_2))^2 + 3\,\mathbb{E}[\mathbf{b}]^2 \tag{166}$$

$$\preceq 3\mathbf{z}_1^2 + 3\|\mathbb{E}[\mathbf{W}]\|_2^2 \tfrac{1}{M_2}\|\phi_2(\mathbf{z}_2)\|_2^2\mathbf{1} + 3\,\mathbb{E}[\mathbf{b}]^2 \tag{167}$$

$$\preceq 3\mathbf{z}_1^2 + 3\|\mathbb{E}[\mathbf{W}]\|_2^2 \tfrac{1}{M_2}\|\mathbf{z}_2\|_2^2\mathbf{1} + 3\,\mathbb{E}[\mathbf{b}]^2 \tag{168}$$

where the squares and inequalities are element-wise. We use the oddness and 1-Lipschitzness of ϕ_1 and ϕ_2 in the first and last inequalities, respectively, and we use the manipulation $\mathbf{A}\mathbf{v} \preceq \|\mathbf{A}\mathbf{v}\|_\infty\mathbf{1} \preceq \|\mathbf{A}\mathbf{v}\|_2\mathbf{1} \preceq \|\mathbf{A}\|_2\|\mathbf{v}\|_2\mathbf{1}$ in the third inequality. Therefore, using that $\|\mathbb{E}[\mathbf{b}]\|_\infty \leq \|\mathbb{E}[\mathbf{b}]\|_2$, we have

$$\|\mathbb{E}[\phi_1^2(\mathbf{z}_1')]\|_\infty \leq 3\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty + 3\|\mathbb{E}[\mathbf{W}]\|_2^2 \tfrac{1}{M_2}\mathbb{E}[\|\mathbf{z}_2\|_2^2] + 3\|\mathbb{E}[\mathbf{b}]\|_\infty^2 \tag{169}$$

$$= 3\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty + 3c^2 \tag{170}$$

We can now plug the bounds on the $\mathbf{z}_1'$ terms, given by Equations (157), (143), and (170), into Equation (150):

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq \sqrt{2}c\left((1+\sqrt{3})\mathbb{E}[\|\mathbf{z}_1\|_2^2]^{1/2}+\sqrt{3}c\right)+3\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+3c^2 \tag{171}$$

$$\leq \sqrt{2}(1+\sqrt{3})\mathbb{E}[\|\mathbf{z}_1\|_2^2]^{1/2}c+(3+\sqrt{6})(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+c^2) \tag{172}$$

$$\leq \sqrt{2}(1+\sqrt{3})\sqrt{M_1}\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty^{1/2}c+(3+\sqrt{6})(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+c^2), \tag{173}$$

where we add an extra $2\sqrt{3}\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty$ term to more simply group the terms in second inequality and we covert the 2-norm of $\mathbf{z}_2$ to an ∞-norm in the third inequality, as in Equation (154). Next, since $2ab+a^2+b^2=(a+b)^2\leq 2a^2+2b^2 \implies ab\leq \frac{1}{2}(a^2+b^2)$, we can simplify the following expression in the first term above as

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty^{1/2}c \leq \tfrac{1}{2}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+c^2). \tag{174}$$

Therefore,

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq \tfrac{1}{2}\sqrt{2}(1+\sqrt{3})\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+c^2)+(3+\sqrt{6})(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+c^2) \tag{175}$$

$$\leq \left(\tfrac{1}{2}\sqrt{2}(1+\sqrt{3})\sqrt{M_1}+(3+\sqrt{6})\right)(\tfrac{1}{2}\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+c^2) \tag{176}$$

If $M_1\geq 3$, then $(3+\sqrt{6})\leq \frac{1}{2}\sqrt{2}(1+\sqrt{3})\sqrt{M_1}$, so we have

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq \sqrt{2}(1+\sqrt{3})\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+c^2 \tag{177}$$

$$\leq 4\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty^{1/2}+\|\mathbb{E}[\mathbf{W}]\|_2^2\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty+\|\mathbb{E}[\mathbf{b}]\|_2^2). \tag{178}$$

Otherwise, if $M_1<3$, then the inequality holds with the slightly worse constant $\sqrt{2}(1+\sqrt{3})+(3+\sqrt{6})\approx 7.4\leq 8$:

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq 8\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty^{1/2}+\|\mathbb{E}[\mathbf{W}]\|_2^2\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty+\|\mathbb{E}[\mathbf{b}]\|_2^2). \tag{179}$$

□

Remark 28. The symmetry condition for $(\mathbf{W},\mathbf{z}_2,\mathbf{b})$ is satisfied if $\mathbf{W}$ and $\mathbf{b}$, not $\mathbf{z}_2$, are element-wise independent and symmetric around their means. For such $(\mathbf{W},\mathbf{z}_2,\mathbf{b})$, the symmetry condition is also satisfied for the triple $(\mathbf{W},\mathbf{z}_2,\frac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1}+\mathbf{b})$ where $\mathbf{1}$ is the vector of all ones and $\alpha\in\mathbb{R}$. In that case, a similar conclusion holds: if instead

$$\mathbf{z}_1=\tfrac{1}{\sqrt{M_2}}\mathbf{W}(\phi_2+\alpha)(\mathbf{z}_2)+\mathbf{b}=\tfrac{1}{\sqrt{M_2}}\mathbf{W}\phi_2(\mathbf{z}_2)+(\tfrac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1}+\mathbf{b}), \tag{180}$$

then, if $M_1\geq 3$, by Lemma 27

$$\|\mathbb{E}[\phi_1(\mathbf{z}_1)\phi_1(\mathbf{z}_1)^\mathsf{T}]\|_2 \leq 4\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+\|\mathbb{E}[\mathbf{W}]\|_2^2\,\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty+\|\mathbb{E}[\tfrac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1}+\mathbf{b}]\|_2^2) \tag{181}$$

$$\leq 4\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+\|\mathbb{E}[\mathbf{W}]\|_2^2\,\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty+\tfrac{2\alpha^2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_2^2\|\mathbf{1}\|_2^2+2\|\mathbb{E}[\mathbf{b}]\|_2^2) \tag{182}$$

$$\overset{\text{(i)}}{\leq} 4\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+\|\mathbb{E}[\mathbf{W}]\|_2^2(2\alpha^2+\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty)+2\|\mathbb{E}[\mathbf{b}]\|_2^2) \tag{183}$$

$$\leq 8\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+\|\mathbb{E}[\mathbf{W}]\|_2^2(\alpha^2+\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty)+\|\mathbb{E}[\mathbf{b}]\|_2^2) \tag{184}$$

$$\overset{\text{(ii)}}{\leq} 8\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+K^2\left((\alpha^2+\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty)\vee 1\right)) \tag{185}$$

$$\leq 8\sqrt{M_1}(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty+K^2\left(\alpha^2+1\vee\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty\right)) \tag{186}$$

where in (i) we use that $\|\mathbf{1}\|_2^2=M_2$ and in (ii) we use that $\|\mathbb{E}[\mathbf{W}]\|_2^2+\|\mathbb{E}[\mathbf{b}]\|_\infty^2\leq K^2$. If $M_1<3$, then the inequality holds with the slightly worse constant of 12.

Towards developing a recursion, we now express $\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty$ in terms of $\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty$.

Lemma 29. Assume the conditions of the Lemma 27. Then

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty \leq \tfrac{2}{M_2}\|\mathbb{V}_\mathrm{d}[\mathrm{vec}(\mathbf{W})]\|_\infty\|\mathbb{E}[\mathbf{z}_2^2]\|_1+\tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_\mathrm{F}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2+2\|\mathbb{E}[\mathbf{b}^2]\|_\infty \tag{187}$$

$$\leq 2(\sqrt{2}+K)^2(1\vee\tfrac{1}{M_2}\|\mathbb{E}[\mathbf{z}_2^2]\|_1)+\tfrac{2}{M_2}K^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2 \tag{188}$$

and

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty \leq 2\|\mathbb{V}_\mathrm{d}[\mathrm{vec}(\mathbf{W})]\|_\infty\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty+\tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_\mathrm{F}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2+2\|\mathbb{E}[\mathbf{b}^2]\|_\infty \tag{189}$$

$$\leq 2(\sqrt{2}+K)^2(1\vee\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty)+\tfrac{2}{M_2}K^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2. \tag{190}$$

Proof. Let $m \in [M_2]$, and let $\mathbf{w}_m$ denote the m^{th} row of $\mathbf{W}$, so that $z_{1,m} = \frac{1}{\sqrt{M_2}}\mathbf{w}_m^T\phi_2(\mathbf{z}_2) + b_m$. Using $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\mathbb{E}[z_{1,m}^2] \leq \tfrac{2}{M_2}\,\mathbb{E}[(\mathbf{w}_m^T\phi_2(\mathbf{z}_2))^2] + 2\,\mathbb{E}[b_m^2] \tag{191}$$
$$\leq \tfrac{2}{M_2}\langle \mathbb{E}[\mathbf{w}_m^2], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle + C + 2\,\mathbb{E}[b_m^2] \tag{192}$$
$$= \tfrac{2}{M_2}\langle \mathbb{V}_{\mathrm{d}}[\mathbf{w}_m], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle + C + \tfrac{2}{M_2}\langle \mathbb{E}^2[\mathbf{w}_m], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle + 2\,\mathbb{E}[b_m^2], \tag{193}$$

where C is the sum of the off-diagonal terms of $\frac{2}{M_2}\mathbb{E}[(\mathbf{w}_m^T\phi_2(\mathbf{z}_2))^2]$. Hence, by the triangle inequality,

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty \leq \tfrac{2}{M_2}\max_m\{|\langle \mathbb{V}_{\mathrm{d}}[\mathbf{w}_m], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle|\} + \max_m\{|C + \tfrac{2}{M_2}\langle \mathbb{E}^2[\mathbf{w}_m], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle|\} + 2\,\|\mathbb{E}[\mathbf{b}^2]\|_\infty. \tag{194}$$

Consider the middle term in Equation (194). We have

$$C + \tfrac{2}{M_2}\langle \mathbb{E}^2[\mathbf{w}_m], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle = \tfrac{2}{M_2}\langle \mathbb{E}[\mathbf{w}_m], \mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\mathbb{E}[\mathbf{w}_m]\rangle \tag{195}$$
$$\leq \tfrac{2}{M_2}\|\mathbb{E}[\mathbf{w}_m]\|_2^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2, \tag{196}$$

where the second inequality follows from the Cauchy-Schwarz inequality and the definition of the operator norm. Therefore,

$$\max_m\{|C + \tfrac{2}{M_2}\langle \mathbb{E}^2[\mathbf{w}_m], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle|\} \leq \tfrac{2}{M_2}\max_m\{\|\mathbb{E}[\mathbf{w}_m]\|_2^2\}\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2 \tag{197}$$
$$\leq \tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2. \tag{198}$$

To bound the first term in Equation (194), use the Hölder inequality with the conjugate pair $(\infty, 1)$:

$$\tfrac{2}{M_2}\max_m\{|\langle \mathbb{V}_{\mathrm{d}}[\mathbf{w}_m], \mathbb{E}[\phi_2^2(\mathbf{z}_2)]\rangle|\} \leq \tfrac{2}{M_2}\max_m\{\|\mathbb{V}_{\mathrm{d}}[\mathbf{w}_m]\|_\infty\|\mathbb{E}[\phi_2^2(\mathbf{z}_2)]\|_1\} \tag{199}$$
$$= \tfrac{2}{M_2}\max_m\{\|\mathbb{V}_{\mathrm{d}}[\mathbf{w}_m]\|_\infty\}\|\mathbb{E}[\phi_2^2(\mathbf{z}_2)]\|_1 \tag{200}$$
$$= \tfrac{2}{M_2}\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty\|\mathbb{E}[\phi_2^2(\mathbf{z}_2)]\|_1. \tag{201}$$

Notice that since ϕ_2 is 1-Lipschitz and odd, we can write

$$\|\mathbb{E}[\phi_2^2(\mathbf{z}_2)]\|_1 = \|\mathbb{E}[|\phi_2(\mathbf{z}_2) - \phi_2(\mathbf{0})|^2]\|_1 \leq \|\mathbb{E}[|\mathbf{z}_2 - \mathbf{0}|^2]\|_1 = \|\mathbb{E}[\mathbf{z}_2^2]\|_1. \tag{202}$$

Plugging into Equation (194), we have

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty \leq \tfrac{2}{M_2}\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty\|\mathbb{E}[\mathbf{z}_2^2]\|_1 + \tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2 + 2\|\mathbb{E}[\mathbf{b}^2]\|_\infty \tag{203}$$
$$\leq 2(\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty + \|\mathbb{E}[\mathbf{b}^2]\|_\infty)(1 \vee \tfrac{1}{M_2}\|\mathbb{E}[\mathbf{z}_2^2]\|_1) + \tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2. \tag{204}$$

The result follows by applying Lemma 9 and Proposition 10. Similarly, using $\|\mathbb{E}[\mathbf{z}_2^2]\|_1 \leq M_2\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty$, we can again plug into Equation (194) to obtain

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty \leq 2\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty + \tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2 + 2\|\mathbb{E}[\mathbf{b}^2]\|_\infty \tag{205}$$
$$\leq 2(\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty + \|\mathbb{E}[\mathbf{b}^2]\|_\infty)(1 \vee \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty) + \tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^\mathsf{T}]\|_2. \tag{206}$$

□

Remark 30. Like in Remark 28, apply Lemma 29 to the triple $(\mathbf{W}, \mathbf{z}_2, \frac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1} + \mathbf{b})$ where $\mathbf{W}$ and $\mathbf{b}$, not $\mathbf{z}_2$, are element-wise independent and symmetric around their means. To write down the result, we need to estimate $\|\mathbb{E}[(\frac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1} + \mathbf{b})^2]\|_\infty$. To begin with, note that

$$(\tfrac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1} + \mathbf{b})^2 = (\tfrac{\alpha}{\sqrt{M_2}}(\mathbf{W} - \mathbb{E}[\mathbf{W}])\mathbf{1} + \tfrac{\alpha}{\sqrt{M_2}}\mathbb{E}[\mathbf{W}]\mathbf{1} + \mathbf{b})^2 \tag{207}$$
$$\preceq 3\tfrac{\alpha^2}{M_2}((\mathbf{W} - \mathbb{E}[\mathbf{W}])\mathbf{1})^2 + 3\tfrac{\alpha^2}{M_2}(\mathbb{E}[\mathbf{W}]\mathbf{1})^2 + 3\mathbf{b}^2 \tag{208}$$
$$\preceq 3\tfrac{\alpha^2}{M_2}((\mathbf{W} - \mathbb{E}[\mathbf{W}])\mathbf{1})^2 + 3\tfrac{\alpha^2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_\infty^2\|\mathbf{1}\|_\infty^2\mathbf{1} + 3\mathbf{b}^2 \tag{209}$$
$$\overset{\text{(i)}}{\preceq} 3\tfrac{\alpha^2}{M_2}((\mathbf{W} - \mathbb{E}[\mathbf{W}])\mathbf{1})^2 + 3\alpha^2\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2\mathbf{1} + 3\mathbf{b}^2 \tag{210}$$

where the squares and inequalities are element-wise and in (i) we use that $\|\mathbb{E}[\mathbf{W}]\|_\infty \le \sqrt{M_2}\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}$. For the first term, consider the m^{th} element and use independence of the elements of $\mathbf{W}$:

$$\mathbb{E}[\tfrac{1}{M_2}((\mathbf{W}-\mathbb{E}[\mathbf{W}])\mathbf{1})^2]_m = \tfrac{1}{M_2}\mathbb{E}\big[\big(\textstyle\sum_{m'=1}^{M_2}(W_{m,m'}-\mathbb{E}[W_{m,m'}])\big)^2\big] \tag{211}$$
$$= \tfrac{1}{M_2}\textstyle\sum_{m'=1}^{M_2}\mathbb{E}[(W_{m,m'}-\mathbb{E}[W_{m,m'}])^2] \tag{212}$$
$$\le \|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty. \tag{213}$$

Therefore,

$$\|\mathbb{E}[(\tfrac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1}+\mathbf{b})^2]\|_\infty \le 3\alpha^2\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty + 3\alpha^2\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2 + 3\|\mathbb{E}[\mathbf{b}^2]\|_\infty, \tag{214}$$

so, by Lemma 27 (specifically Equation (187)),

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty$$
$$\le 2\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty + \tfrac{2}{M_2}\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^{\mathsf{T}}]\|_2 + 2\|\mathbb{E}[(\tfrac{\alpha}{\sqrt{M_2}}\mathbf{W}\mathbf{1}+\mathbf{b})^2]\|_\infty, \tag{215}$$
$$\le 2\|\mathbb{V}_{\mathrm{d}}[\mathrm{vec}(\mathbf{W})]\|_\infty(3\alpha^2+\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty) + 2\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2(3\alpha^2+\tfrac{1}{M_2}\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^{\mathsf{T}}]\|_2) + 6\|\mathbb{E}[\mathbf{b}^2]\|_\infty \tag{216}$$
$$\le 6(\sqrt{2}+K)^2(\alpha^2+1\vee\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty) + 6K^2(\alpha^2+\tfrac{1}{M_2}\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^{\mathsf{T}}]\|_2). \tag{217}$$

Similarly, by Equation (189), we can write this inequality in terms of $\frac{1}{M_2}\|\mathbb{E}[\mathbf{z}_2^2]\|_1$ instead of $\|\mathbb{E}[\mathbf{z}_2^2]\|_\infty$:

$$\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty \le 6(\sqrt{2}+K)^2(\alpha^2+1\vee\tfrac{1}{M_2}\|\mathbb{E}[\mathbf{z}_2^2]\|_1) + 6K^2(\alpha^2+\tfrac{1}{M_2}\|\mathbb{E}[\phi_2(\mathbf{z}_2)\phi_2(\mathbf{z}_2)^{\mathsf{T}}]\|_2). \tag{218}$$

Lemma 31. For $\alpha = 0$ and $1 \le \ell \le L$, we have

$$\|\mathbb{E}[\mathbf{z}_\ell^2]\|_\infty \le (\gamma_{\alpha,M}(1\vee K^2))^\ell(1\vee\tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2), \tag{219}$$

with $\gamma_{\alpha,M}$ taking the value $\frac{2}{3}(13+2\sqrt{43}) \in (17,18)$ if $M \ge 36$ and $2(6+\sqrt{38}) \in (24,25)$ if $1 \le M < 36$. On the other hand, for $\alpha \ne 0$ and $1 \le \ell \le L$, we have

$$\alpha^2 + \|\mathbb{E}[\mathbf{z}_\ell^2]\|_\infty \le (\gamma_{\alpha,M}(1\vee K^2))^\ell(\alpha^2+1\vee\tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2), \tag{220}$$

with $\gamma_{\alpha,M}$ taking the value $28+\sqrt{793} \in (56,57)$ if $M \ge 36$ and $48+\sqrt{2353} \in (96,97)$ if $1 \le M < 36$.

Proof of case $\alpha = 0$. For notational convenience, define $a_\ell = 1\vee\|\mathbb{E}[\mathbf{z}_\ell^2]\|_\infty$ and $b_\ell = \|\mathbb{E}[\phi(\mathbf{z}_\ell)\phi(\mathbf{z}_\ell)^{\mathsf{T}}]\|_2$. Also take $a_0 = 1\vee\|\mathbf{x}\|_2^2$. Apply Lemma 29 with ϕ_2 given by the identity function and $\mathbf{z}_2 = \mathbf{x}$. Then

$$a_1 \le 2\left((\sqrt{2}+K)^2(1\vee\tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}^2\|_1) + \tfrac{1}{D_{\mathrm{i}}}K^2\|\mathbf{x}\mathbf{x}^{\mathsf{T}}\|_2\right) \tag{221}$$
$$\le 2\left(8(1\vee K)^2(1\vee\tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2) + \tfrac{1}{D_{\mathrm{i}}}K^2\|\mathbf{x}\|_2^2\right) \tag{222}$$
$$\le 18(1\vee K^2)(1\vee\tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2) \tag{223}$$

where the second inequality follows from $\sqrt{2}+K \le \sqrt{2}(1+K) \le 2\sqrt{2}(1\vee K)$ and $\|\mathbf{x}\mathbf{x}^{\mathsf{T}}\|_2 = \|\mathbf{x}\|_2^2$.

By Lemma 29 we have

$$a_\ell \le 2\left((\sqrt{2}+K)^2 a_{\ell-1} + \tfrac{1}{M}K^2 b_{\ell-1}\right). \tag{224}$$

Further, by Lemma 27, we have

$$b_{\ell-1} \le 4\sqrt{M}(a_{\ell-1} + K^2\, a_{\ell-2}). \tag{225}$$

Combining these estimates yields

$$a_\ell \le 2\left((\sqrt{2}+K)^2 a_{\ell-1} + \tfrac{4}{\sqrt{M}}K^2\left(a_{\ell-1}+K^2 a_{\ell-2}\right)\right) \tag{226}$$
$$\le 2\left(8(1\vee K^2)a_{\ell-1} + \tfrac{4}{\sqrt{M}}K^2\left(a_{\ell-1}+K^2 a_{\ell-2}\right)\right) \tag{227}$$
$$\le (16+\tfrac{8}{\sqrt{M}})(1\vee K^2)a_{\ell-1} + \tfrac{8}{\sqrt{M}}(1\vee K^4)a_{\ell-2}. \tag{228}$$

Assuming $M \geq 36$, we further simplify

$$a_\ell \leq (16 + \tfrac{4}{3})(1 \vee K^2)a_{\ell-1} + \tfrac{4}{3}(1 \vee K^4)a_{\ell-2}. \tag{229}$$

We choose $M \geq 36$ for convenience when subsequently applying this lemma. We will return to the general case of $M \geq 1$ later. First, notice Equation (229) is a homogeneous second-order linear recurrence relation. By finding the roots of the characteristic polynomial associated to this recurrence relation,

$$a_\ell \leq c_0 \Big(\gamma(1 \vee K^2)\Big)^\ell \tag{230}$$

for some c_0 and $\gamma = \frac{2}{3}(13 + 2\sqrt{43} \in (17, 18)$, which can be proved by induction. Checking initial conditions, we see that we can take $c_0 = 1 \vee \|\mathbf{x}\|^2$, yielding the bound

$$a_\ell \leq \Big(\gamma(1 \vee K^2)\Big)^\ell (1 \vee \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2). \tag{231}$$

On the other hand, if $1 \leq M < 36$, by the same reasoning the bound holds with constant $\gamma = 2(6 + \sqrt{38}) \in (24, 25)$. □

Proof of case $\alpha \neq 0$. The proof for the case $\alpha \neq 0$ proceeds like the case $\alpha = 0$, but uses Remarks 28 and 30 instead of Lemmas 27 and 29. Analogously define $a_\ell = \alpha^2 + 1 \vee \|\mathbb{E}[\mathbf{z}_\ell^2]\|_\infty$, $a_0 = \alpha^2 + 1 \vee \|\mathbf{x}\|_2^2$, and $b_\ell = \|\mathbb{E}[\phi(\mathbf{z}_\ell)\phi(\mathbf{z}_\ell)^\mathsf{T}]\|_2$. To begin with, apply Remark 30 with ϕ_2 given by the identity function and $\mathbf{z}_2 = \mathbf{x}$:

$$\begin{aligned}
a_1 &\leq \alpha^2 + 6(\sqrt{2} + K)^2(\alpha^2 + 1 \vee \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}^2\|_1) + 6K^2(\alpha^2 + \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\mathbf{x}^\mathsf{T}\|_2) && (232)\\
&\leq 48(1 \vee K^2)(\alpha^2 + 1 \vee \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2) + 7K^2(\alpha^2 + \|\mathbf{x}\|_2^2) && (233)\\
&\leq 55(1 \vee K^2)(\alpha^2 + 1 \vee \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2). && (234)
\end{aligned}$$

For $1 < \ell \leq L$,

$$a_\ell \leq \alpha^2 + 48(1 \vee K^2)a_{\ell-1} + 6K^2(\alpha^2 + \tfrac{1}{M}b_{\ell-1}). \tag{235}$$

By Remark 28,

$$b_{\ell-1} \leq 8\sqrt{M}(\|\mathbb{E}[\mathbf{z}_{\ell-1}^2]\|_\infty + K^2 a_{\ell-2}). \tag{236}$$

Plugging the expression for $b_{\ell-1}$ into the expression for a_ℓ,

$$\begin{aligned}
a_\ell &\leq \alpha^2 + 48(1 \vee K^2)a_{\ell-1} + 6K^2(\alpha^2 + \tfrac{8}{\sqrt{M}}(\|\mathbb{E}[\mathbf{z}_{\ell-1}^2]\|_\infty + K^2 a_{\ell-2})) && (237)\\
&\leq \alpha^2 + 48(1 \vee K^2)a_{\ell-1} + 6(1 \vee 8M^{-1/2})K^2(\alpha^2 + \|\mathbb{E}[\mathbf{z}_{\ell-1}^2]\|_\infty + K^2 a_{\ell-2}) && (238)\\
&\leq \alpha^2 + 48(1 \vee K^2)a_{\ell-1} + 6(1 \vee 8M^{-1/2})K^2(a_{\ell-1} + K^2 a_{\ell-2}) && (239)\\
&\leq \alpha^2 + (48 + 6(1 \vee 8M^{-1/2}))(1 \vee K^2)a_{\ell-1} + 6(1 \vee 8M^{-1/2})(1 \vee K^4)a_{\ell-2} && (240)\\
&\leq (48 + 6(1 \vee 8M^{-1/2}))(1 \vee K^2)a_{\ell-1} + (1 + 6(1 \vee 8M^{-1/2}))(1 \vee K^4)a_{\ell-2}. && (241)
\end{aligned}$$

As in the $\alpha = 0$ case, we first consider the case $M \geq 36$, giving

$$a_\ell \leq 56(1 \vee K^2)a_{\ell-1} + 9(1 \vee K^4)a_{\ell-2}. \tag{242}$$

This is again a homogeneous second-order linear recurrence relation. Solving for the roots of the characteristic polynomial, we find that

$$a_\ell \leq c_0 \Big(\gamma'(1 \vee K^2)\Big)^\ell, \tag{243}$$

for some c_0 and $\gamma' = 28 + \sqrt{793} \in (56, 57)$. Comparing with the bound on a_1, we see that $c_0 = \alpha^2 + 1 \vee \frac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2$. On the other hand, if $1 \leq M < 36$, by the same reasoning the bound holds with $\gamma' = 48 + \sqrt{2353} \in (96, 97)$. □

Finally, we end with a lemma which can be used to bound the covariance between two post-activations. The following lemmas in this section are only applied in Appendix G but we include them here because of their similarity.

Lemma 32. Let $\phi_1, \phi_2 \colon \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and odd and $\alpha \in \mathbb{R}$. Let the triple $(\mathbf{W}, \mathbf{z}_2, \mathbf{b}) \in \mathbb{R}^{M_1 \times M_2} \times \mathbb{R}^{M_2} \times \mathbb{R}^{M_1}$ be (possibly dependent) random variables such that, for all Rademacher vectors $\boldsymbol{\varepsilon} \in \{-1, 1\}^{M_1}$,

$$(\mathbf{W} - \mathbb{E}[\mathbf{W}], \mathbf{z}_2, \mathbf{b} - \mathbb{E}[\mathbf{b}]) \overset{\mathrm{d}}{=} (\mathrm{diag}(\boldsymbol{\varepsilon})(\mathbf{W} - \mathbb{E}[\mathbf{W}]), \mathbf{z}_2, \boldsymbol{\varepsilon} \circ (\mathbf{b} - \mathbb{E}[\mathbf{b}])). \tag{244}$$

Consider $\mathbf{z}_1 = \frac{1}{\sqrt{M_2}} \mathbf{W}(\phi_2 + \alpha)(\mathbf{z}_2) + \mathbf{b}$. Let $m, m' \in [M_1]$, $m \neq m'$. Then

$$|\mathbb{E}[\phi_1(z_{1,m})\phi_1(z_{1,m'})]| \leq 5(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty^{1/2} +)(\alpha^2 + 1 \vee \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty)^2 (K \vee K^2). \tag{245}$$

Proof. The proof of this lemma uses exactly the same approach as Lemma 27. Set

$$\mathbf{z}_1' = \mathbf{z}_1 - \tfrac{1}{\sqrt{M_2}} \mathbb{E}[\mathbf{W}](\phi_2 + \alpha)(\mathbf{z}_2) - \mathbb{E}[\mathbf{b}]. \tag{246}$$

Consider

$$|\mathbb{E}[\phi_1(z_{1,m})\phi_1(z_{1,m'})]| \leq \sqrt{\mathbb{E}[(\phi_1(z_{1,m}) - \phi_1(z'_{1,m}))^2]\mathbb{E}[\phi_1^2(z_{1,m'})]} + |\mathbb{E}[\phi_1(z'_{1,m})\phi_1(z_{1,m'})]| \tag{247}$$

$$\leq \sqrt{ab} + |\mathbb{E}[\phi_1(z'_{1,m})\phi_1(z_{1,m'})]| \tag{248}$$

where we denote

$$a = \|\mathbb{E}[\mathbf{z}_1^2]\|_\infty, \quad b = \|\mathbb{E}[(\mathbf{z}_1 - \mathbf{z}_1')^2]\|_\infty. \tag{249}$$

Similarly,

$$|\mathbb{E}[\phi_1(z'_{1,m})\phi_1(z_{1,m'})]| \leq \sqrt{\mathbb{E}[\phi_1^2(z'_{1,m})]\mathbb{E}[(\phi_1(z_{1,m'}) - \phi_1(z'_{1,m'}))^2]} + |\mathbb{E}[\phi_1(z'_{1,m})\phi_1(z'_{1,m'})]|. \tag{250}$$

Now bound

$$\phi_1^2(\mathbf{z}_1') \preceq (\mathbf{z}_1')^2 \preceq 2\mathbf{z}_1^2 + 2(\mathbf{z}_1 - \mathbf{z}_1')^2, \tag{251}$$

so

$$\sqrt{\mathbb{E}[\phi_1^2(z'_{1,m})]\mathbb{E}[(\phi_1(z_{1,m'}) - \phi_1(z'_{1,m'}))^2]} \leq \sqrt{(2a + 2b)b}. \tag{252}$$

Therefore,

$$|\mathbb{E}[\phi_1(z_{1,m})\phi_1(z_{1,m'})]| \leq \sqrt{ab} + \sqrt{(2a + 2b)b} + \cancel{|\mathbb{E}[\phi_1(z'_{1,m})\phi_1(z'_{1,m'})]|} \tag{253}$$

where the expectation on the RHS is zero by the assumed symmetry condition and oddness of ϕ. Breaking up the square root, we find that

$$|\mathbb{E}[\phi_1(z_{1,m})\phi_1(z_{1,m'})]| \leq (\sqrt{2} + 1)\sqrt{ab} + \sqrt{2}b. \tag{254}$$

We finally estimate b:

$$\mathbb{E}[(\mathbf{z}_1 - \mathbf{z}_1')^2] \preceq \|\mathbb{E}[\mathbf{W}]\|_2^2 \tfrac{1}{M_2} \mathbb{E}[\|(\phi_2 + \alpha)^2(\mathbf{z}_2)\|_2]\mathbf{1} + \|\mathbb{E}[\mathbf{b}]\|_2^2 \mathbf{1}. \tag{255}$$

Therefore

$$b \leq \|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2 \|\mathbb{E}[(\phi_2 + \alpha)^2(\mathbf{z}_2)]\|_\infty + \|\mathbb{E}[\mathbf{b}]\|_2^2 \tag{256}$$

$$\leq 2\|\mathbb{E}[\mathbf{W}]\|_{\mathrm{F}}^2 (\alpha^2 + \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty) + \|\mathbb{E}[\mathbf{b}]\|_2^2 \tag{257}$$

$$\leq 2K^2(\alpha^2 + 1 \vee \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty), \tag{258}$$

which implies that

$$|\mathbb{E}[\phi_1(z_{1,m})\phi_1(z_{1,m'})]| \leq ((\sqrt{2} + 1)\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty^{1/2} + \sqrt{2})(\sqrt{b} \vee b), \quad b = 2K^2(\alpha^2 + 1 \vee \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty). \tag{259}$$

Simplify the estimate to get the result:

$$|\mathbb{E}[\phi_1(z_{1,m})\phi_1(z_{1,m'})]| \leq 5(\|\mathbb{E}[\mathbf{z}_1^2]\|_\infty^{1/2} +)(\alpha^2 + 1 \vee \|\mathbb{E}[\mathbf{z}_2^2]\|_\infty)^2 (K \vee K^2). \tag{260}$$

□

Lemma 33. Let ϕ be 1-Lipschitz odd and $\alpha \in \mathbb{R}$. Then,

$$\begin{aligned}&\tfrac{1}{M}\operatorname{tr}\left(\mathbb{E}_Q[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^\mathsf{T}](\mathbb{E}_Q[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}] - \mathbb{E}_P[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}])\right)\\ &\quad\leq \tfrac{56+32\sqrt{2}}{\sqrt{M}}\rho_{\alpha,M}^L L^{1/2}K(K^2\vee 1)^{L+\frac{1}{2}}\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right),\end{aligned} \tag{261}$$

Proof. For convenience, define

$$\mathbf{A} = \mathbb{E}_Q[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}] - \mathbb{E}_P[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}] \tag{262}$$

Recall $\mathbf{\Sigma}_Q = \mathbb{E}_Q[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^\mathsf{T}] - \mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}$. By subtracting and adding $\mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T} + \mathbf{I}$, we have

$$|\operatorname{tr}\left(\mathbb{E}_Q[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^\mathsf{T}]\mathbf{A}\right)| = |\operatorname{tr}\left((\mathbf{\Sigma}_Q - \mathbf{I} + \mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T} + \mathbf{I})\mathbf{A}\right)| \tag{263}$$

$$= |\operatorname{tr}\left((\mathbf{\Sigma}_Q - \mathbf{I})\mathbf{A} + \mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbf{A} + \mathbf{A}\right)| \tag{264}$$

$$\leq |\operatorname{tr}\left((\mathbf{\Sigma}_Q - \mathbf{I})\mathbf{A}\right)| + |\operatorname{tr}\left(\mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbf{A}\right)| + |\operatorname{tr}\left(\mathbf{A}\right)|. \tag{265}$$

We deal with each of the three terms in Equation (265) separately.

Consider the first term in Equation (265), since $\mathbf{\Sigma}_Q - \mathbf{I} = \operatorname{diag}(\boldsymbol{\sigma}_Q^2 - \mathbf{1})$ is a diagonal matrix, by Cauchy-Schwarz

$$|\operatorname{tr}\left((\mathbf{\Sigma}_Q - \mathbf{I})\mathbf{A}\right)| = |\langle \boldsymbol{\sigma}_Q^2 - \mathbf{1}, \mathbb{E}_Q[\phi^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi^2(\mathbf{z}_L)]\rangle| \tag{266}$$

$$\leq \|\boldsymbol{\sigma}_Q^2 - \mathbf{1}\|_2 \|\mathbb{E}_Q[\phi^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi^2(\mathbf{z}_L)]\|_2 \tag{267}$$

$$\overset{\text{(i)}}{\leq} ((2+K)K)\left(8\sqrt{2}\rho_{\alpha,M}^{L-\frac{1}{2}}L^{1/2}K(K^2\vee 1)^{L-\frac{1}{2}}\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right)\right) \tag{268}$$

$$\overset{\text{(ii)}}{\leq} 32\sqrt{2}\rho_{\alpha,M}^{L-\frac{1}{2}}L^{1/2}K(K^2\vee 1)^{L+\frac{1}{2}}\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right), \tag{269}$$

where in (i) we use Lemma 9 and Lemma 35 (proved later) while in (ii) we use $2+K \leq 2(1+K) \leq 4(K\vee 1) = 4(K^2\vee 1)^{1/2}$ and $K \leq K\vee 1 = (K^2\vee 1)^{1/2}$.

Next, consider the second term in Equation (265), since $\operatorname{tr}(\mathbf{u}\mathbf{v}^T) = \mathbf{v}^T\mathbf{u}$ for $\mathbf{u},\mathbf{v}\in\mathbb{R}^M$,

$$|\operatorname{tr}\left(\mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbf{A}\right)| = |\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbf{A}\mathbb{E}_Q[\mathbf{w}_{L+1}]| \tag{270}$$

$$\leq \|\mathbb{E}_Q[\mathbf{w}_{L+1}]\|_2^2\|\mathbf{A}\|_2 \tag{271}$$

$$\leq K^2\|\mathbf{A}\|_2, \tag{272}$$

where the first inequality follows from Cauchy-Schwarz and the definition of the operator norm while the second inequality follows from Lemma 9. To bound $\|\mathbf{A}\|_2$, notice by Lemma 36 (proved later), we have

$$\|\mathbb{E}_Q[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}]\| \leq 24\sqrt{M}\left(\gamma_{\alpha,M}(K^2\vee 1)\right)^L\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right) \tag{273}$$

and

$$\|\mathbb{E}_P[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}]\| \leq 24\sqrt{M}\left(\gamma_{\alpha,M}\right)^L\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right) \tag{274}$$

$$\leq 24\sqrt{M}\left(\gamma_{\alpha,M}(K^2\vee 1)\right)^L\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right) \tag{275}$$

Therefore, using the triangle inequality,

$$\|\mathbf{A}\|_2 \leq \|\mathbb{E}_Q[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}]\|_2 + \|\mathbb{E}_P[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}]\|_2 \tag{276}$$

$$\leq 48\sqrt{M}\left(\gamma_{\alpha,M}(K^2\vee 1)\right)^L\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right). \tag{277}$$

Putting in the bound on $\|A\|_2$, a bound on the second term in Equation (265) is

$$|\operatorname{tr}\left(\mathbb{E}_Q[\mathbf{w}_{L+1}]\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbf{A}\right)| \leq 48\sqrt{M}K^2\left(\gamma_{\alpha,M}(K^2\vee 1)\right)^L\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right) \tag{278}$$

$$\leq 48\sqrt{M}\gamma_{\alpha,M}^L K(K^2\vee 1)^{L+\frac{1}{2}}\left(\alpha^2+1\vee\tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right), \tag{279}$$

where we use $K \leq K \vee 1 = (K^2 \vee 1)^{1/2}$

Finally, consider the third term in Equation (265). We have

$$\begin{aligned}
\operatorname{tr}(\mathbf{A}) &= \left|\operatorname{tr}\left(\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}] - \mathbb{E}_P[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^{\mathsf{T}}]\right)\right| && (280)\\
&= \left|\langle \mathbf{1}, \mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\rangle\right| && (281)\\
&\leq \|\mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\|_1 && (282)\\
&\leq \sqrt{M}\|\mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\|_2 && (283)\\
&\leq 8\sqrt{2}\sqrt{M}\rho_{\alpha,M}^{L-\frac{1}{2}} L^{1/2} K(K^2 \vee 1)^{L-\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right) && (284)
\end{aligned}$$

where the last inequality follows from Lemma 35.

Here are the three bounds for reference:

$$32\sqrt{2}\rho_{\alpha,M}^{L-\frac{1}{2}} L^{1/2} K(K^2 \vee 1)^{L+\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right), \tag{285}$$

$$48\sqrt{M}\gamma_{\alpha,M}^{L} K(K^2 \vee 1)^{L+\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right), \text{ and} \tag{286}$$

$$8\sqrt{2}\sqrt{M}\rho_{\alpha,M}^{L-\frac{1}{2}} L^{1/2} K(K^2 \vee 1)^{L-\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right). \tag{287}$$

Finally, plugging into Equation (265)

$$\tfrac{1}{M}\left|\operatorname{tr}\left(\mathbb{E}_Q[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^{\mathsf{T}}]\mathbf{A}\right)\right| \leq \tfrac{56+32\sqrt{2}}{\sqrt{M}}\rho_{\alpha,M}^{L} L^{1/2} K(K^2 \vee 1)^{L+\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right), \tag{288}$$

where we use $1 \leq \gamma_{\alpha,M} \leq \rho_{\alpha,M}$, $L \geq 1$, and $M \geq 1$. □

Lemma 34. Let ϕ be 1-Lipschitz odd and $\alpha \in \mathbb{R}$. Then,

$$\tfrac{1}{M}\left|\mathbb{E}[\langle \mathbf{w}, \alpha\mathbf{1}\rangle\langle \mathbf{w}, \phi(\mathbf{z}_L)\rangle]\right| \leq 10\tfrac{|\alpha|}{\sqrt{M}} L(2 + |\alpha| + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(2+2c)^{L-1} K(K \vee 1)^{L+1}. \tag{289}$$

Proof. For convenience let $\mathbf{w} = \mathbf{w}_{L+1}$. Note that

$$\begin{aligned}
\tfrac{1}{M}\left|\mathbb{E}[\langle \mathbf{w}, \alpha\mathbf{1}\rangle\langle \mathbf{w}, \phi(\mathbf{z}_L)\rangle]\right| &= \tfrac{|\alpha|}{M}\left|\mathbf{1}\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathsf{T}}]\mathbb{E}[\phi(\mathbf{z}_L)]\right| && (290)\\
&\leq \tfrac{|\alpha|}{\sqrt{M}}\|\mathbb{V}[\mathbf{w}] + \mathbb{E}[\mathbf{w}]\mathbb{E}[\mathbf{w}]^{\mathsf{T}}\|_2\|\mathbb{E}[\phi(\mathbf{z}_L)]\|_2 && (291)\\
&\leq \tfrac{|\alpha|}{\sqrt{M}}(\|\mathbb{V}[\mathbf{w}]\|_2 + \|\mathbb{E}[\mathbf{w}]\mathbb{E}[\mathbf{w}]^{\mathsf{T}}\|_2)\|\mathbb{E}[\phi(\mathbf{z}_L)]\|_2 && (292)\\
&= \tfrac{|\alpha|}{\sqrt{M}}(\|\mathbb{V}_{\mathrm{d}}[\mathbf{w}]\|_\infty + \|\mathbb{E}[\mathbf{w}]\|_2^2)\|\mathbb{E}[\phi(\mathbf{z}_L)]\|_2. && (293)
\end{aligned}$$

Using Lemma 9,

$$\|\mathbb{V}_{\mathrm{d}}[\mathbf{w}]\|_\infty + \|\mathbb{E}[\mathbf{w}]\|_2^2 \leq \sigma_{\max}^2 + K^2 \leq (1+K)^2 + K^2 \leq 5(K \vee 1)^2. \tag{294}$$

Therefore, using Lemma 13,

$$\tfrac{1}{M}\left|\mathbb{E}[\langle \mathbf{w}, \alpha\mathbf{1}\rangle\langle \mathbf{w}, \phi(\mathbf{z}_L)\rangle]\right| \leq 10\tfrac{|\alpha|}{\sqrt{M}} L(2 + |\alpha| + \tfrac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2)(2+2c)^{L-1} K(K \vee 1)^{L+1} \tag{295}$$

□

E.2 Diagonal Terms

The main technical result needed for bounding the difference in the diagonal terms is the following bound on the Frobenius norm between the difference between these matrices.

Lemma 35. We have,

$$\|\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)^2] - \mathbb{E}_P[\phi_{\mathrm{o}}(\mathbf{z}_L)^2]\|_2 \leq 8\sqrt{2}\rho_{\alpha,M}^{L-\frac{1}{2}} L^{1/2} K(K^2 \vee 1)^{L-\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right), \tag{296}$$

where $K = \sqrt{2\,\mathrm{KL}(Q,P)}$ and

$$\rho_{\alpha,M} = \begin{cases} 12(2+\sqrt{2\pi}) \in (54,55) & \alpha = 0, M < 36 \\ \frac{2}{3}(13+2\sqrt{43}) \in (17,18) & \alpha = 0, M \geq 36 \\ 28+\sqrt{793} \in (56,57) & \alpha \neq 0, M < 36 \\ 48+\sqrt{2353} \in (96,97) & \alpha \neq 0, M \geq 36. \end{cases} \tag{297}$$

Proof. We would like to combine the two expectations into a single expectation. To do this, we couple $\phi(\mathbf{z}_\ell)^2$ under P and Q by using the reparameterization trick and having them share the same noise. In particular, for $2 \leq \ell \leq L$, define

$$\mathbf{z}_\ell^P = \tfrac{1}{\sqrt{M}}\boldsymbol{\mathcal{E}}\phi_{\mathrm{o}}(\mathbf{z}_{\ell-1}^P) + \tfrac{\alpha}{\sqrt{M}}\boldsymbol{\mathcal{E}}\mathbf{1} + \boldsymbol{\varepsilon}, \tag{298}$$

$$\mathbf{z}_L^Q = \tfrac{1}{\sqrt{M}}(\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q)\phi_{\mathrm{o}}(\mathbf{z}_{\ell-1}^Q) + \tfrac{\alpha}{\sqrt{M}}(\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q)\mathbf{1} + (\mathbf{s}_\ell^Q \circ \boldsymbol{\varepsilon} + \mathbf{m}_\ell^Q). \tag{299}$$

where $\boldsymbol{\mathcal{E}}$ is a matrix of i.i.d. standard Gaussian random variables, $\boldsymbol{\varepsilon}$ is a vector of i.i.d. standard Gaussian random variables, $\mathbf{M}_L^Q$ is a matrix consisting of the mean of each weight in layer L under Q, $\mathbf{m}_L^Q$ is vector consisting of the mean of each bias in layer L under Q, $\mathbf{S}_L^Q$ and $\mathbf{s}_L^Q$ are a matrix and vector containing the standard deviations of each weight or bias respectively in layer L under Q. We then can rewrite,

$$\|\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)^2] - \mathbb{E}_P[\phi_{\mathrm{o}}(\mathbf{z}_L)^2]\|_2 = \|\mathbb{E}[\phi_{\mathrm{o}}(\mathbf{z}_L^Q)^2 - \phi_{\mathrm{o}}(\mathbf{z}_L^P)^2]\|_2 \tag{300}$$

$$= \sqrt{\sum_{m=1}^M \left(\mathbb{E}[(\phi_{\mathrm{o}}(z_{m,L}^Q) + \phi_{\mathrm{o}}(z_{m,L}^P))(\phi_{\mathrm{o}}(z_{m,L}^Q) - \phi_{\mathrm{o}}(z_{m,L}^P))]\right)^2} \tag{301}$$

$$\leq \sqrt{\sum_{m=1}^M \mathbb{E}[(\phi_{\mathrm{o}}(z_{m,L}^Q) + \phi_{\mathrm{o}}(z_{m,L}^P))^2]\mathbb{E}[(\phi_{\mathrm{o}}(z_{m,L}^Q) - \phi_{\mathrm{o}}(z_{m,L}^P))^2]} \tag{302}$$

$$\leq \sqrt{\left\|\mathbb{E}[(\phi_{\mathrm{o}}(\mathbf{z}_L^Q) + \phi_{\mathrm{o}}(\mathbf{z}_L^P))^2]\right\|_\infty}\sqrt{\sum_{m=1}^M \mathbb{E}[(\phi_{\mathrm{o}}(z_{m,L}^Q) - \phi_{\mathrm{o}}(z_{m,L}^P))^2]} \tag{303}$$

$$= \sqrt{\|\mathbb{E}[(\phi_{\mathrm{o}}(\mathbf{z}_L^Q) + \phi_{\mathrm{o}}(\mathbf{z}_L^P))^2]\|_\infty}\sqrt{\mathbb{E}[\|\phi_{\mathrm{o}}(\mathbf{z}_L^Q) - \phi_{\mathrm{o}}(\mathbf{z}_L^P)\|_2^2]} \tag{304}$$

$$\leq \sqrt{\|\mathbb{E}[(\phi_{\mathrm{o}}(\mathbf{z}_L^Q) + \phi_{\mathrm{o}}(\mathbf{z}_L^P))^2]\|_\infty}\sqrt{\mathbb{E}[\|\mathbf{z}_L^Q - \mathbf{z}_L^P\|_2^2]} \tag{305}$$

The first inequality is an element-wise application of Cauchy-Schwarz viewing the expectation as an inner product, the second inequality is a bound of the form $\sum |a_i b_i| \leq \sup |a_i| \sum |b_i|$, and the third inequality uses the Lipschitz property of ϕ_{o}. We next bound the square of each of the two terms in Equation (305).

Bounding $\|\mathbb{E}[(\phi_{\mathbf{o}}(\mathbf{z}_L^Q) + \phi_{\mathbf{o}}(\mathbf{z}_L^P))^2]\|_\infty$. Using the inequality $(a_1+a_2)^2 \leq 2(a_1^2+a_2^2)$ and the triangle inequality we have

$$\|\mathbb{E}[(\phi_{\mathrm{o}}(\mathbf{z}_L^Q) + \phi_{\mathrm{o}}(\mathbf{z}_L^P))^2]\|_\infty \leq \|\mathbb{E}[2\phi_{\mathrm{o}}^2(\mathbf{z}_L^Q) + 2\phi_{\mathrm{o}}^2(\mathbf{z}_L^P)]\|_\infty \tag{306}$$

$$\leq 2\|\mathbb{E}[\phi_{\mathrm{o}}^2(\mathbf{z}_L^Q)]\|_\infty + 2\|\mathbb{E}[\phi_{\mathrm{o}}^2(\mathbf{z}_L^P)]\|_\infty \tag{307}$$

$$\leq 2\|\mathbb{E}[(\mathbf{z}_L^Q)^2]\|_\infty + 2\|\mathbb{E}[(\mathbf{z}_L^P)^2]\|_\infty, \tag{308}$$

where the last inequality follows from the oddness and Lipschitz property of ϕ_{o}:

$$\|\mathbb{E}[\phi_{\mathrm{o}}^2(\mathbf{z}_L^Q)]\|_\infty = \|\mathbb{E}[(\phi_{\mathrm{o}}(\mathbf{z}_L^Q) - \phi_{\mathrm{o}}(\mathbf{0}))^2]\|_\infty \leq \|\mathbb{E}[(\mathbf{z}_L^Q)^2]\|_\infty. \tag{309}$$

Likewise, $\|\mathbb{E}[\phi^2(\mathbf{z}_L^P)]\|_\infty \leq \|\mathbb{E}[(\mathbf{z}_L^P)^2]\|_\infty$. We can upper by applying Lemma 31 to each of the two terms:

$$\|\mathbb{E}_Q[(\mathbf{z}_L)^2]\|_\infty \leq \left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right) - \alpha^2 \tag{310}$$

and

$$\|\mathbb{E}_P[(\mathbf{z}_L)^2]\|_\infty \leq \gamma_{\alpha,M}^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2\right) - \alpha^2. \tag{311}$$

Therefore, the first term in Equation (305) is upper bounded by

$$\|\mathbb{E}[(\phi_\mathrm{o}(\mathbf{z}_L^Q) + \phi_\mathrm{o}(\mathbf{z}_L^P))^2]\|_\infty \leq 2\gamma_{\alpha,M}^L((K^2 \vee 1)^L + 1)(\alpha^2 + 1 \vee \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2) - 4\alpha^2 \tag{312}$$

$$\leq 4\gamma_{\alpha,M}^L(K^2 \vee 1)^L(\alpha^2 + 1 \vee \tfrac{1}{D_\mathrm{i}}\|\mathbf{x}\|_2^2), \tag{313}$$

where we use that $(K^2 \vee 1)^L + 1 \leq 2(K^2 \vee 1)$ and drop the $-4\alpha^2$ term for simplicity later on.

Bounding $\mathbb{E}[\|\mathbf{z}_L^Q - \mathbf{z}_L^P\|_2^2]$ It remains to upper bound the second term in Equation (305). To do so, we will setup a linear, non-homogenous recursion relation. To start, notice for $2 \leq \ell \leq L$, by adding and subtracting $\boldsymbol{\mathcal{E}}\phi_\mathrm{o}(\mathbf{z}_{\ell-1}^Q)$ and then applying the triangle inequality, we have

$$\begin{aligned}\|\mathbf{z}_\ell^Q - \mathbf{z}_\ell^P\|_2 \leq& \tfrac{1}{\sqrt{M}}\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_2\|\phi_\mathrm{o}(\mathbf{z}_{\ell-1}^Q)\|_2 + \alpha\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_2 \\ &+ \tfrac{1}{\sqrt{M}}\|\boldsymbol{\mathcal{E}}\|_2\|\phi_\mathrm{o}(\mathbf{z}_{\ell-1}^Q) - \phi_\mathrm{o}(\mathbf{z}_{\ell-1}^P)\|_2 + \|\mathbf{s}_\ell^Q \circ \boldsymbol{\varepsilon} + \mathbf{m}_\ell^Q - \boldsymbol{\varepsilon}\|_2,\end{aligned} \tag{314}$$

where we note that $\|\mathbf{1}\|_2 = \sqrt{M}$ cancels the $\frac{1}{\sqrt{M}}$ term in the second term. To obtain a tighter bound, we consider the cases of $\alpha = 0$ and $\alpha \neq 0$ separately by defining $c_\alpha = 3$ if $\alpha = 0$ and $c_\alpha = 4$ if $\alpha \neq 0$. Then, for any $a_1, a_2, a_3, a_4 \in \mathbb{R}$, we have $(a_1 + a_2 + a_3 + \alpha a_4)^2 \leq c_\alpha(a_1^2 + a_2^2 + a_3^2 + \alpha^2 a_4^2)$. Applying this expression to Equation (314), squaring both sides, and then taking the expectation we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{z}_\ell^Q - \mathbf{z}_\ell^P\|_2^2] \leq& \tfrac{c_\alpha}{M}\mathbb{E}[\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_2^2]\mathbb{E}[\|\phi_\mathrm{o}(\mathbf{z}_{\ell-1}^Q)\|_2^2] + c_\alpha\alpha^2\mathbb{E}[\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_2^2] \\ &+ \tfrac{c_\alpha}{M}\mathbb{E}[\|\boldsymbol{\mathcal{E}}\|_2^2]\mathbb{E}[\|\phi_\mathrm{o}(\mathbf{z}_{\ell-1}^Q) - \phi_\mathrm{o}(\mathbf{z}_{\ell-1}^P)\|_2^2] + c_\alpha\mathbb{E}[\|\mathbf{s}_\ell^Q \circ \boldsymbol{\varepsilon} + \mathbf{m}_\ell^Q - \boldsymbol{\varepsilon}\|_2^2].\end{aligned} \tag{315}$$

We now turn to bounding each of the terms in Equation (315). First, by Lemma 26,

$$\tfrac{1}{M}\mathbb{E}[\|\boldsymbol{\mathcal{E}}\|_2^2] \leq \eta_M. \tag{316}$$

Next, we have

$$\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_2 \leq \|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_\mathrm{F} \tag{317}$$

$$\leq \|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}\|_\mathrm{F} + \|\mathbf{M}_\ell^Q\|_\mathrm{F} \tag{318}$$

$$\implies \mathbb{E}[\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_2^2] \leq 2\mathbb{E}[\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}\|_\mathrm{F}^2 + 2\mathbb{E}[\|\mathbf{M}_\ell^Q\|_\mathrm{F}^2]. \tag{319}$$

We can then bound each term by $\mathrm{KL}(Q, P)$. To do this, first notice we can write

$$\mathbb{E}[\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}\|_\mathrm{F}^2] = \mathbb{E}\left[\sum_m \sum_{m'} (\boldsymbol{\sigma}_\ell^Q - 1)^2 \epsilon_{m,m'}^2\right] = \sum_m \sum_{m'} (\boldsymbol{\sigma}_\ell^Q - 1)^2 \mathbb{E}\left[\epsilon_{m,m'}^2\right] = \sum_m \sum_{m'} (\boldsymbol{\sigma}_\ell^Q - 1)^2. \tag{320}$$

By an application of Lemma 9, we then have,

$$\mathbb{E}[\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}\|_\mathrm{F}^2] \leq K^2 \text{ and } \mathbb{E}[\|\mathbf{M}_\ell^Q\|_\mathrm{F}^2] \leq K^2, \tag{321}$$

where, as before, $K = \sqrt{2\,\mathrm{KL}(Q, P)}$. Therefore, we obtain

$$\mathbb{E}[\|\mathbf{S}_\ell^Q \circ \boldsymbol{\mathcal{E}} + \mathbf{M}_\ell^Q - \boldsymbol{\mathcal{E}}\|_2^2] \leq 4K^2.$$

We can apply an identical argument to $\|\mathbf{s}_\ell^Q \circ \boldsymbol{\varepsilon} + \mathbf{m}_\ell^Q - \boldsymbol{\varepsilon}\|_2$ to conclude that

$$\mathbb{E}[\|\mathbf{s}_\ell^Q \circ \boldsymbol{\varepsilon} + \mathbf{m}_\ell^Q - \boldsymbol{\varepsilon}\|_2^2] \leq 4K^2. \tag{322}$$

Finally, we have that

$$\mathbb{E}[\|\phi_\mathrm{o}(\mathbf{z}_{\ell-1}^Q)\|_2^2] = \mathbb{E}\left[\sum_{m=1}^M \phi_\mathrm{o}^2(z_{\ell-1,m}^Q)\right] = \sum_{m=1}^M \mathbb{E}[\phi_\mathrm{o}^2(z_{\ell-1,m}^Q)] \leq M\|\mathbb{E}[\phi_\mathrm{o}^2(\mathbf{z}_{\ell-1}^Q)]\|_\infty \leq M\|\mathbb{E}[(\mathbf{z}_{\ell-1}^Q)^2]\|_\infty, \tag{323}$$

where the last inequality follows from the oddness and Lipschitzness of ϕ_{o}. Dividing by M and applying Lemma 31 to the last expression, we can conclude that

$$\tfrac{1}{M}\mathbb{E}[\|\phi_{\text{o}}(\mathbf{z}_{\ell-1}^Q)\|_2^2] \leq \left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^{\ell-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2\right) - \alpha^2. \tag{324}$$

We now plug these expressions into Equation (314), noting the cancellation of α^2 terms, to obtain

$$\mathbb{E}[\|\mathbf{z}_\ell^Q - \mathbf{z}_\ell^P\|_2^2] \leq 4c_\alpha K^2\left(\left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^{\ell-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2\right) + 1\right) + c_\alpha\eta_M\,\mathbb{E}[\|\mathbf{z}_{\ell-1}^Q - \mathbf{z}_{\ell-1}^P\|_2^2]. \tag{325}$$

We now set up the recursion, with base case

$$\mathbb{E}[\|\mathbf{z}_1^Q - \mathbf{z}_1^P\|_2^2] \leq 4c_\alpha K^2\left(\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2\right) + 1\right). \tag{326}$$

Taken together, Equation (325) and Equation (326) define a linear, non-homogeneous recursion in $\mathbb{E}[\|\mathbf{z}_\ell^Q - \mathbf{z}_\ell^P\|_2^2]$ with variable coefficients. By unrolling the recursion, we find that[3]

$$\mathbb{E}[\|\mathbf{z}_L^Q - \mathbf{z}_L^P\|_2^2] \leq \sum_{\ell=1}^{L}(c_\alpha\eta_M)^{L-\ell}\left(4c_\alpha K^2\left(\left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^{\ell-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2\right) + 1\right)\right) \tag{327}$$

$$\leq (c_\alpha\eta_M \vee \gamma_{\alpha,M})^{L-1}4c_\alpha K^2\sum_{\ell=1}^{L}(K^2 \vee 1)^{\ell-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2 + 1\right), \tag{328}$$

where in the second inequality we factor out $(\gamma_{\alpha,M}(K^2 \vee 1))^{\ell-1}$ (which is possible since it is greater than 1) and then combine $(4c)^{L-\ell}\gamma_{\alpha,M}^{\ell-1} \leq (4c \vee \gamma_{\alpha,M})^{L-1}$. We can bound the sum by L times the largest term in the sum, which is the $\ell = L$ term since $K^2 \vee 1 \geq 1$:

$$\mathbb{E}[\|\mathbf{z}_L^Q - \mathbf{z}_L^P\|_2^2] \leq L\left[(c_\alpha\eta_M \vee \gamma_{\alpha,M})^{L-1}4c_\alpha K^2(K^2 \vee 1)^{L-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2 + 1\right)\right] \tag{329}$$

$$\leq 32(c_\alpha\eta_M \vee \gamma_{\alpha,M})^{L-1}LK^2(K^2 \vee 1)^{L-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2\right) \tag{330}$$

where the second inequality uses that $1 \vee \frac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2 + 1 \leq 2(1 \vee \frac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2)$ and $c_\alpha \leq 4$. Define $\rho_{\alpha,M} = c_\alpha\eta_M \vee \gamma_{\alpha,M}$. Recall from Lemma 26 that $\eta_M \in (5, 19)$ depends on if $M \geq 36$. Recall also from Lemma 31 that $\gamma_{\alpha,M} \in (17, 97)$ depends if $M \geq 36$ and if $\alpha = 0$. Comparing cases, we see

$$\rho_{\alpha,M} = \begin{cases} 12(2+\sqrt{2\pi}) \in (54, 55) & \alpha = 0, M < 36 \\ \frac{2}{3}(13 + 2\sqrt{43}) \in (17, 18) & \alpha = 0, M \geq 36 \\ 28 + \sqrt{793} \in (56, 57) & \alpha \neq 0, M < 36 \\ 48 + \sqrt{2353} \in (96, 97) & \alpha \neq 0, M \geq 36. \end{cases} \tag{331}$$

Notice $\rho_{\alpha,M}$ is equal to $\gamma_{\alpha,M}$ except in the case of $\alpha = 0$ and $M < 36$.

$$\mathbb{E}[\|\mathbf{z}_L^Q - \mathbf{z}_L^P\|_2^2] \leq 32\rho_{\alpha,M}^{L-1}LK^2(K^2 \vee 1)^{L-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2\right). \tag{332}$$

Combining terms. Plugging Equations (313) and (332) into Equation (305) and simplifying we have

$$\|\mathbb{E}_Q[\phi_{\text{o}}(\mathbf{z}_L)^2] - \mathbb{E}_P[\phi_{\text{o}}(\mathbf{z}_L)^2]\|_2 \leq 8\sqrt{2}\rho_{\alpha,M}^{L-\frac{1}{2}}L^{1/2}K(K^2 \vee 1)^{L-\frac{1}{2}}\left(\alpha^2 + 1 \vee \tfrac{1}{D_{\text{i}}}\|\mathbf{x}\|_2^2\right). \tag{333}$$

□

[3]Suppose that $x_n \leq b_n + ax_{n-1}$ for $n \geq 2$ and $x_1 \leq b_1$. We then show that $x_n \leq \sum_{n'=1}^n a^{n-n'}b_n$ for all $n \geq 1$. For the base case, note that the case $n = 1$ is true because $x_1 \leq b_1$ For the induction step, suppose that the case $n \in \mathbb{N}$ is true. Then $x_{n+1} \leq b_{n+1} + a\sum_{n'=1}^n a^{n-n'}b_n \leq a^{n+1-(n+1)}b_{n+1} + \sum_{n'=1}^n a^{n+1-n'}b_n = \sum_{n'=1}^{n+1} a^{n+1-n'}b_n$, so the case $n+1$ is also true. We conclude that $x_n \leq \sum_{n'=1}^n a^{n-n'}b_n$ for all $n \geq 1$.

Given this lemma, we can succinctly prove Lemma 24.

Proof of Lemma 24. We have

$$|D_Q^i - D_P^i| = \frac{1}{M}\left|\mathrm{tr}\big((\boldsymbol{\Sigma}_Q - \mathbf{I})\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}] + (\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}] - \mathbb{E}_P[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}])\big)\right| \tag{334}$$

$$\leq \tfrac{1}{M}\left|\mathrm{tr}\big((\boldsymbol{\Sigma}_Q - \mathbf{I})\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}]\big)\right| + \tfrac{1}{M}\left|\mathrm{tr}\big(\mathbb{E}_Q[\phi(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}] - \mathbb{E}_P[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}]\big)\right| \tag{335}$$

Next we bound each of the two terms in Equation (335) separately, starting with the first term. First, note that $\boldsymbol{\Sigma}_Q - \mathbf{I}$ is a diagonal matrix and let $\boldsymbol{\sigma}^2 = \mathrm{diag}(\boldsymbol{\Sigma}_Q)$. Since the trace of a matrix product is the element-wise inner product of the matrices, the first term in Equation (335) becomes

$$\left|\mathrm{tr}\big((\boldsymbol{\Sigma}_Q - \mathbf{I})\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}]\big)\right| = \left|\langle \boldsymbol{\sigma}_Q^2 - \mathbf{1}, \mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]]\rangle\right| \tag{336}$$

$$\leq \|\boldsymbol{\sigma}_Q^2 - \mathbf{1}\|_1 \|\mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\|_\infty \tag{337}$$

$$\leq \sqrt{M}\|\boldsymbol{\sigma}_Q^2 - \mathbf{1}\|_2 \|\mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\|_\infty \tag{338}$$

where the first inequality is an application of Hölder's inequality. Now $\|\boldsymbol{\sigma}_Q^2 - \mathbf{1}\|_2 \leq (2+K)\,K$ by Lemma 9 (iv.), where $K = \sqrt{2\,\mathrm{KL}(Q,P)}$. To upper bound $\|\mathbb{E}_Q[\phi^2(\mathbf{z}_L)]\|_\infty$, notice we can use the oddness and Lipschitz property of ϕ_{o} to write

$$\|\mathbb{E}[\phi_{\mathrm{o}}^2(\mathbf{z}_L^Q)]\|_\infty = \|\mathbb{E}[(\phi_{\mathrm{o}}(\mathbf{z}_L^Q) - \phi_{\mathrm{o}}(\mathbf{0}))^2]\|_\infty \leq \|\mathbb{E}[(\mathbf{z}_L^Q)^2]\|_\infty. \tag{339}$$

We can upper bound this using Lemma 31:

$$\|\mathbb{E}_Q[(\mathbf{z}_L)^2]\|_\infty \leq \left(\gamma_\alpha(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right). \tag{340}$$

Therefore, the first term in Equation (335) is upper bounded by

$$\left|\mathrm{tr}\big((\boldsymbol{\Sigma}_Q - \mathbf{I})\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}]\big)\right| \leq \sqrt{M}\,(2+K)\,K\left(\gamma_\alpha(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right) \tag{341}$$

$$\leq \sqrt{M}\gamma_\alpha^L K\left(K^2 \vee 1\right)^{L+\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right), \tag{342}$$

where we use that $K + 2 \leq 3(K^2 \vee 1)^{1/2}$ in the second inequality.

The second term in Equation (335) can be upper bounded by

$$\left|\mathrm{tr}\big(\mathbb{E}_Q[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}] - \mathbb{E}_P[\phi_{\mathrm{o}}(\mathbf{z}_L)\phi_{\mathrm{o}}(\mathbf{z}_L)^\mathsf{T}]\big)\right| = \left|\langle \mathbf{1}, \mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\rangle\right| \tag{343}$$

$$\leq \|\mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\|_1 \tag{344}$$

$$\leq \sqrt{M}\|\mathbb{E}_Q[\phi_{\mathrm{o}}^2(\mathbf{z}_L)] - \mathbb{E}_P[\phi_{\mathrm{o}}^2(\mathbf{z}_L)]\|_2 \tag{345}$$

$$\leq \sqrt{M}8\sqrt{2}\rho_{\alpha,M}^{L-\frac{1}{2}} L^{1/2} K(K^2 \vee 1)^{L-\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right) \tag{346}$$

where the last inequality follows from Lemma 35.

Plugging the bounds on both terms into Equation (335) and simplifying we have

$$|D_Q^i(\mathbf{x}) - D_P^i(\mathbf{x})| \leq \tfrac{1+8\sqrt{2}}{\sqrt{M}}\rho_{\alpha,M}^L L^{1/2} K(K^2 \vee 1)^{L+\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right) \tag{347}$$

$$\leq \tfrac{16+\sqrt{2}}{\sqrt{M}}\rho_{\alpha,M}^L L^{1/2}\,\mathrm{KL}(Q,P)^{1/2}(2\,\mathrm{KL}(Q,P) \vee 1)^{L+\frac{1}{2}} \left(\alpha^2 + 1 \vee \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2\right) \tag{348}$$

where we use $1 \leq L^{1/2}$, $\rho_{\alpha,M}^{L-\frac{1}{2}} \leq \rho_{\alpha,M}^L$, $\gamma_{\alpha,M} \leq \rho_{\alpha,M}$, and $(K^2 \vee 1)^{L-\frac{1}{2}} \leq (K^2 \vee 1)^{L+\frac{1}{2}}$ to more simply group the terms.

□

E.3 Off-Diagonal Terms

The key result in bounding the off-diagonal is an upper bound on the operator norm of the last layer outer product.

Lemma 36.

$$\|\mathbb{E}_Q[\phi_o(\mathbf{z}_L(\mathbf{x}))\phi_o(\mathbf{z}_L(\mathbf{x}))^\mathsf{T}]\|_2 \leq 24\sqrt{M}\left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right), \tag{349}$$

where $K = \sqrt{\mathrm{KL}(Q,P)}$ and $\gamma_{\alpha,M} \in (17, 97)$ is defined in Lemma 31.

Proof. By Remark 28,

$$\|\mathbb{E}_Q[\phi_o(\mathbf{z}_L(\mathbf{x}))\phi_o(\mathbf{z}_L(\mathbf{x}))^\mathsf{T}]\|_2 \leq 12\sqrt{M}\left(\|\mathbb{E}[\mathbf{z}_L^2]\|_\infty + K^2\left((\alpha^2 + \|\mathbb{E}[\mathbf{z}_{L-1}^2]\|_\infty) \vee 1\right)\right). \tag{350}$$

Note that in the $\alpha = 0$ case we could save a factor of 2 by instead applying Lemma 27, but for simplicity we consider any $\alpha \in \mathbb{R}$ here. We could also reduce the constant by separately considering the case of $M \geq 3$ but we do not for simplicity. We now upper bound each of the two terms above. By Lemma 31,

$$\alpha^2 + \|\mathbb{E}[(\mathbf{z}_L)^2]\|_\infty \leq \left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right). \tag{351}$$

Similarly,

$$K^2\left((\alpha^2 + \|\mathbb{E}[\mathbf{z}_{L-1}^2]\|_\infty) \vee 1\right) \leq K^2\left(\left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^{L-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right) \vee 1\right) \tag{352}$$

$$\leq (K^2 \vee 1)\left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^{L-1}\left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right) \tag{353}$$

$$\leq \left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right), \tag{354}$$

where we use that $\gamma_{\alpha,M} \geq 1$ in the last two inequalities. Adding the two terms together in Equation (350), noting that we drop the α^2 term in the Equation (351) to more simply group the terms, we have

$$|\mathbb{E}_Q[\phi_o(\mathbf{z}_L(\mathbf{x}))\phi_o(\mathbf{z}_L(\mathbf{x}))^\mathsf{T}]\|_2 \leq 24\sqrt{M}\left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right). \tag{355}$$

□

Given Lemma 36, we can prove Lemma 25 using the definition of operator norm.

Proof of Lemma 25. We have

$$|O_Q^i(\mathbf{x})| = \frac{1}{M}|\mathbb{E}_Q[\mathbf{w}_{L+1}]^\mathsf{T}\mathbb{E}_Q[\phi_o(\mathbf{z}_L)\phi_o(\mathbf{z}_L)^\mathsf{T}]\mathbb{E}_Q[\mathbf{w}_{L+1}]| \leq \frac{1}{M}\|\mathbb{E}_Q[\mathbf{w}_{L+1}]\|_2^2\|\mathbb{E}_Q[\phi_o(\mathbf{z}_L)\phi_o(\mathbf{z}_L)^\mathsf{T}]\|_2, \tag{356}$$

by the definition of the operator norm. Lemma 9 tells us that $\|\mathbb{E}_Q[\mathbf{w}_{L+1}]\|_2^2 \leq K^2 = 2\,\mathrm{KL}(Q,P)$. Combining this estimate with Lemma 36 gives

$$|O_Q^i(\mathbf{x})| \leq \tfrac{1}{M}\left(K^2\right)\left(24\sqrt{M}\left(\gamma_{\alpha,M}(K^2 \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right)\right) \tag{357}$$

$$\leq \tfrac{48}{\sqrt{M}}\,\mathrm{KL}(Q,P)\left(\gamma_{\alpha,M}(2\,\mathrm{KL}(Q,P) \vee 1)\right)^L \left(\alpha^2 + 1 \vee \tfrac{1}{D_i}\|\mathbf{x}\|_2^2\right). \tag{358}$$

□

F Quantitative Bounds on the KL divergence for General Likelihoods

The sketch of the proof given to upper bound $\mathrm{KL}(Q^*, P)$ in the main text relied on that any convergent sequence is bounded. Unfortunately, this is not quantitative, in that it does not allow an upper bound on a computable upper bound on $\mathrm{KL}(Q^*, P)$. In this section, we show how a modification of the approach described in Section 4.4 can yield a computable upper bound on $\mathrm{KL}(Q^*, P)$.

We take the same assumptions as in Section 4.4:

(i) The likelihood factorizes over data points, i.e. $\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))$, for some function p;

(ii) there exists a C such that $\log p(\mathbf{y}_n|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)) \leq C$;

(iii) for any fixed $\mathbf{y}_n$, $\log p(\mathbf{y}_n|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))$ can be lower bounded by a quadratic function in $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)$.

Also, the proof follows the same initial approach: by the optimality of Q^*, we have

$$0 \leq \mathrm{ELBO}(Q^*) - \mathrm{ELBO}(P) \tag{359}$$
$$= \mathbb{E}_{\boldsymbol{\theta}\sim Q^*}[\log \mathcal{L}(\boldsymbol{\theta})] - \mathrm{KL}(Q^*, P) - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\log \mathcal{L}(\boldsymbol{\theta})]. \tag{360}$$

Rearranging and using the assumptions on $\log \mathcal{L}(\boldsymbol{\theta})$,

$$\mathrm{KL}(Q^*, P) \leq \mathbb{E}_{\boldsymbol{\theta}\sim Q^*}[\log \mathcal{L}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\log \mathcal{L}(\boldsymbol{\theta})] \tag{361}$$
$$\leq CN - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\textstyle\sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))] \tag{362}$$
$$\leq CN - \mathbb{E}_{\boldsymbol{\theta}\sim P}[\textstyle\sum_{n=1}^{N} h_n(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))] \tag{363}$$

where h_n is quadratic. Since h_n is quadratic, $\mathbb{E}_P[h_n(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n))]$ is a linear combination of the first and second moments of $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)$. As the top layer weight matrix has mean 0, $\mathbb{E}_P[\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)] = 0$ independent of width. We therefore need only to upper bound $\mathbb{E}_P[\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_n)^2]$ independent of width.

It suffices to prove this in the case $D_{\mathrm{o}} = 1$. In the general case, the resulting bound can simply be summed over output dimension. For the variance, note that

$$\mathbb{V}[f(\mathbf{x})] = \tfrac{1}{M}\mathbb{V}[\langle \mathbf{w}_{L+1}, \phi(\mathbf{z}_L)\rangle] + \mathbb{V}[b_{L+1}] \tag{364}$$
$$= \tfrac{1}{M}\mathbb{E}[\mathrm{tr}[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^{\mathsf{T}}\phi(\mathbf{z}_L)\phi^{\mathsf{T}}(\mathbf{z}_L)]] + 1 \tag{365}$$
$$= \tfrac{1}{M}\mathbb{E}[\|\phi(\mathbf{z}_L)\|_2^2] + 1 \tag{366}$$
$$\leq \tfrac{1}{M}\mathbb{E}[\|\mathbf{z}_L\|_2^2] + 1. \tag{367}$$

Similarly,

$$\tfrac{1}{M}\mathbb{E}[\|\mathbf{z}_\ell\|_2^2] = \tfrac{1}{M^2}\mathbb{E}[\mathrm{tr}[\mathbf{W}_\ell\mathbf{W}_\ell^{\mathsf{T}}\phi(\mathbf{z}_{\ell-1})\phi^{\mathsf{T}}(\mathbf{z}_{\ell-1})]] + \tfrac{1}{M}\mathbb{E}[\|\mathbf{b}_\ell\|_2^2] \tag{368}$$
$$= \tfrac{1}{M}\mathbb{E}[\|\phi(\mathbf{z}_{\ell-1})\|_2^2] + 1 \tag{369}$$
$$\leq \tfrac{1}{M}\mathbb{E}[\|\mathbf{z}_{\ell-1}\|_2^2] + 1. \tag{370}$$

Therefore,

$$\mathbb{V}[f(x)] \leq L + 1 + \tfrac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2. \tag{371}$$

Example 37 (Gaussian Likelihood)**.** Consider a Gaussian likelihood so that,

$$\log p(y_n|f(\mathbf{x}_n)) = -\log 2\pi\sigma^2 - \tfrac{1}{2\sigma^2}(y - f(\mathbf{x}_n))^2. \tag{372}$$

We then have $\log p(y_n|f(\mathbf{x}_n)) \leq -\log 2\pi\sigma^2 =: C$. Also,

$$\mathbb{E}_P[\log p(y_n|f(\mathbf{x}_n))] = CN - \frac{\sum_{n=1}^{N} y_n^2 + \mathbb{V}[f(\mathbf{x}_n)]}{2\sigma^2} \tag{373}$$

Hence by Equation (371),

$$\mathrm{KL}(Q^*, P) \leq \frac{(L+1)N + \sum_{n=1}^{N} y_n^2 + \|\mathbf{x}_n\|_2^2}{2\sigma^2}. \tag{374}$$

Example 38 (Student t likelihood)**.** For a Student t likelihood (used in robust regression) with $\nu > 0$ degrees of freedom, we have

$$\log p(y_n|f(\mathbf{x}_n)) = c(\nu) - \frac{\nu+1}{2}\log\left(1 + \frac{(f(\mathbf{x}_n) - y_n)^2}{\nu}\right). \tag{375}$$

As the second term is non-negative, this is upper bounded by $c(\nu)$. Applying $\log(1+a) \leq a$, we have

$$\log p(y_n|f(\mathbf{x}_n)) \geq c(\nu) - \frac{\nu+1}{2}\frac{(f(x_n) - y_n)^2}{\nu} \tag{376}$$

which provides the desired quadratic lower bound.

Example 39 (Logistic likelihood)**.** For a logistic likelihood, we have

$$\log p(y_n|f(\mathbf{x}_n)) = y_n \log\left(\frac{1}{1+e^{-f(\mathbf{x}_n)}}\right) + (1-y_n)\log\left(\frac{e^{-f(\mathbf{x}_n)}}{1+e^{-f(\mathbf{x}_n)}}\right) \tag{377}$$

As both terms are non-positive, we have $\log p(y_n|f(\mathbf{x}_n)) \leq 0$. As $g(a) = \log(1+e^{-a})$ is three times differentiable, we have that for any $a \in \mathbb{R}$, there exists a $\xi_a \in \mathbb{R}$ such that,

$$\log(1+e^{-a}) = \log 2 - \frac{a}{2} + \frac{a^2}{8} + \frac{g^{(3)}(\xi_a)}{3!}a^3. \tag{378}$$

As $\operatorname{sign}(g^{(3)}(\xi_a)) = \operatorname{sign}(\xi_a) = \operatorname{sign}(a) = \operatorname{sign}(a^3)$, the final term is non-negative, so

$$\log(1+e^{-a}) \leq \log 2 - \frac{a}{2} + \frac{a^2}{8}. \tag{379}$$

Hence

$$\log p(y_n = 1|f(\mathbf{x}_n)) \geq -\log 2 + \frac{f(\mathbf{x}_n)}{2} - \frac{f(\mathbf{x}_n)^2}{8} \text{ and} \tag{380}$$

$$\log p(y_n = 0|f(\mathbf{x}_n)) = \log p(y_n = 1|f(\mathbf{x}_n)) - f(\mathbf{x}_n) \geq -\log 2 - \frac{f(\mathbf{x}_n)}{2} - \frac{f(\mathbf{x}_n)^2}{8}. \tag{381}$$

G Proof of Convergence in Distribution of Finite Marginals of the Variational Posterior for Odd Activation Functions

In this section we derive a generalization of Theorem 1 that incorporates a bias.

Theorem 40. Consider N one-dimensional data points $(\mathbf{x}_n, y_n)_{n=1}^N$ ($D_\mathrm{i} \in \mathbb{N}$ and $D_\mathrm{o} = 1$) and let $\overline{y}$ be the average of the observed values. Let Q^* be the optimal mean-field variational posterior for a neural network with L hidden layers and M neurons per hidden layer. Suppose $\phi_\mathrm{e} = \alpha$ for some $\alpha \in \mathbb{R}$ and $\phi\colon \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz. Also suppose that the likelihood is Gaussian with variance parameter σ^2. Then, along any finite-dimensional distribution, as $M \to \infty$, $f \sim Q^*$ converges weakly to the sum of the NNGP and an independent Gaussian with distribution

$$\mathcal{N}\left(\left(\beta \times 1 + (1-\beta) \times \frac{1}{1+\alpha^2+\frac{\sigma^2}{N}}\right)\overline{y}, \frac{\sigma^2}{\sigma^2+N}\right) \quad \text{where} \quad \beta = \frac{\alpha^2}{\alpha^2+\frac{\sigma^2}{N}}. \tag{382}$$

Observe that the mean of the independent Gaussian is a convex combination of the maximum likelihood solution $\overline{y}$ of fitting a univariate Gaussian to the data and the posterior mean $(1+\alpha^2+\sigma^2/N)^{-1}\overline{y}$ of observing the data as noisy observations for the final bias with noise variance $\sigma^2 + N\alpha^2$. The coefficient of the convex combination, $\beta \in [0,1)$, measures the strength of α relative to the observation noise and number of observations. As $\alpha \to \infty$, the mean converges to the maximum likelihood solution; and as $\alpha \to 0$, the mean converges to the posterior mean.

Proof. The proof conceptually proceeds in two steps. In the first conceptual step, we show that the parameters of all layers but the last layer converge to the prior. Intuitively, this means that the overall solution converges to mean-field inference in the Bayesian linear regression model with features $x \mapsto \mathbb{E}[\phi(z_{L,m})(x)]$. In the second conceptual step, we show that this mean-field solution of the Bayesian linear regression model converges to the prior, thereby completing the proof. Whereas conceptually the proof proceeds in two steps, below we split up these two steps into six steps, as follows:

Step 1: We prove that the parameters of the hidden layers converge to the prior.

Step 2: By assuming two claims, Claim 1 and Claim 2, we almost conclude the proof. It only remains to compute the limiting distribution of the final bias.

Step 3: We prove Claim 1.

Step 4: Using Claim 1, we compute the limiting distribution of the final bias.

Step 5: Using Claim 1 and the limiting distribution of the final bias, we prove Claim 2.

Step 6: We reconcile Step 2 with the limiting distribution of the bias to finally conclude the proof.

Throughout the proof and unlike in the theorem statement, we do *not* incorporate α in ϕ, but write $\phi_\alpha = \phi + \alpha$ for the version of ϕ that does include α. Moreover, we write $\boldsymbol{\theta}_{L+1}$ for the parameters of the final layer and let $\boldsymbol{\theta}_{1:L}$ be all remaining parameters, so $\boldsymbol{\theta} = (\boldsymbol{\theta}_{L+1}, \boldsymbol{\theta}_{1:L})$. We decompose the parameters of the final layer as $\boldsymbol{\theta}_{L+1} = (\mathbf{w}_{L+1}, b_{L+1})$, so also $\boldsymbol{\theta} = (\mathbf{w}_{L+1}, b_{L+1}, \boldsymbol{\theta}_{1:L})$.

We will consider the limits $M \to \infty$ and $\mathrm{KL} \to 0$. Using results from earlier sections in the appendix, we establish the following bounds which ignore proportionality constants irrelevant for the limits $M \to \infty$ and $\mathrm{KL} \to 0$:

$$\|\mathbb{E}_Q[\phi(\mathbf{z}_{L+1})]\|_2 \overset{\text{(Lemma 13)}}{\lesssim} \sqrt{\mathrm{KL}(Q_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}})}, \tag{383}$$

$$|\mathbb{E}_Q[\tfrac{1}{\sqrt{M}}\langle \mathbf{w}_{L+1}, \phi(\mathbf{z}_{L+1})\rangle]| \overset{\text{(Theorem 11)}}{\lesssim} \tfrac{1}{\sqrt{M}}\,\mathrm{KL}(Q_{\mathbf{w}_{L+1},\boldsymbol{\theta}_{1:L}}, P_{\mathbf{w}_{L+1},\boldsymbol{\theta}_{1:L}}), \tag{384}$$

$$|\mathbb{E}_Q[b_{L+1}]| \overset{\text{(Lemma 9)}}{\lesssim} \sqrt{\mathrm{KL}(Q_{b_{L+1}}, P_{b_{L+1}})}, \tag{385}$$

$$\tfrac{1}{M}\|\mathbb{E}_Q[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}]\|_2 \overset{\text{(Lemma 36)}}{\lesssim} \tfrac{1}{\sqrt{M}}, \tag{386}$$

$$\|\mathbb{E}_Q[\phi_\alpha(\mathbf{z}_L)^2]\|_\infty \overset{\text{(Lemma 31)}}{\lesssim} 1, \tag{387}$$

$$|\mathbb{E}_Q[\phi(z_{L,m})\phi(z_{L,m'})]| \overset{\text{(Lemmas 31 and 32, } m \neq m')}{\lesssim} \sqrt{\mathrm{KL}(Q_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}})}, \tag{388}$$

$$\tfrac{1}{M}|\mathrm{tr}\left(\mathbb{E}_Q[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^\mathsf{T}](\mathbb{E}_Q[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}] - \mathbb{E}_P[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^\mathsf{T}])\right)| \overset{\text{(Lemma 33)}}{\lesssim} \tfrac{1}{\sqrt{M}}\sqrt{\mathrm{KL}(Q_{\mathbf{w}_{L+1},\boldsymbol{\theta}_{1:L}}, P_{\mathbf{w}_{L+1},\boldsymbol{\theta}_{1:L}})}, \tag{389}$$

$$\tfrac{1}{M}|\mathbb{E}[\langle \mathbf{w}_{L+1}, \alpha\mathbf{1}\rangle\langle \mathbf{w}_{L+1}, \phi(\mathbf{z}_L)\rangle]| \overset{\text{(Lemma 34)}}{\lesssim} \tfrac{1}{\sqrt{M}}\sqrt{\mathrm{KL}(Q_{\mathbf{w}_{L+1},\boldsymbol{\theta}_{1:L}}, P_{\mathbf{w}_{L+1},\boldsymbol{\theta}_{1:L}})}. \tag{390}$$

Recall that any KL divergence between parameters of an optimal variational posterior and the prior is bounded uniformly over M (Appendix F). This means, e.g., that the mean of any weight of any variational posterior can be considered as an unknown but bounded constant (Appendix C).

Step 1 (convergence of hidden layers). Let P be the prior and let Q^* be the optimal mean-field posterior. Decompose the KL divergence as follows:

$$\mathrm{KL}(Q^*, P) = \mathrm{KL}(Q^*_{\boldsymbol{\theta}_{L+1}}, P_{\boldsymbol{\theta}_{L+1}}) + \mathrm{KL}(Q^*_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}}) = \mathbb{E}_{Q^*}[\log p(\mathbf{y} \mid \boldsymbol{\theta})] - \mathrm{ELBO}(Q^*). \tag{391}$$

Let Q' be the modification of Q^* where the distribution of $\boldsymbol{\theta}_{1:L}$ is set to the prior. Then

$$\mathrm{KL}(Q^*_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}}) = \mathbb{E}_{Q^*}[\log p(\mathbf{y} \mid \boldsymbol{\theta})] - \mathrm{ELBO}(Q^*) - \mathrm{KL}(Q^*_{\boldsymbol{\theta}_{L+1}}, P_{\boldsymbol{\theta}_{L+1}}) \tag{392}$$

$$= \mathbb{E}_{Q^*}[\log p(\mathbf{y} \mid \boldsymbol{\theta})] - \mathrm{ELBO}(Q^*) + \mathrm{ELBO}(Q') - \mathbb{E}_{Q'}[\log p(\mathbf{y} \mid \boldsymbol{\theta})]. \tag{393}$$

Therefore, by optimality of Q^*,

$$\mathrm{KL}(Q^*_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}})$$

$$\leq \mathbb{E}_{Q^*}[\log p(\mathbf{y} \mid \boldsymbol{\theta})] - \mathbb{E}_{Q'}[\log p(\mathbf{y} \mid \boldsymbol{\theta})] \tag{394}$$

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mathbb{E}_{Q^*}[(y_n - f(\mathbf{x}_n))^2] - \mathbb{E}_{Q'}[(y_n - f(\mathbf{x}_n))^2]) \tag{395}$$

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mathbb{E}_{Q^*}[f^2(\mathbf{x}_n) - 2y_n f(\mathbf{x}_n)] - \mathbb{E}_{Q^*}[\mathbb{E}_P[f^2(\mathbf{x}_n) - 2y_n f(\mathbf{x}_n) \mid \boldsymbol{\theta}_{L+1}]]) \tag{396}$$

$$\leq \frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(|\mathbb{E}_{Q^*}[f^2(\mathbf{x}_n) - \mathbb{E}_P[f^2(\mathbf{x}_n) \mid \boldsymbol{\theta}_{L+1}]]| + 2|y_n||\mathbb{E}_{Q^*}[f(\mathbf{x}_n) - \mathbb{E}_P[f(\mathbf{x}_n) \mid \boldsymbol{\theta}_{L+1}]]|\Big). \tag{397}$$

Define $\tilde{b} = \frac{1}{\sqrt{M}}\langle \mathbf{w}_{L+1}, \alpha\mathbf{1}\rangle + b_{L+1}$ and let $\tilde{f}(\mathbf{x}_n) = f(\mathbf{x}_n) - \tilde{b}$. Note that $\tilde{b}$ is $\sigma(\boldsymbol{\theta}_{L+1})$-measurable (i.e. $\tilde{b}$ is deterministic after conditioning on the top layer parameters). Rearrange as

$$\begin{aligned}
&\mathbb{E}_{Q^*}[f^2(\mathbf{x}_n) - \mathbb{E}_P[f^2(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}]] \\
&\quad = \mathbb{E}_{Q^*}[\tilde{f}^2(\mathbf{x}_n) + 2\tilde{b}\tilde{f}(\mathbf{x}_n) + \tilde{b}^2 - \mathbb{E}_P[\tilde{f}^2(\mathbf{x}_n) + 2\tilde{b}\tilde{f}(\mathbf{x}_n) + \tilde{b}^2 \,|\, \boldsymbol{\theta}_{L+1}]] \qquad (398)\\
&\quad = \mathbb{E}_{Q^*}[\tilde{f}^2(\mathbf{x}_n) - \mathbb{E}_P[\tilde{f}^2(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}]] + 2\,\mathbb{E}_{Q^*}[\tilde{b}(\tilde{f}(\mathbf{x}_n) - \mathbb{E}_P[\tilde{f}(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}])]. \qquad (399)
\end{aligned}$$

This gives

$$\begin{aligned}
&\mathrm{KL}(Q^*_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}}) \\
&\quad \le \frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(|\mathbb{E}_{Q^*}[\tilde{f}^2(\mathbf{x}_n) - \mathbb{E}_P[\tilde{f}^2(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}]]| + 2|y_n||\mathbb{E}_{Q^*}[\tilde{f}(\mathbf{x}_n) - \mathbb{E}_P[\tilde{f}(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}]]| \\
&\qquad\qquad + 2\,|\mathbb{E}_{Q^*}[\tilde{b}(\tilde{f}(\mathbf{x}_n) - \mathbb{E}_P[\tilde{f}(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}])]|\Big). \qquad (400)
\end{aligned}$$

By oddness of ϕ, it can be seen that $\mathbb{E}_P[\tilde{f}(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}] = 0$. Therefore,

$$\begin{aligned}
&\mathrm{KL}(Q^*_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}}) \\
&\quad \le \frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(|\mathbb{E}_{Q^*}[\tilde{f}^2(\mathbf{x}_n) - \mathbb{E}_P[\tilde{f}^2(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}]]| + 2|y_n||\mathbb{E}_{Q^*}[\tilde{f}(\mathbf{x}_n)]| + 2|\mathbb{E}_{Q^*}[\tilde{b}\tilde{f}(\mathbf{x}_n)]|\Big). \qquad (401)
\end{aligned}$$

Here

$$\mathbb{E}_{Q^*}[\tilde{f}^2(\mathbf{x}_n) - \mathbb{E}_P[\tilde{f}^2(\mathbf{x}_n) \,|\, \boldsymbol{\theta}_{L+1}]] = \tfrac{1}{M}\,\mathrm{tr}\,\mathbb{E}_{Q^*}[\mathbf{w}_{L+1}\mathbf{w}_{L+1}^{\mathsf{T}}](\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^{\mathsf{T}}] - \mathbb{E}_P[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^{\mathsf{T}}]), \qquad (402)$$

which is $O(1/\sqrt{M})$ by Equation (389), and

$$|\mathbb{E}_{Q^*}[\tilde{b}\tilde{f}(x_n)]| \le |\mathbb{E}_{Q^*}[b_{L+1}]||\mathbb{E}_{Q^*}[\tilde{f}(x_n)]| + \tfrac{1}{M}|\mathbb{E}_{Q^*}[\langle \mathbf{w}_{L+1}, \alpha\mathbf{1}\rangle\langle \mathbf{w}_{L+1}, \phi(\mathbf{z}_L)\rangle]|, \qquad (403)$$

which is $O(1/\sqrt{M})$ by Equations (384) and (385) applied to the first term and Equation (390) applied to the second term. Therefore, $\mathrm{KL}(Q^*_{\boldsymbol{\theta}_{1:L}}, P_{\boldsymbol{\theta}_{1:L}}) = O(1/\sqrt{M})$.

Step 2 (beginning of conclusion of proof). Let

$$d_M = \sum_{n=1}^{N}(y_n - \mathbb{E}_{Q^*}[b_{L+1}]). \qquad (404)$$

Although d_M depends on $Q^*_{b_{L+1}}$ which in turn depends on M, note that $\mathbb{E}_{Q^*}[b_{L+1}]$ and therefore d_M can be treated like unknown but bounded constants. Set $c_M = \frac{\alpha d_M}{\alpha^2 N + \sigma^2}$. We make two claims:

$$\mathrm{KL}(Q^*_{\mathbf{w}_{L+1} - \frac{c_M}{\sqrt{M}}\mathbf{1}}, P_{\mathbf{w}_{L+1}}) \to 0, \qquad \text{(Claim 1)}$$

$$c_M \to c_\infty \qquad \text{(Claim 2)}$$

where $Q^*_{\mathbf{w}_{L+1} - \frac{c_M}{\sqrt{M}}\mathbf{1}}$ is the distribution of $\mathbf{w}_{L+1} - \frac{c_M}{\sqrt{M}}\mathbf{1}$ under Q^* and $c_\infty \in \mathbb{R}$ is some constant. The claims will be proven in the next parts. Assuming the claims, denote $\mathbf{w}'_{L+1} = \mathbf{w}_{L+1} - \frac{c_M}{\sqrt{M}}\mathbf{1}$ and $\phi_\alpha = \phi + \alpha$ and decompose

$$\tfrac{1}{\sqrt{M}}\langle \mathbf{w}_{L+1}, \phi_\alpha(\mathbf{z}_L)\rangle = \tfrac{1}{\sqrt{M}}\langle \mathbf{w}'_{L+1}, \phi_\alpha(\mathbf{z}_L)\rangle + \tfrac{c_M}{M}\langle \mathbf{1}, \phi_\alpha(\mathbf{z}_L)\rangle. \qquad (405)$$

By Chebyshev's inequality, we have $\frac{c_M}{M}\langle \mathbf{1}, \phi(\mathbf{z}_L)\rangle \to 0$ in probability under Q^*:

$$\begin{aligned}
\mathbb{E}_{Q^*}[\tfrac{1}{M}\langle \mathbf{1}, \phi(\mathbf{z}_L)\rangle] &\le \tfrac{1}{\sqrt{M}}\|\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)]\|_2 \qquad (406)\\
\mathbb{V}_{Q^*}[\tfrac{1}{M}\langle \mathbf{1}, \phi(\mathbf{z}_L)\rangle] &= \tfrac{1}{M^2}\langle \mathbf{1}, \mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^{\mathsf{T}}]\mathbf{1}\rangle - \tfrac{1}{M^2}\langle \mathbf{1}, \mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)]\rangle^2 \qquad (407)\\
&\le \tfrac{1}{M}\|\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)\phi(\mathbf{z}_L)^{\mathsf{T}}]\|_2 + \tfrac{1}{M}\|\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)]\|_2^2, \qquad (408)
\end{aligned}$$

which are both $O(1/\sqrt{M})$ using Equations (383) and (386). Since $\frac{c_M}{M}\langle \mathbf{1}, \phi(\mathbf{z}_L)\rangle \to 0$ in probability, $\frac{c_M}{M}\langle \mathbf{1}, \phi_\alpha(\mathbf{z}_L)\rangle = \frac{c_M}{M}\langle \mathbf{1}, \phi(\mathbf{z}_L)\rangle + \alpha c_M \to \alpha c_\infty$ in probability by Claim 2.

Suppressing the dependence on $\mathbf{x}$, define $g_\theta = \frac{1}{\sqrt{M}}\langle \mathbf{w}'_{L+1}, \phi_\alpha(\mathbf{z}_L)\rangle$ and note that it is a deterministic function of $(\mathbf{w}'_{L+1}, \boldsymbol{\theta}_{1:L})$. Also write $f_\theta = \frac{1}{\sqrt{M}}\langle \mathbf{w}_{L+1}, \phi_\alpha(\mathbf{z}_L)\rangle$ and note that it the *same* deterministic function of $(\mathbf{w}_{L+1}, \boldsymbol{\theta}_{1:L})$. (That it is the same deterministic function will allow us to use the data processing inequality below.) Let $I \in \mathbb{N}$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_I) \in (\mathbb{R}^{D_i})^I$. Let $Q_g^{(M)}$ be the finite-dimensional distribution of g at $\mathbf{X}$ under the optimal mean-field solution Q^* at width M and let $P_f^{(M)}$ be the finite-dimensional distribution of f at $\mathbf{X}$ under the prior P at width M. Using Pinsker's inequality and the data processing inequality for KL-divergences we have

$$\mathrm{TV}(Q_g^{(M)}, P_f^{(M)}) \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}(Q_g^{(M)}, P_f^{(M)})} \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}(Q^{(M)}_{\mathbf{w}'_{L+1}, \boldsymbol{\theta}_{1:L}}, P^{(M)}_{\mathbf{w}_{L+1}, \boldsymbol{\theta}_{1:L}})}, \tag{409}$$

which goes to zero as $M \to \infty$ by the previous part and Claim 1. Let d be the Lévy–Prokhorov metric on $\mathbb{R}^I$ with the Borel σ-algebra. Since $\mathbb{R}^I$ is separable, the Lévy–Prokhorov metric metrizes weak convergence. Moreover, since $\mathbb{R}^I$ is separable, the Lévy–Prokhorov metric is upper bounded by the total variation distance. By triangle inequality we then bound the distance between $Q_g^{(M)}$ and the NNGP, which we denote by P_{NN}:

$$d(Q_g^{(M)}, P_{\mathrm{NN}}) \le d(Q_g^{(M)}, P_f^{(M)}) + d(P_f^{(M)}, P_{\mathrm{NN}}) \le \mathrm{TV}(Q_g^{(M)}, P_f^{(M)}) + d(P_f^{(M)}, P_{\mathrm{NN}}). \tag{410}$$

As $M \to \infty$, the first term converges to zero by Equation (409) and the second term converges to zero because $P_f^{(M)}$ converges to the NNGP (Matthews et al., 2018). Hence, $Q_g^{(M)}$ converges weakly to the NNGP, P_{NN}. Since $f = g + \frac{c_M}{M}\langle \mathbf{1}, \phi_\alpha(\mathbf{z}_L)\rangle$ and we previously showed that $\frac{c_M}{M}\langle \mathbf{1}, \phi_\alpha(\mathbf{z}_L)\rangle \to \alpha c_\infty$ in probability under Q^*, we conclude that, along any finite-dimensional distribution, f converges to the NNGP plus the constant αc_∞. It remains to add the limiting distribution of the bias, which we do in the last step of the proof.

Step 3 (proof of Claim 1). We previously claimed that

$$\mathrm{KL}(Q^*_{\mathbf{w}_{L+1} - \frac{c_M}{\sqrt{M}}\mathbf{1}}, P_{\mathbf{w}_{L+1}}) \to 0. \tag{411}$$

We now prove this claim. Let $q^*(\boldsymbol{\theta}) = q^*(\mathbf{w}_{L+1})q^*(b_{L+1})q^*(\boldsymbol{\theta}_{1:L})$ be the density of Q^* w.r.t. the Lebesgue measure. Let $\mathcal{L}(q(\mathbf{w}_{L+1}), q(b_{L+1}), q(\boldsymbol{\theta}_{1:L}))$ be the ELBO:

$$\begin{aligned}&\mathcal{L}(q(\mathbf{w}_{L+1}), q(b_{L+1}), q(\boldsymbol{\theta}_{1:L}))\\ &\quad= \mathbb{E}_q[\log p(\mathbf{y} \mid \boldsymbol{\theta})] - \mathrm{KL}(q(\mathbf{w}_{L+1}), p(\mathbf{w}_{L+1})) - \mathrm{KL}(q(b_{L+1}), p(b_{L+1})) - \mathrm{KL}(q(\boldsymbol{\theta}_{1:L}), p(\boldsymbol{\theta}_{1:L})).\end{aligned} \tag{412}$$

Because Q^* is optimal, $q^*(\mathbf{w}_{L+1})$ maximizes the function $q(\mathbf{w}_{L+1}) \mapsto \mathcal{L}(q(\mathbf{w}_{L+1}), q^*(b_{L+1}), q^*(\boldsymbol{\theta}_{1:L}))$. We therefore parametrize $q(\mathbf{w}_{L+1}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\nu}))$ and set the gradients of $(\boldsymbol{\mu}, \boldsymbol{\nu}) \mapsto \mathcal{L}(q(\mathbf{w}_{L+1}), q^*(b_{L+1}), q^*(\boldsymbol{\theta}_{1:L}))$ to zero to find equations which characterize the mean and variance of $q^*(\mathbf{w}_{L+1}) = \mathcal{N}(\boldsymbol{\mu}^*, \mathrm{diag}(\boldsymbol{\nu}^*))$. Consider the joint density $q(\boldsymbol{\theta}) = q(\mathbf{w}_{L+1})q^*(b_{L+1})q^*(\boldsymbol{\theta}_{1:L})$. Denote $\phi_\alpha = \phi + \alpha$. Compute

$$\begin{aligned}&\mathbb{E}_q[\log p(\mathbf{y} \mid \boldsymbol{\theta})] - \mathrm{KL}(q(\mathbf{w}_{L+1}), p(\mathbf{w}_{L+1}))\\ &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(\tfrac{1}{M}\langle \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \mathrm{diag}(\boldsymbol{\nu}), \mathbb{E}_{Q^*}[\phi_\alpha(\mathbf{z}_L)\phi_\alpha(\mathbf{z}_L)^\mathsf{T}]\rangle - 2\tfrac{1}{\sqrt{M}}\langle \boldsymbol{\mu}, \mathbb{E}_{Q^*}[(y_n - b_{L+1})\phi_\alpha(\mathbf{z}_L)]\rangle + \mathbb{E}_{Q^*}[y_n - b_{L+1}]^2\Big)\\ &\quad - \frac{1}{2}\|\boldsymbol{\mu}\|_2^2 - \frac{1}{2}\sum_{m=1}^{M}(\nu_m - 1 - \log(\nu_m)).\end{aligned} \tag{413}$$

Denote

$$\mathbf{A} = \frac{1}{M}\sum_{n=1}^{N}\mathbb{E}_{Q^*}[\phi_\alpha(\mathbf{z}_L(\mathbf{x}_n))\phi_\alpha(\mathbf{z}_L(\mathbf{x}_n))^\mathsf{T}], \quad \mathbf{b} = \frac{1}{\sqrt{M}}\sum_{n=1}^{N}(y_n - \mathbb{E}_q[b_{L+1}])\mathbb{E}_{Q^*}[\phi_\alpha(\mathbf{z}_L(\mathbf{x}_n))]. \tag{414}$$

Then the part of the ELBO depending on $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ can be written as

$$-\frac{1}{2\sigma^2}\big[\langle \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \mathrm{diag}(\boldsymbol{\nu}), \mathbf{A}\rangle - 2\langle \boldsymbol{\mu}, \mathbf{b}\rangle + (y_n - \mathbb{E}_{Q^*}[b_{L+1}])^2\big] - \frac{1}{2}\|\boldsymbol{\mu}\|_2^2 - \frac{1}{2}\sum_{m=1}^{M}(\nu_m - 1 - \log(\nu_m)). \tag{415}$$

Setting the gradient with respect to $\boldsymbol{\mu}$ to zero gives

$$-\frac{1}{\sigma^2}\mathbf{A}\boldsymbol{\mu}^* + \frac{1}{\sigma^2}\mathbf{b} - \boldsymbol{\mu}^* = 0 \implies \boldsymbol{\mu}^* = (\mathbf{A} + \sigma^2\mathbf{I})^{-1}\mathbf{b}. \tag{416}$$

Similarly, setting the gradient with respect ν_m to zero gives

$$-\frac{1}{2\sigma^2}A_{mm} - \frac{1}{2} + \frac{1}{2}\frac{1}{\nu_m^*} = 0 \implies \nu_m^* = \frac{1}{1 + \frac{1}{\sigma^2}A_{mm}}. \tag{417}$$

Therefore,

$$\mathrm{KL}(Q^*_{\mathbf{w}_{L+1} - \frac{c_M}{\sqrt{M}}\mathbf{1}}, P_{\mathbf{w}_{L+1}}) = \frac{1}{2}\|\boldsymbol{\mu}^* - \tfrac{c_M}{\sqrt{M}}\mathbf{1}\|_2^2 + \frac{1}{2}\sum_{m=1}^{M}(\nu_m^* - 1 - \log(\nu_m^*)) \tag{418}$$

$$= \frac{1}{2\sigma^2}\|\boldsymbol{\mu}^* - \tfrac{c_M}{\sqrt{M}}\mathbf{1}\|_2^2 + \frac{1}{2}\sum_{m=1}^{M}\left(\log(1 + \tfrac{1}{\sigma^2}A_{mm}) - \frac{\frac{1}{\sigma^2}A_{mm}}{1 + \frac{1}{\sigma^2}A_{mm}}\right) \tag{419}$$

$$\overset{\text{(i)}}{\leq} \frac{1}{2\sigma^2}\|\boldsymbol{\mu}^* - \tfrac{c_M}{\sqrt{M}}\mathbf{1}\|_2^2 + \frac{1}{2}M\left(\log(1 + \tfrac{1}{M}\tfrac{M}{\sigma^2}A_{m^*}) - \frac{\frac{1}{M}\frac{M}{\sigma^2}A_{m^*}}{1 + \frac{1}{M}\frac{M}{\sigma^2}A_{m^*}}\right) \tag{420}$$

$$\overset{\text{(ii)}}{\leq} \frac{1}{2\sigma^2}\|\boldsymbol{\mu}^* - \tfrac{c_M}{\sqrt{M}}\mathbf{1}\|_2^2 + \frac{1}{2\sigma^2}\frac{M}{\sigma^2}A_{m^*}^2 \tag{421}$$

where in (i) $A_{m^*} = \max_{m\in[M]} A_{mm}$ and we use that $a - 1 - \log(a)$ is increasing in a for $a < 1$ and that $\nu_m^* < 1$. In (ii) we used the inequality[4]

$$x\left(\log(1 + \tfrac{c}{x}) - \frac{\frac{c}{x}}{1 + \frac{c}{x}}\right) \leq \frac{c^2}{x} \quad \text{for all } c \geq 0 \text{ and } x > 0 \tag{422}$$

with $c = \frac{M}{\sigma^2}A_{m^*} \geq 0$ and $x = M > 0$. By Equation (387), $MA_{m^*}^2 = O(1/M)$, so the claim is shown if

$$\|(\mathbf{A} + \sigma^2\mathbf{I})^{-1}\mathbf{b} - \tfrac{c_M}{\sqrt{M}}\mathbf{1}\|_2 \to 0. \tag{423}$$

Using that

$$\|(\mathbf{A} + \sigma^2\mathbf{I})^{-1}\mathbf{b} - \tfrac{c_M}{\sqrt{M}}\mathbf{1}\|_2 \leq \|(\mathbf{A} + \sigma^2\mathbf{I})^{-1}\|_2\|\mathbf{b} - \tfrac{c_M}{\sqrt{M}}(\mathbf{A} + \sigma^2\mathbf{I})\mathbf{1}\|_2 \leq \sigma^{-2}\|\mathbf{b} - \tfrac{c_M}{\sqrt{M}}(\mathbf{A} + \sigma^2\mathbf{I})\mathbf{1}\|_2, \tag{424}$$

it suffices to show that $\sqrt{M}\|\mathbf{b} - \frac{c_M}{\sqrt{M}}(\mathbf{A} + \sigma^2\mathbf{I})\mathbf{1}\|_\infty \to 0$:

$$\max_{m\in[M]}\left|\sum_{n=1}^{N}(y_n - \mathbb{E}_q[b_{L+1}])\mathbb{E}_{Q^*}[\phi_\alpha(z_{L,m}(\mathbf{x}_n))]\right. \tag{425}$$
$$\left. - \frac{\alpha d_M}{\alpha^2 N + \sigma^2}\left(\sigma^2 + \frac{1}{M}\sum_{m'=1}^{M}\sum_{n=1}^{N}\mathbb{E}_{Q^*}[\phi_\alpha(z_{L,m}(\mathbf{x}_n))\phi_\alpha(z_{L,m'}(\mathbf{x}_n))]\right)\right| \to 0.$$

Expand

$$\frac{1}{M}\sum_{m'=1}^{M}\sum_{n=1}^{N}\mathbb{E}_{Q^*}[\phi_\alpha(z_{L,m}(\mathbf{x}_n))\phi_\alpha(z_{L,m'}(\mathbf{x}_n))]$$
$$= \frac{1}{M}\sum_{m'=1}^{M}\sum_{n=1}^{N}(\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))\phi(z_{L,m'}(\mathbf{x}_n))] + \alpha(\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))] + \mathbb{E}_{Q^*}[\phi(z_{L,m'}(\mathbf{x}_n))]) + \alpha^2) \tag{426}$$

$$= \frac{1}{M}\sum_{n=1}^{N}\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))^2] + \frac{1}{M}\sum_{m'=1}^{M}\sum_{n=1}^{N}\alpha^2 \tag{427}$$
$$+ \frac{1}{M}\sum_{m'\neq m}^{M}\sum_{n=1}^{N}\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))\phi(z_{L,m'}(\mathbf{x}_n))] + \frac{1}{M}\sum_{m'=1}^{M}\sum_{n=1}^{N}\alpha(\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))] + \mathbb{E}_{Q^*}[\phi(z_{L,m'}(\mathbf{x}_n))]).$$

[4]Note that the inequality is equivalent to $\log(1 + x) - \frac{x}{1+x} \leq x^2$ for all $x > 0$, which follows from $\log(1 + x) < x$ for all $x > 0$.

Collecting the first two terms of the expansion, note that, by definition of d_M and Equation (387),

$$\max_{m\in[M]}\left|\sum_{n=1}^{N}(y_n-\mathbb{E}_{Q^*}[b_{L+1}])\alpha-\frac{\alpha d_M}{\alpha^2N+\sigma^2}\left(\sigma^2+\frac{1}{M}\sum_{n=1}^{N}\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))^2]+\frac{1}{M}\sum_{n=1}^{N}\sum_{m'=1}^{M}\alpha^2\right)\right|\to 0. \tag{428}$$

It therefore remains to show that the remainder also goes, which follows from the following three limits:

$$\max_{m\in[M]}\left|\sum_{n=1}^{N}(y_n-\mathbb{E}_{Q^*}[b_{L+1}])\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))]\right|\overset{\text{Equations (383) and (385) with } \mathrm{KL}(Q^*_{\theta_{1:L}},P_{\theta_{1:L}})\to 0}{\to} 0, \tag{429}$$

$$\max_{m\in[M]}\left|\frac{1}{M}\sum_{n=1}^{N}\sum_{m'\neq m}^{M}\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))\phi(z_{L,m'}(\mathbf{x}_n))]\right|\overset{\text{Equation (388) with } \mathrm{KL}(Q^*_{\theta_{1:L}},P_{\theta_{1:L}})\to 0}{\to} 0, \tag{430}$$

$$\max_{m\in[M]}\left|\frac{\alpha}{M}\sum_{n=1}^{N}\sum_{m'=1}^{M}\mathbb{E}_{Q^*}[\phi(z_{L,m}(\mathbf{x}_n))+\phi(z_{L,m'}(\mathbf{x}_n))]\right|\overset{\text{Equation (383) with } \mathrm{KL}(Q^*_{\theta_{1:L}},P_{\theta_{1:L}})\to 0}{\to} 0. \tag{431}$$

Step 4 (convergence of the bias). The starting point is to note that

$$\begin{aligned}&\tfrac{1}{\sqrt{M}}\mathbb{E}_{Q^*}[\langle\mathbf{w}_{L+1},\phi_\alpha(\mathbf{z}_L)\rangle]\\&\quad=\tfrac{1}{\sqrt{M}}\langle\mathbb{E}_{Q^*}[\mathbf{w}_{L+1}-\tfrac{c_M}{\sqrt{M}}\mathbf{1}],\mathbb{E}_{Q^*}[\phi_\alpha(\mathbf{z}_L)]\rangle+\tfrac{c_M}{M}\langle\mathbf{1},\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)]\rangle+\alpha c_M\end{aligned} \tag{432}$$

where

$$\alpha c_M=\frac{\alpha^2 d_M}{\alpha^2N+\sigma^2}=\frac{\alpha^2}{\alpha^2N+\sigma^2}\sum_{n=1}^{N}(y_n-\mathbb{E}_{Q^*}[b_{L+1}])=:\beta(\overline{y}-\mathbb{E}_{Q^*}[b_{L+1}]) \tag{433}$$

with

$$\overline{y}=\frac{1}{N}\sum_{n=1}^{N}y_n,\qquad \beta=\frac{\alpha^2}{\alpha^2+\sigma^2/N}. \tag{434}$$

Moreover,

$$\begin{aligned}|\tfrac{1}{\sqrt{M}}\langle\mathbb{E}_{Q^*}[\mathbf{w}_{L+1}-\tfrac{c_M}{\sqrt{M}}\mathbf{1}],\mathbb{E}_{Q^*}[\phi_\alpha(\mathbf{z}_L)]\rangle|&\le\tfrac{1}{\sqrt{M}}(2\,\mathrm{KL}(Q^*_{\mathbf{w}_{L+1}-\frac{c_M}{\sqrt{M}}\mathbf{1}},P_{\mathbf{w}_{L+1}}))^{1/2}\|\mathbb{E}_{Q^*}[\phi_\alpha(\mathbf{z}_L)]\|_2\\&\le\tfrac{1}{\sqrt{M}}(2\,\mathrm{KL}(Q^*_{\mathbf{w}_{L+1}-\frac{c_M}{\sqrt{M}}\mathbf{1}},P_{\mathbf{w}_{L+1}}))^{1/2}(\sqrt{M}|\alpha|+\|\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)]\|_2),\end{aligned}$$

which is $o(1)$ by the Claim 1 and Equation (383); and

$$|\tfrac{c_M}{M}\langle\mathbf{1},\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)]\rangle|\le\tfrac{c_M}{\sqrt{M}}\|\mathbb{E}_{Q^*}[\phi(\mathbf{z}_L)]\|_2,$$

which is also $o(1)$ by Equation (383). We conclude that

$$\tfrac{1}{\sqrt{M}}\mathbb{E}_{Q^*}[\langle\mathbf{w}_{L+1},\phi_\alpha(\mathbf{z}_L)\rangle]=\beta(\overline{y}-\mathbb{E}_{Q^*}[b_{L+1}])+o(1). \tag{435}$$

We proceed like in the second part: Because Q^* is optimal, $q^*(b_{L+1})$ maximizes the function $q(b_{L+1})\mapsto\mathcal{L}(q^*(\mathbf{w}_{L+1}),q(b_{L+1}),q^*(\boldsymbol{\theta}_{1:L}))$. We therefore parametrize $q(b_{L+1})=\mathcal{N}(\mu_M,\nu_M)$ and set the gradients of $(\mu_M,\nu_M)\mapsto\mathcal{L}(q(\mathbf{w}^*_{L+1}),q(b_{L+1}),q^*(\boldsymbol{\theta}_{1:L}))$ to zero to find equations which characterize the mean and variance of $q^*(b_{L+1})=\mathcal{N}(\mu^*_M,\nu^*_M)$. Consider the joint density $q(\boldsymbol{\theta})=q^*(\mathbf{w}_{L+1})q(b_{L+1})q^*(\boldsymbol{\theta}_{1:L})$. Denote $\phi_\alpha=\phi+\alpha$. Compute

$$\begin{aligned}&\mathbb{E}_q[\log p(\mathbf{y}\mid\boldsymbol{\theta})]-\mathrm{KL}(q(b_{L+1}),p(b_{L+1}))\\&\quad=-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(\nu_M+\mu_M^2-2\mu_M(y_n-\tfrac{1}{\sqrt{M}}\mathbb{E}_{Q^*}[\langle\mathbf{w}_{L+1},\phi_\alpha(\mathbf{z}_L)\rangle])+\mathbb{E}_{Q^*}[(y_n-\tfrac{1}{\sqrt{M}}\langle\mathbf{w}_{L+1},\phi_\alpha(\mathbf{z}_L)\rangle)^2]\Big)\\&\qquad-\frac{1}{2}\mu_M^2-\frac{1}{2}(\nu_M-1-\log(\nu_M))\end{aligned} \tag{436}$$

$$\begin{aligned}&=-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big[\nu_M+\mu_M^2-2\mu_M(y_n-\beta(\overline{y}-\mu^*_M)+o(1))+\mathbb{E}_{Q^*}[(y_n-\tfrac{1}{\sqrt{M}}\langle\mathbf{w}_{L+1},\phi_\alpha(\mathbf{z}_L)\rangle)^2]\Big]\\&\qquad-\frac{1}{2}\mu_M^2-\frac{1}{2}(\nu_M-1-\log(\nu_M)).\end{aligned} \tag{437}$$

Setting the derivative w.r.t. ν_M to zero and solving gives $\nu_M^* = \sigma^2/(\sigma^2+N)$. Setting the derivative w.r.t. μ_M to zero, we find

$$0 = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}[2\mu_M^* - 2(y_n - \beta(\overline{y} - \mu_M^*) + o(1))] - \mu_M^* \tag{438}$$

$$= -\frac{N}{2\sigma^2}[2\mu_M^* - 2(y_n - \beta(\overline{y} - \mu_M^*))] - \mu_M^* + o(1) \tag{439}$$

$$= -\frac{N}{\sigma^2}[\mu_M^* - (\overline{y} - \beta(\overline{y} - \mu_M^*))] - \mu_M^* + o(1) \tag{440}$$

$$= -\frac{N}{\sigma^2}[(1-\beta)\mu_M^* - (1-\beta)\overline{y}] - \mu_M^* + o(1) \tag{441}$$

$$= -(1-\beta)\frac{N}{\sigma^2}[\mu_M^* - \overline{y}] - \mu_M^* + o(1) \tag{442}$$

$$= -\frac{1}{\alpha^2 + \frac{\sigma^2}{N}}[\mu_M^* - \overline{y}] - \mu_M^* + o(1) \tag{443}$$

$$\implies \mu_M^* = \frac{1}{(\alpha^2 + \frac{\sigma^2}{M})^{-1} + 1}\left(\frac{1}{\alpha^2 + \frac{\sigma^2}{M}}\overline{y} + o(1)\right). \tag{444}$$

Therefore, taking $M \to \infty$, under Q^*,

$$b_{L+1} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\frac{1}{1+\alpha^2+\frac{\sigma^2}{N}}\overline{y}, \frac{\sigma^2}{\sigma^2+N}\right). \tag{445}$$

Step 5 (proof of Claim 2). By the previous step, we have that

$$\alpha c_M = \beta(\overline{y} - \mu_M^*) \tag{446}$$

$$\to \beta\left(1 - \frac{1}{1+\alpha^2+\frac{\sigma^2}{N}}\right)\overline{y} \tag{447}$$

$$= c_\infty, \tag{448}$$

which proves Claim 2.

Step 6 (end of conclusion of proof). In Step 2, we showed that, under Q^*, along any finite-dimensional distribution, $\frac{1}{M}\langle \mathbf{w}_{L+1}, \phi_\alpha(\mathbf{z}_L)\rangle$ converges to the NNGP plus the constant αc_∞. We now add the limiting distribution of the final bias to this, which concludes the proof. For this, we note that αc_∞ simply adds to the mean of the bias:

$$\alpha c_M + \mu_M^* \to \left(\beta \times 1 + (1-\beta) \times \frac{1}{1+\alpha^2+\frac{\sigma^2}{N}}\right)\overline{y}, \tag{449}$$

which agrees with the theorem statement. □

H Example Showing that the Mean of the Variational Posterior Need not Converge if Activation Functions are not Odd

We begin by recalling that any function $\phi : \mathbb{R} \to \mathbb{R}$ can be decomposed into the sum of an even and an odd function in a (unique) way. In particular, we have,

$$\phi(a) = \frac{\phi(a) + \phi(-a)}{2} + \frac{\phi(a) - \phi(-a)}{2}. \tag{450}$$

We define $\phi_e(a) = \frac{\phi(a)+\phi(-a)}{2}$ and $\phi_o(a) = \frac{\phi(a)-\phi(-a)}{2}$ to be the even and odd parts of ϕ respectively.

The main goal of this section is to prove the following theorem, which shows that for certain activations including ReLU, the variational posterior need not converge to the prior as the width tends to infinity.

Theorem 41. We consider a Bayesian neural network prior with a standard Gaussian distribution over the weights defines in Equations (1) to (3). Let $\mathcal{Q}$ denote the set of all mean field variational distributions over feed-forward neural networks with activation function ϕ, L hidden layers, and M neurons per layer. Assume the following conditions:

i. ϕ is continuous.
ii. $\phi(a) = O(|a| + 1)$. This condition is equivalent to the linear envelop condition Matthews et al. (2018).
iii. There exist $a, a' \in \mathbb{R}$ such that $\phi_e(a) \neq \phi_e(a')$. In words, it is not true that ϕ is equal to the sum of an odd function and a constant.
iv. There exists an $a \in \mathbb{R}$ such that $\phi(a)^2 + \phi(-a)^2 \neq 2\phi(0)^2$. This is equivalent to ϕ^2 is not equal to an odd function plus a constant.

For a dataset $D = (\mathbf{x}_n, y_n)_{\mathrm{n}=1}^N$, with $y_n \in \mathbb{R}^{D_\mathrm{i}}$ and a given homoscedastic Gaussian likelihood let $Q^* \in \mathcal{Q}$ be the optimal variational posterior for this dataset and likelihood. Then, there exists a dataset D and a homoscedastic Gaussian likelihood such that Q^* satisfies

$$|\mathbb{E}_{Q^*}[f_\theta(\mathbf{x})] - \mathbb{E}_{Q^*}[f_\theta(\mathbf{x}')]| \geq c \tag{451}$$

where c is a constant independent of M.

Our proof strategy will be to find a sequence of variational distributions (indexed by M) with a mean function that does not tend to a constant, and show that there exists a dataset such that sequence of ELBOs defined by this sequence converges to a number that is higher than the ELBOs of any sequence of variational distributions that have a mean that tends to a constant function.

To this end, we introduce the following sequence of distributions, that will serve as our candidate set of distributions with non-constant means and 'good' ELBOs:

Definition 42. For $C \in \mathbb{R}$, define Q^C to be the mean field Gaussian distribution over weights of a neural network with L hidden layers and M neurons such that $\mathbf{W}_{\mathrm{L\text{-}1}}, \ldots, \mathbf{W}_0, \mathbf{b}_\mathrm{L}, \ldots, \mathbf{b}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{W}_\mathrm{L} \sim \mathcal{N}(\frac{\sqrt{C}}{\sqrt{M}}\mathbf{1}, \mathbf{I})$.

H.1 Sketch of Construction

We now sketch the ideas behind the counterexample. The KL divergence $\mathrm{KL}(Q^C, P) = \frac{C}{2}$ (Proposition 44). Hence it suffices to show that Q^C has a expected log likelihood term at least $\frac{C}{2}$ better than any constant predictor. We note that under Q^C, the expected value of each post-activation in the last hidden layer will be the same by exchangeability. In the case of odd activations, by symmetry this was 0, but for other activations this expectation will generally depend on $\mathbf{x}$. By choosing the final weight layer to be parallel to $\mathbf{1}$, we will make the variation in the mean as large as possible (as the mean of the last layer is parallel to the expected value of the post-activations). If the y values happen to fall on this line, Q^C will obtain a much better mean-square error than any constant predictor. We can upper bound the variance of Q^C at the data, as this is the same as the prior variance. The proof is then completed by choosing values for parameters to show that Q^C is better than any hypothetical variational approximation with near constant mean.

H.2 Preliminary Definitions and Results

In this section, we define several quantities and state the necessary preliminaries to construct the counter-example. We include the proofs when they are brief, but defer the proof of Proposition 46, which is more involved until after constructing the counterexample.

We first define a function to represent the expected value of the post-activations in the final hidden layer,

Definition 43. Define $\lambda_M \colon \mathbb{R}^{D_\mathrm{i}} \to \mathbb{R}$ by

$$\lambda_M(\mathbf{x}) = \mathbb{E}_P[\phi(\mathbf{z}_{L-1}^M)]. \tag{452}$$

Further, define

$$\lambda(\mathbf{x}) = \mathbb{E}[\mathbf{z}_{L-1}(\mathbf{x})], \qquad \mathbf{z}_{L-1} \sim \mathcal{N}(0, k^{L-1}(\mathbf{x}, \mathbf{x})) \tag{453}$$

where k^{L-1} is defined by the recursion,

$$k^0(\mathbf{x},\mathbf{x}) = 1 + \|\mathbf{x}\|_2^2, \qquad k^\ell(\mathbf{x},\mathbf{x}) = 1 + h(k^{\ell-1}(\mathbf{x},\mathbf{x})), \tag{454}$$

with $h\colon (0,\infty) \to \mathbb{R}$ defined by $h(a) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(az)^2]$.

In words, $\lambda_M(\mathbf{x})$ is the expected value of the output of the each neuron in the final hidden layer of the network under the prior at input $\mathbf{x}$ (after applying the activation) for a network of width M and intuitively $\lambda(\mathbf{x})$ is the limit of λ_M as $M \to \infty$ (this will be carefully proven in Proposition 45). $k^\ell(\cdot,\cdot)$ is the kernel function associated to the BNN with this activation under the prior as $M \to \infty$.

In order to construct the counterexample, we need three preliminary results. The first is the following calculation,

Proposition 44. Let P denote the prior for a network with L hidden layers and M neurons per hidden layer, i.e. $P = \mathcal{N}(\mathbf{0},\mathbf{I})$. Then, $\mathrm{KL}(Q^C, P) = \frac{C}{2}$.

Proof. By independence and the form of the variational posterior, we have,

$$\mathrm{KL}(Q^C,P) = \mathrm{KL}\Big(\mathcal{N}(\sqrt{C/M}\mathbf{1},\mathbf{I}_\mathrm{M}),\mathcal{N}(\mathbf{0},\mathbf{I}_\mathrm{M})\Big) = \frac{1}{2}\left\|\sqrt{C/M}\mathbf{1}\right\|_2^2 = \frac{C}{2M}\|\mathbf{1}\|_2^2 = \frac{C}{2}. \qquad \square$$

Proposition 45. Suppose ϕ satisfies conditions i. and ii. in Theorem 41. Then, for all $\mathbf{x} \in \mathbb{R}^{D_\mathrm{i}}$, $\lambda_M(\mathbf{x}) \to \lambda(\mathbf{x})$ and $1 + \mathbb{E}[\phi(\mathbf{z}_{L-1}^M(\mathbf{x})])^2] \to k(\mathbf{x},\mathbf{x})$.

Proposition 45 is essentially a corollary of results in Matthews et al. (2018).

Proof. Fix $\mathbf{x} \in \mathbb{R}^{D_\mathrm{i}}$. Let $z \sim \mathcal{N}(0, k^{L-1}(\mathbf{x},\mathbf{x}))$ and let $\xrightarrow{\mathrm{d}}$ denote convergence in distribution. By Matthews et al. (2018, Theorem 4) under P, and hence also under Q, $z_{L-1,m}^M(\mathbf{x}) \xrightarrow{\mathrm{d}} z$ for each m. Since ϕ is continuous, by the continuous mapping theorem (Billingsley, 2008, Theorem 25.7), $\phi(z_{L-1,m}^M(\mathbf{x})) \xrightarrow{\mathrm{d}} \phi(z)$. To strengthen convergence in distribution to convergence of the means, we note that $(\phi(z_{L-1,m}^M(\mathbf{x})))_{M=1}^\infty$ is uniformly integrable (Lemma 21 by Matthews et al., 2018) and apply Billingsley (2008, Theorem 25.12): $\lambda_M(\mathbf{x}) = \mathbb{E}[\phi(z_{L-1,m}^M(\mathbf{x}))] \to \mathbb{E}[\phi(z)] = \lambda(\mathbf{x})$. Noting that $k^L(\mathbf{x},\mathbf{x}) = 1 + \mathbb{E}[\phi(z)^2]$, the proof to show the second limit is exactly the same. $\square$

The final two propositions we need will show states that for activations satisfying conditions i-iv. in Theorem 41, $\lambda(x)$ takes at least two values:

Proposition 46. Suppose ϕ is continuous and $\phi(a)^2 + \phi(-a^2) \neq c$. Define $\kappa^\ell : \mathbb{R}^{D_\mathrm{i}} \to \mathbb{R}$ by $\kappa^\ell(\mathbf{x}) = k^\ell(\mathbf{x},\mathbf{x})$ with $k^\ell(\mathbf{x},\mathbf{x})$ defined by Equation (454). Then for all $\ell \in \mathbb{N} \cup \{0\}$, $\kappa^\ell(\mathbb{R}^{D_\mathrm{i}})$ contains an open interval.

Proposition 47. Suppose ϕ satisfies condition i-iii. of Theorem 41. Then for any open interval $I \subset (0,\infty)$ there exists an $a, a' \in I$ such that

$$\gamma(a) \neq \gamma(a'). \tag{455}$$

with $\gamma : (0,\infty) \to \mathbb{R}$ defined by $\gamma(a) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(az)]$. Further $\gamma(I)$ contains an open interval.

Remark 48. Note that taken together, these imply that if ϕ satisfies $i-iv.$ then the image of λ contains an open interval. This can be seen by applying Proposition 46 with $\ell = L-1$, to conclude the diagonal of k^{L-1} contains and open interval, then noting that $\lambda(\mathbf{x}) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(\sqrt{k^{L-1}(\mathbf{x},\mathbf{x})}z]$, so we may apply Proposition 47.

H.3 Construction

Construction of counterexample in Theorem 41. Suppose we have a data-set consisting of two points, $((\mathbf{x},y),(\mathbf{x}',y'))$, with $\mathbf{x},\mathbf{x}'$ such that $\lambda(\mathbf{x}) \neq \lambda(\mathbf{x}')$. The existence of such an $\mathbf{x},\mathbf{x}'$ is guaranteed by Propositions 46 and 47, see Remark 48.

Choose $y = \sqrt{C}\lambda(\mathbf{x}), y' = \sqrt{C}\lambda(\mathbf{x}')$ with $C \in (0, \infty)$ to be chosen later. We supress the dependce on C for convenience and write $Q := Q^C$. Let Q^* be any posterior with ELBO better than Q, then

$$\frac{1}{2\sigma^2}\left(\mathbb{E}_{Q^*}[(y - f_\theta(\mathbf{x}))^2] + \mathbb{E}_{Q^*}[(y' - f_\theta(\mathbf{x}'))^2]\right) \tag{456}$$

$$\leq \mathrm{KL}(Q, P) - \mathrm{KL}(Q^*, P) + \frac{1}{2\sigma^2}\left(\mathbb{E}_{Q}[(y - f_\theta(\mathbf{x}))^2] + \mathbb{E}_{Q}[(y' - f_\theta(\mathbf{x}'))^2]\right) \tag{457}$$

$$\leq \mathrm{KL}(Q, P) + \frac{1}{2\sigma^2}\left(\mathbb{E}_{Q}[(y - f_\theta(\mathbf{x}))^2] + \mathbb{E}_{Q}[(y' - f_\theta(\mathbf{x}'))^2]\right) \tag{458}$$

$$= \frac{C}{2} + \frac{1}{2\sigma^2}\left(\mathbb{E}_{Q}[(y - f_\theta(\mathbf{x}))^2] + \mathbb{E}_{Q}[(y' - f_\theta(\mathbf{x}'))^2]\right). \tag{459}$$

where the first inequality comes from rearranging both ELBOs, the second uses non-negativity of the KL divergence and makes use of Proposition 44.

Using convexity of the squared function to apply Jensen's inequality to the left hand side, and multiplying both sides by $2\sigma^2$ we have,

$$(y - \mathbb{E}_{Q^*}[f_\theta(\mathbf{x})])^2 + (y' - \mathbb{E}_{Q^*}[f_\theta(\mathbf{x}')])^2 \leq \mathbb{E}_{Q^*}[(y - f_\theta(\mathbf{x}))^2] + \mathbb{E}_{Q^*}[(y' - f_\theta(\mathbf{x}'))^2] \tag{460}$$

$$\leq \sigma^2 C + \mathbb{E}_{Q}[(y - f_\theta(\mathbf{x}))^2] + \mathbb{E}_{Q}[(y' - f_\theta(\mathbf{x}'))^2]. \tag{461}$$

Since $\lambda(\mathbf{x}) \neq \lambda(\mathbf{x}')$ there exists a $\beta > 0$ such that $|\lambda(\mathbf{x}) - \lambda(\mathbf{x}')| \geq \beta/\sqrt{2}$. Using our choice of y, y'

$$\mathbb{E}_{Q^*}[(y - f_\theta(\mathbf{x}))^2] = \mathbb{E}_{Q^*}[\sqrt{C}\lambda(\mathbf{x}) - f_\theta(\mathbf{x})]^2 \tag{462}$$

$$= (\sqrt{C}\lambda(\mathbf{x}) - \sqrt{C}\lambda_M(\mathbf{x}))^2 + \kappa_M(\mathbf{x}), \tag{463}$$

where κ_M is the variance function for the width M neural net prior. By Proposition 45, for M sufficiently large, we have

$$(\sqrt{C}\lambda(\mathbf{x}) - \sqrt{C}\lambda_M(\mathbf{x}))^2 + \kappa_M(\mathbf{x}) \leq \kappa(\mathbf{x}) + \beta^2/2, \tag{464}$$

where κ is variance function of the limiting NNGP kernel. The same argument can be applied to $\mathbf{x}'$. This gives us the upper bound, for M sufficiently large,

$$\mathbb{E}_{Q}[(y - f_\theta(\mathbf{x}))^2] \leq \kappa(\mathbf{x}) + \kappa(\mathbf{x}') + \beta^2. \tag{465}$$

Combining with our earlier equation, we have

$$(\sqrt{C}\lambda(\mathbf{x}) - \mathbb{E}_{Q^*}[f_\theta(\mathbf{x})])^2 + (\sqrt{C}\lambda(\mathbf{x}') - \mathbb{E}_{Q^*}[f_\theta(\mathbf{x}')])^2 \leq \sigma^2 C + \kappa(\mathbf{x}) + \kappa(\mathbf{x}') + \beta^2. \tag{466}$$

Therefore,

$$|\mathbb{E}_{Q^*}[]f_\theta(\mathbf{x})] - \mathbb{E}_{Q^*}[f_\theta(\mathbf{x}')]|$$

$$= |(\mathbb{E}_{Q^*}[f_\theta(\mathbf{x})] - \sqrt{C}\lambda(\mathbf{x})) - (\mathbb{E}_{Q^*}[f_\theta(\mathbf{x}')] - \sqrt{C}\lambda(\mathbf{x}')) + (\sqrt{C}\lambda(\mathbf{x}) - \sqrt{C}\lambda(\mathbf{x}')| \tag{467}$$

$$\geq |\sqrt{C}\lambda(\mathbf{x}) - \sqrt{C}\lambda(\mathbf{x}')| - |(\mathbb{E}_{Q^*}[f_\theta(\mathbf{x})] - \sqrt{C}\lambda(\mathbf{x})) - (\mathbb{E}_{Q^*}[f_\theta(\mathbf{x}')] - \sqrt{C}\lambda(\mathbf{x}'))| \tag{468}$$

$$\overset{\text{(i)}}{\geq} |\sqrt{C}\lambda(\mathbf{x}) - \sqrt{C}\lambda(\mathbf{x}')| - \sqrt{2}\sqrt{(\mathbb{E}_{Q^*}[f_\theta(\mathbf{x})] - \sqrt{C}\lambda(\mathbf{x}))^2 + (\mathbb{E}_{Q^*}[f_\theta(\mathbf{x}')] - \sqrt{C}\lambda(\mathbf{x}'))^2} \tag{469}$$

$$\geq \sqrt{C}\beta/\sqrt{2} - \sqrt{2}\sqrt{\sigma^2 C + \kappa(\mathbf{x}) + \kappa(\mathbf{x}') + \beta^2} \tag{470}$$

using in (i) $|(x-a)-(y-b)| \leq |x-a| + |y-b| \leq \sqrt{2}\sqrt{(x-a)^2 + (y-b)^2}$. To finish the proof, choose $\sigma^2 = 1/C$ and take C large enough that the first term in Equation (470) is larger than the second. □

H.4 Proof of Propositions 46 and 47

Having completed the construction of the counterexample, it remains to prove Propositions 46 and 47 in order to verify that we can indeed select two points $\mathbf{x}, \mathbf{x}'$ such that $\lambda(\mathbf{x}) \neq \lambda(\mathbf{x}')$. Note that for any typical activation, this could simply be verified numerically by working through the recursion for kernel functions, so the main purpose of the following proofs is generality.

Proposition 49. Define $\alpha\colon (0,\infty) \to \mathbb{R}$ by $\alpha(a) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(az)]$. Suppose ϕ satisfies conditions i, ii of Theorem 41. Then α is continuous.

Proof. Let $(a_i)_{i\geq 1} \subseteq \mathbb{R}$ be convergent to some $a \in \mathbb{R}$. We show that $\mathbb{E}[\phi(a_i z)] \to \mathbb{E}[\phi(az)]$. If we can interchange limit and expectation, then the result follows from continuity of ϕ. To show that we can interchange limit and expectation, we demonstrate an integrable dominating function. Since $(a_i)_{i\geq 1}$ is convergent, it is bounded, hence contained in some interval $[-K, K] \subseteq \mathbb{R}$. Using that $\phi(a) = O(|a|+1)$, let $|\phi(a)| \leq C|a|$ for all $|a| \geq R$ for some $C > 0$. Let M be the maximum of $|\phi|$ on $[-R, R]$, which is finite because $|\phi|$ is continuous. Then estimate

$$|\phi(a_i z)| \leq \sup_{a\in[-K,K]} |\phi(az)| \leq M + C \sup_{a\in[-K,K]} |az| \leq M + CK|z|, \tag{471}$$

which is integrable because $z \sim \mathcal{N}(0,1)$. □

Proof of Proposition 47. Our proof follows themaker (https://math.stackexchange.com/users/ 114509/themaker). Without loss of generality, we may assume $I = (s,t)$ is bounded with $s > 0$.

Towards contradiction, suppose there exists a $c \in \mathbb{R}$ such that $\alpha(a) = c$ for all $a \in I$. This supposition can be written as,

$$\int \phi(az) e^{-z^2/2} dz = c' \quad \forall a \in I, \tag{472}$$

where $c' = c\sqrt{2\pi}$. Define $u = az$, $b = 1/(2a^2)$ and $I' = \left(\frac{1}{2t^2}, \frac{1}{2s^2}\right)$. Then we can rewrite Equation (472) as

$$\int \phi(u) e^{-bu^2} du = c' \quad \forall b \in I'. \tag{473}$$

We have $|u^{2n}\phi(u)e^{-bu^2}| \leq u^{2n}|\phi(u)|e^{-bu^2} \leq u^{2n}|\phi(u)|e^{-\frac{u^2}{2t^2}}$, which is integrable since $\phi(u) = O(|x|+1)$. Hence we may take n derivatives and apply Leibniz's rule,

$$\frac{\partial^n}{\partial^n b} \int \phi(u) e^{-bu^2} du = \int u^{2n} \phi(u) e^{-bu^2} du = 0 \quad \forall b \in I', \forall n \in \mathbb{N}. \tag{474}$$

For a (arbitrary) $b \in I'$, define $w(u) = e^{-bu^2}$ and $L^2(\mathbb{R}, w)$ to be the Hilbert space with inner product $\langle f, g\rangle = \int f(u)g(u)w(u)du$. Following Mykie (https://math.stackexchange.com/users/832/ mykie) (2012), we note that the set of compactly supported functions is dense in $L^2(\mathbb{R}, w)$, and by the Weirstrauss theorem the set of polynomial is dense in the set of compact functions. As a dense subset of a dense set is again dense, we conclude the set of polynomial is dense in $L^2(\mathbb{R}, w)$.

Writing $\phi(u) = \phi_e(u) + \phi_o(u)$, we have,

$$\int_{-\infty}^{\infty} u^{2i-1} \phi_e(u) w(u) du = 0 \quad \forall i \in \mathbb{N} \tag{475}$$

as the integrand is odd and we integrate over a symmetric domain. Also,

$$\int u^{2n} \frac{\phi(u) + \phi(-u)}{2} w(u) du = \frac{1}{2}\int u^{2i}\phi(u)w(u)du + \frac{1}{2}\int u^{2n}\phi(-u)w(u)du \tag{476}$$

$$= \frac{1}{2}\int u^{2i}\phi(u)w(u)du + \frac{1}{2}\int u^{2n}\phi(u)w(u)du \tag{477}$$

$$= 0 + 0, \quad \forall i \in \mathbb{N}, \tag{478}$$

by Equation (474). Since the polynomial are a basis, we have $\phi_e(u) = \sum_{i=0}^{\infty} \alpha_i u^i$ for some $(\alpha_i)_{i=1}^{\infty}$. Taking the inner product of both sides with respect to u^i, shows that $\alpha_i = 0$ for all $i > 0$, hence $\phi_e(u) = c''$ for some $c'' \in \mathbb{R}$, with equality in $L^2(\mathbb{R}, w)$. But as ϕ_e is continuous, this implies it is equal to a c'' everywhere.

As I contains an open set, it contains a ball. As α is continuous (Proposition 49) the image of this ball under α is connected i.e. an interval. As the interval contains at least two points, it contains an open interval. □

Proof of Proposition 46. The proof proceeds by induction.

Base case: We have $\kappa^0(\mathbf{x}) = 1 + \|\mathbf{x}\|_2^2$. As $\mathbb{R}^{D_\text{i}}$ contains an open set, it contains an open ball. The image of this open ball under the map κ, which is continuous, must be connected, hence an interval. Therefore, if it contains at least two points, it contains an open interval. But in any ball, there are two points with different norms, so this must be the case.

Inductive step: Defining $h(a) = \mathbb{E}[\phi(az)]$ and following the recursion for kernels for deep networks Matthews et al. (2018, Lemma 2), we have $\kappa^\ell(\mathbb{R}^{D_\text{i}}) = 1 + h(\kappa^{\ell-1}(\mathbb{R}^{D_\text{i}}))$. By the inductive hypothesis, $\kappa^{\ell-1}(\mathbb{R}^{D_\text{i}})$ contains an open interval, I, the image of which must be connected as h is continuous (ϕ^2 is continuous and is $O(|a|^2+1)$ so we may apply Proposition 49). Hence, it suffices to show that ϕ^2 satisfies the conditions of Proposition 47. This follows from the assumption that $\phi(a)^2 + \phi(-a^2) \neq c$. □

I Proof of Constants for Mean Result in Single Hidden-Layer Network

We have,

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})]\|_2 = \tfrac{1}{\sqrt{M}}\|\mathbb{E}[\mathbf{W}_2]\mathbb{E}[\phi(\tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)]\| \leq \|\mathbb{E}[\mathbf{W}_2]\|_\text{F}\|\mathbb{E}[\phi(\tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b})]\|_2 \tag{479}$$

Defining $\mathbf{W}' = \mathbf{W}_1 - \mathbb{E}[\mathbf{W}_1]$ and $\mathbf{b}' = \mathbf{b}_1 - \mathbb{E}[\mathbf{b}_1]$, we have

$$\begin{aligned}
\|\mathbb{E}[\phi(\tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b})]\|_2 &= \|\mathbb{E}[\phi(\tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)] - \mathbb{E}[\phi(\mathbf{W}'\mathbf{x} + \mathbf{b}')]\|_2 && (480)\\
&\leq \mathbb{E}\|\phi(\tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)] - \phi(\tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}'\mathbf{x} + \mathbf{b}')\| && (481)\\
&\leq \mathbb{E}\|(\tfrac{1}{\sqrt{D_\text{i}}}\mathbf{W}_1 - \mathbf{W})\mathbf{x} + (\mathbf{b} - \mathbf{b}')\| && (482)\\
&\leq \|\mathbb{E}[\mathbf{W}_1]\|_\text{F}\tfrac{1}{\sqrt{D_\text{i}}}\|\mathbf{x}\|_2 + \|\mathbb{E}[\mathbf{b}_1]\|_2 && (483)
\end{aligned}$$

Combining Equation (479) and Equation (483),

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})]\|_2 \leq \|\mathbb{E}[\mathbf{W}_2]\|_\text{F}\|\mathbb{E}[\mathbf{W}_1]\|_\text{F}\tfrac{1}{\sqrt{D_\text{i}}}\|\mathbf{x}\|_2 + \|\mathbb{E}[\mathbf{W}_2]\|_\text{F}\|\mathbb{E}[\mathbf{b}_1]\|_2. \tag{484}$$

We also know from Lemma 9 that

$$\|\mathbb{E}[\mathbf{W}_1]\|_\text{F}^2 + \|\mathbb{E}[\mathbf{W}_2]\|_\text{F}^2 + \|\mathbb{E}[\mathbf{b}_1]\|_2 \leq \text{KL}(Q, P). \tag{485}$$

Combining Equation (484) and Equation (485) we obtain the following upper bound phrased as an optimization problem which is convex,

$$\begin{aligned}
\|\mathbb{E}[\mathbf{f}(\mathbf{x})]\|_2 \leq \max_\alpha \quad & \alpha_1\alpha_2\tfrac{1}{\sqrt{D_\text{i}}}\|\mathbf{x}\|_2 + \alpha_1\alpha_3 && (486)\\
\text{s.t. } & \frac{1}{2}\sum \alpha_i^2 = \text{KL}(Q, P), \quad \alpha_i \geq 0. && (487)
\end{aligned}$$

We can solve this optimization via Lagrange multipliers. We form the Lagrangian,

$$\alpha_1\alpha_2\tfrac{1}{\sqrt{D_\text{i}}}\|\mathbf{x}\|_2 + \alpha_1\alpha_3 - \lambda\left(\frac{1}{2}\sum \alpha_i^2 - \text{KL}(Q, P)\right). \tag{488}$$

For convenience, name $c = \frac{1}{\sqrt{D_\text{i}}}\|\mathbf{x}\|_2$. Differentiating with respect to each variable and setting to 0 gives,

$$\begin{aligned}
c\alpha_2 + \alpha_3 - \lambda\alpha_1 &= 0 && (489)\\
\alpha_2 &= \frac{c}{\lambda}\alpha_1 && (490)\\
\alpha_3 &= \frac{1}{\lambda}\alpha_1. && (491)
\end{aligned}$$

Plugging these back in to the constraint,

$$\frac{1}{2}\alpha_1^2\left(1 + \frac{1}{\lambda^2} + \frac{c^2}{\lambda^2}\right) = \text{KL}(Q, P) \tag{492}$$

Solving for λ yields,

$$\lambda = \alpha_1 \sqrt{\frac{1+c^2}{\mathrm{KL}(Q,P) - \frac{1}{2}\alpha_1^2}} \tag{493}$$

Hence

$$\alpha_2 = c\sqrt{\frac{\mathrm{KL}(Q,P) - \frac{1}{2}\alpha_1^2}{1+c^2}}, \quad \alpha_3 = \sqrt{\frac{\mathrm{KL}(Q,P) - \frac{1}{2}\alpha_1^2}{1+c^2}}. \tag{494}$$

We can now plug these back into the remaining constraint to give,

$$\sqrt{1+c^2}\sqrt{\mathrm{KL}(Q,P) - \frac{1}{2}\alpha_1^2} - \alpha_1^2 \sqrt{\frac{1+c^2}{\mathrm{KL}(Q,P) - \frac{1}{2}\alpha_1^2}} = 0 \tag{495}$$

Simplifying slightly,

$$\left(\mathrm{KL}(Q,P) - \frac{1}{2}\alpha_1^2\right)^2 = \alpha_1^4 \tag{496}$$

We recognize this as a quadratic form in α_1^2, which can be solved yielding $\alpha_1 = \sqrt{\frac{2}{3}\,\mathrm{KL}(Q,P)}$. So $\alpha_2 = c\sqrt{\frac{\frac{2}{3}\,\mathrm{KL}(Q,P)}{1+c^2}}$ and $\alpha_3 = \sqrt{\frac{\frac{2}{3}\,\mathrm{KL}(Q,P)}{1+c^2}}$. The result then follows from a short calculation.

J Convergence of Mean of Linear Networks

We consider the case of networks with only affine layers. In particular, we suppose $\phi(a) = a$ for all $a \in \mathbb{R}$. In this case, we can prove upper and lower bounds on the discrepancy between the mean function at two points in the input space.

J.1 Upper Bounds

Then for two points $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{D_\mathrm{i}}$,

$$\mathbb{E}[\mathbf{f}(\mathbf{x})] - \mathbb{E}[\mathbf{f}(\mathbf{x}')] = \mathbb{E}[\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}')] \tag{497}$$

$$= D_\mathrm{i}^{-1/2} M^{-L/2} \mathbb{E}\Big[\prod_{\ell=1}^{L+1} \mathbf{W}_\ell\Big](\mathbf{x} - \mathbf{x}') \tag{498}$$

$$\leq D_\mathrm{i}^{-1/2} M^{-L/2} \left\| \prod_{\ell=1}^{L+1} \mathbb{E}[\mathbf{W}_\ell] \right\|_2 \|\mathbf{x} - \mathbf{x}'\|. \tag{499}$$

Using sub-multiplicativity of operator norm and that $\|\cdot\|_2 \leq \|\cdot\|_F$,

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})] - \mathbb{E}[\mathbf{f}(\mathbf{x}')]\|_2 \leq D_\mathrm{i}^{-1/2} M^{-L/2} \prod_{\ell=1}^{L+1} \|\mathbb{E}[\mathbf{W}_\ell]\|_\mathrm{F} \|\mathbf{x} - \mathbf{x}'\|_2. \tag{500}$$

We can now apply the arithmetic-geometric mean inequality, to conclude

$$\prod_{\ell=1}^{L+1} \|\mathbb{E}[\mathbf{W}_\ell]\|_F \leq \left(\frac{\sum_{\ell=1}^{L+1} \|\mathbb{E}[\mathbf{W}_\ell]\|_F}{L+1}\right)^{L+1} \tag{501}$$

Using the ℓ^2-ℓ^1-inequality, we have, and Lemma 9,

$$\sum_{\ell=1}^{L+1} \|\mathbb{E}[\mathbf{W}_\ell]\|_F \leq \sqrt{L+1} \sqrt{\sum_{\ell=1}^{L+1} \|\mathbb{E}[\mathbf{W}_\ell]\|_F^2} \leq \sqrt{L+1}\sqrt{2\,\mathrm{KL}(Q,P)}. \tag{502}$$

Combining Equation (502), Equation (501) and Equation (500) we obtain,

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})] - \mathbb{E}[\mathbf{f}(\mathbf{x}')]\|_2 \leq D_{\mathrm{i}}^{-1/2} M^{-L/2} \left(\frac{2\,\mathrm{KL}(Q,P)}{L+1}\right)^{\frac{L+1}{2}} \|\mathbf{x} - \mathbf{x}'\|_2 \tag{503}$$

J.2 Lower Bounds

We consider the case when $\mathbf{x}' = \mathbf{0}$ and $\mathbf{x} = \mathbf{e}_1$. We consider the Q with variance of each parameter equal to 1 and mean of all bias parameters equal to 0. We select $\mathbb{E}[W_\ell]$ to be the matrix with entry $1, 1$, c and entries 0 elsewhere. Then, $\frac{1}{2}(L+1)c^2 = \mathrm{KL}(Q,P)$, so $c = \frac{\sqrt{2\,\mathrm{KL}(Q,P)}}{L+1}$. Also,

$$\|\mathbb{E}[\mathbf{f}(\mathbf{x})] - \mathbb{E}[\mathbf{f}(\mathbf{0})]\|_2 = D_{\mathrm{i}}^{-1/2} M^{-L/2} c^{L+1} = D_{\mathrm{i}}^{-1/2} M^{-L/2} (2\,\mathrm{KL}(Q,P))^{\frac{L+1}{2}} (L+1)^{-(L+1)}. \tag{504}$$

This bound differs from the upper bound by a factor of $(L+1)^{-\frac{L+1}{2}}$.

K Lower Bound on Convergence for Non-Linear Networks

Theorem 50. Assume the following:

(i) $D_{\mathrm{o}} = 1$.

(ii) ϕ is a sum of an odd function and a constant: ϕ_{e} is constant.

(iii) ϕ is twice continuously differentiable with $\|\phi'\|_\infty \leq 1$ and $\|\phi''\|_\infty < \infty$.

(iv) ϕ^2 is not a sum of an odd function and a constant: $\phi^2(a) + \phi^2(-a) \neq 2\phi^2(0)$ for some $a \in \mathbb{R}$.

Then, if ϕ_{o} is non-linear, there exist two inputs $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{D_{\mathrm{i}}}$ and a constant $c > 0$ such that, for every $K > 0$, there exists a sequence of mean-field distributions $(Q_M)_{M \geq 1}$ with $\mathrm{KL}(Q_M, P) = K$, one for every network width $M \geq 1$, that achieves

$$\lim_{M \to \infty} \sqrt{M} |\mathbb{E}_{Q_M}[f(\mathbf{x})] - \mathbb{E}_{Q_M}[f(\mathbf{x}')]| = cK. \tag{505}$$

Proof. Consider the distribution of $\mathbf{z}_{L-1}$ under the prior. Let k be the covariance function of the NNGP associated to $\frac{1}{\sqrt{M}}\langle \boldsymbol{\varepsilon}, \phi(\mathbf{z}_{L-1})\rangle + Z$ as $M \to \infty$, where $\boldsymbol{\varepsilon}$ is a vector with i.i.d. $\mathcal{N}(0,1)$ entries and $Z \sim \mathcal{N}(0,1)$. Henceforth, denote $k(\mathbf{x}) = k(\mathbf{x}, \mathbf{x})$.

Note that $\phi' = \phi'_{\mathrm{e}} + \phi'_{\mathrm{o}} = \phi'_{\mathrm{o}}$ is an even function. Suppose that $\mathbf{x} \mapsto \mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)]$ is a constant function. By Propositions 45 and 46 in combination with the assumed conditions on ϕ, it follows that $k(\mathbb{R}^{D_{\mathrm{i}}})$ contains an open interval. Therefore, since ϕ' is a continuous even function and $\mathbf{x} \mapsto \mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)]$ is a constant function, by an argument similar to the proof of Proposition 47, ϕ' must be equal to a constant function. However, since ϕ_{o} is non-linear, $\phi' = \phi'_{\mathrm{o}}$ cannot be equal to a constant function. We conclude that $\mathbf{x} \mapsto \mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)]$ cannot be equal to a constant function: there exist two inputs $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{D_{\mathrm{i}}}$ such that $\mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)] \neq \mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x}')}Z)]$. Let $c > 0$ be the constant $c = |\mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)] - \mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x}')}Z)]|$.

Let $K > 0$. Consider the sequence of mean-field distributions $(Q_M)_{M \geq 1}$ constructed by setting everything equal to the prior except for $\mathbb{E}_{Q_M}[\mathbf{b}_L] = \mu\mathbf{1}$ and $\mathbb{E}_{Q_M}[\mathbf{W}_{L+1}] = \mu\mathbf{1}^{\mathsf{T}}$ with $\mu = \sqrt{K/M}$. Then indeed $\mathrm{KL}(Q_M, P) = K$. Moreover,

$$\tfrac{1}{\sqrt{M}}\mathbf{W}_{L+1}\phi(\tfrac{1}{\sqrt{M}}\mathbf{W}_L\phi(\mathbf{z}_{L-1}) + \mathbf{b}_L) + \mathbf{b}_{L+1} \overset{\mathrm{d}}{=} \tfrac{1}{\sqrt{M}}\langle \mu\mathbf{1} + \boldsymbol{\varepsilon}, \phi(\tfrac{1}{\sqrt{M}}\boldsymbol{\mathcal{E}}\phi(\mathbf{z}_{L-1}) + \mu\mathbf{1} + \boldsymbol{\varepsilon}')\rangle + \boldsymbol{\varepsilon}'' \tag{506}$$

where $\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}'$, $\boldsymbol{\varepsilon}''$, and $\boldsymbol{\mathcal{E}}$ are respectively three vectors and a matrix with i.i.d. $\mathcal{N}(0,1)$ entries and where $\mathbf{z}_L$ is distributed under the prior. Using the observation that the elements of $\mathbf{z}_L$ are identically distributed, compute

$$\mathbb{E}_{Q_M}[f(\mathbf{x})] = \tfrac{\mu}{\sqrt{M}}\mathbb{E}_P[\langle \mathbf{1}, \phi(\tfrac{1}{\sqrt{M}}\boldsymbol{\mathcal{E}}\phi(\mathbf{z}_{L-1}) + \mu\mathbf{1} + \boldsymbol{\varepsilon}')\rangle] = \sqrt{M}\mu\,\mathbb{E}_P[\phi(\tfrac{1}{\sqrt{M}}\langle \boldsymbol{\varepsilon}''', \phi(\mathbf{z}_{L-1})\rangle + Z + \mu)] \tag{507}$$

where $\boldsymbol{\varepsilon}'''$ is a vector with i.i.d. $\mathcal{N}(0,1)$ entries and $Z \sim \mathcal{N}(0,1)$. Note that $\sqrt{M}\mu = \sqrt{K}$ and call $Y_M = \frac{1}{\sqrt{M}}\langle \boldsymbol{\varepsilon}''', \phi(\mathbf{z}_{L-1})\rangle + Z$. From Theorem 4 by Matthews et al. (2018) in combination with the assumed conditions on ϕ, it follows that $Y_M \overset{\mathrm{d}}{\to} \sqrt{k(\mathbf{x})}Z$ where k is the earlier defined covariance function of the corresponding NNGP.

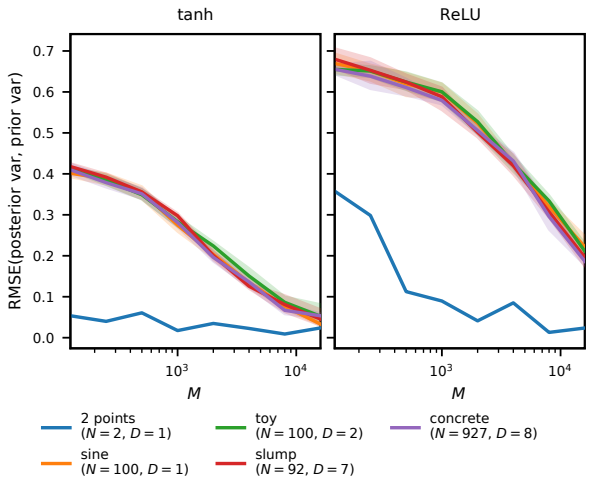

Figure 8: Root mean squared error (RMSE) of the posterior variance to the prior variance.

We finally establish the asymptotic behaviour of $\mathbb{E}_{Q_M}[f(\mathbf{x})] - \mathbb{E}_{Q_M}[f(\mathbf{x}')]$ in two steps. First, by a second-order Taylor expansion of $\mu \mapsto \mathbb{E}_P[\phi(Y_M + \mu)]$ around $\mu = 0$, using (a) $\|\phi'\| < \infty$ and $\|\phi''\|_\infty < \infty$ to interchange derivative and expectation and (b) $\|\phi''\|_\infty < \infty$ and $\mu^2 = o(\frac{1}{\sqrt{M}})$ to determine the order of the error term,

$$\mathbb{E}_{Q_M}[f(\mathbf{x})] = \sqrt{K}\Big[\mathbb{E}_P[\phi(Y_M)] + \mu\,\mathbb{E}_P[\phi'(Y_M)] + o(\tfrac{1}{\sqrt{M}})\Big]. \tag{508}$$

Note that $Y_M \overset{\mathrm{d}}{=} -Y_M$, so $\mathbb{E}_P[\phi(Y_M)] = \phi_\mathrm{e} + \mathbb{E}_P[\phi_\mathrm{o}(Y_M)] = \phi_\mathrm{e}$. Second, since $Y_M \overset{\mathrm{d}}{\to} \sqrt{k(\mathbf{x})}Z$ and ϕ' is continuous and bounded, $\mathbb{E}_P[\phi'(Y_M)] \to \mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)]$. Therefore,

$$\mathbb{E}_{Q_M}[f(\mathbf{x})] = \sqrt{K}\Big[\phi_\mathrm{e} + \mu\Big(\mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)] + o(1)\Big) + o(\tfrac{1}{\sqrt{M}})\Big], \tag{509}$$

so, again using that $\sqrt{M}\mu = \sqrt{K}$,

$$\sqrt{M}(\mathbb{E}_{Q_M}[f(\mathbf{x})] - \mathbb{E}_{Q_M}[f(\mathbf{x}')]) = K\Big(\mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x})}Z)] - \mathbb{E}_P[\phi'(\sqrt{k(\mathbf{x}')}Z)] + o(1)\Big) + o(1) = cK + o(1), \tag{510}$$

which concludes the proof. □

L Experimental Setup

In this section we describe the details of our experiments.

Additional experiments Figure Figure 8 shows the RMSE between the MFVI posterior posterior predictive variance and the prior predictive variance.

Architecture. We only train single-layer ($L = 1$), feed-forward networks of varying width M, as defined in Equations 1-3. Notice we use the NTK parameterization, which scales the post-activations by the network width (Jacot et al., 2018). Depending on the experiment, we use ReLU, tanh, or erf activation functions, where $\operatorname{erf} z = 2/\sqrt{\pi}\int_0^z e^{-t^2}\,dt$ is the error function.

Prior and Likelihood. We use a $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ prior for the neural network weights and a heteroscedastic Gaussian likelihood with a known variance. Except for the counterexample dataset, we set this variance to 0.025.

Variational Family. We use a mean-field Gaussian family for variational inference. In other words, we model each weight and bias parameter of the neural network by an independent Gaussian distribution.

Initialization of Variational Parameters. We initialize the variational mean and variance parameters from a normal-inverse-gamma family. Specifically, for any weight (or bias) of the neural network θ, let $\mathcal{N}(\mu_Q, \sigma_Q^2)$ denote its variational distribution. We randomly initialize $\mu_Q \sim \mathcal{N}(0, 1)$ and $\sigma_Q^2 \sim \mathcal{IG}(\nu + 1, \nu)$. It follows from the laws of total expectation and variance that $\mathbb{E}[\theta] = 0$ and $\mathbb{V}[\theta] = 2$. This allows the for a width-independent initialization of the weights, as is standard for the NTK parameterization, while allowing the hyperparameter ν to control the concentration of σ_Q^2 around its initial mean of one (i.e., $\mathbb{E}[\sigma_Q^2] = 1$ and $\mathbb{V}[\sigma_Q] = 1/(\nu - 1)$). We set $\nu = 100$ in our experiments.

Datasets.

- *2 points* ($N = 2$, $D_{\text{i}} = 1$): This dataset consists of two points: $(-1, -1)$ and $(1, 1)$. This dataset is used in all figures in this paper and is shown in Figure 1.

- *sine* ($N = 100$, $D_{\text{i}} = 1$): This is a synthetic dataset generated by $y = \sin(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, .025)$ and $x \sim \text{Unif}(-5, 5)$.

- *toy* ($N = 100$, $D_{\text{i}} = 2$): This is a synthetic dataset generated by $y = x_0 \sin(x_1) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, .025)$, $x_0 \sim \text{Unif}(-5, 5)$, and $x_1 \sim \text{Unif}(-5, 5)$.

- *counterexample* ($N = 2$, $D_{\text{i}} = 1$): This synthetic dataset consists of two observations, $(0, 8.24)$ and $(1, 11.66)$. It is constructed to meet the conditions of the example discussed in Appendix H. The mean-field posterior predictive of a network with ReLU activation need not converge the prior when trained on this dataset. As part of the construction, we set the observational noise variance of the likelihood to 2.34×10^{-3}. This dataset is shown in Figure 4.

- *slump* ($N = 103$, $D_{\text{i}} = 7$): This is the Concrete Slump Test Data Set, available in the UCI Machine Learning Repository (Yeh, 2007).

- *concrete* ($N = 1030$, $D_{\text{i}} = 8$): This is the Concrete Compressive Strength Data Set, available in the UCI Machine Learning Repository (Yeh, 1998).

All variables (inputs x and observations y) are z-scored standardized (i.e., by subtracting their mean and dividing by their standard deviation). For the synthetic datasets of only two training observations we do not construct test observations. For the larger synthetic datasets, we sample 100 test observations. For the real datasets, we use 10% of the observations as test observations.

Training Procedure. We use 20,000 steps of stochastic gradient descent with a batch size of 100, a learning rate of 0.001, and a momentum of 0.9 for optimization. Note that since the post-activations are already scaled by $1/\sqrt{M}$ in the network definition, we do not scale the learning rate with the network width (see, e.g., Appendix F of (Lee et al., 2019) for a discussion of the learning rates under the NTK parameterization). We use gradient clipping and cosine annealing of the learning rate, with warm restarts every 500 steps (Loshchilov and Hutter, 2017). To evaluate the ELBO, we use the analytical form of the KL divergence and the reparameterization trick (Kingma and Welling, 2014) with 16 samples to approximate the expected log likelihood term.

Optimal bias. In the case of a Gaussian likelihood, we can solve for the optimal variational distribution over bias when all variational parameters are set to the prior. This enables a smaller bound on $KL(Q^*, P)$ in practice. Let $\widetilde{P}$ be the standard prior distribution except with the distribution over the output bias replaced by a normal

distribution $\mathcal{N}(\mu_b, \sigma_b^2)$. We will choose μ_b and σ_b^2 to maximize the ELBO.

$$\text{ELBO}(\widetilde{P}) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\mathbb{E}\left[(y_n - f(x_n))^2\right] - \frac{1}{2}\left(\mu_b^2 + \sigma_b^2 - 1 - \log(\sigma_b^2)\right) \tag{511}$$

$$= C - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\mathbb{E}\left[f(x_n)^2\right] - 2y_n\mathbb{E}\left[f(x_n)\right]\right) - \frac{1}{2}\left(\mu_b^2 + \sigma_b^2 - \log(\sigma_b^2)\right) \tag{512}$$

$$= C' - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(\mathbb{V}[f(x_n)] + \mathbb{E}\left[f(x_n)\right]^2 - 2y_n\mathbb{E}\left[f(x_n)\right]\Big) - \frac{1}{2}\left(\mu_b^2 + \sigma_b^2 - \log(\sigma_b^2)\right) \tag{513}$$

$$= C' - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\sigma_b^2 + \mu_b^2 - 2y_n\mu_b\right) - \frac{1}{2}\left(\mu_b^2 + \sigma_b^2 - \log(\sigma_b^2)\right) \tag{514}$$

$$= C' - \frac{N}{2\sigma^2}\sigma_b^2 - \frac{N}{2\sigma^2}\mu_b^2 - \frac{\mu_b}{\sigma^2}\sum_{n=1}^{N}y_n - \frac{1}{2}\left(\mu_b^2 + \sigma_b^2 - \log(\sigma_b^2)\right), \tag{515}$$

where C and C' are constants. Differentiating with respect to μ_b and setting to 0 we have,

$$\mu_b = \frac{\sum_{n=1}^{N} y_n}{N + \sigma^2} \tag{516}$$

and Similarly, we can differentiate with respect to σ_b^2 and set to 0, to obtain,

$$\sigma_b^2 = \frac{\sigma^2}{N + \sigma^2}. \tag{517}$$

Computing a bound on $KL(Q^*, P)$. By the optimality of Q^* we have $\text{ELBO}(Q^*) \geq \text{ELBO}(\widetilde{P})$. As in step 3 of Section 4.4, it follows that

$$\text{KL}(Q^*, P) \leq \mathbb{E}_{Q^*}[\mathcal{L}(\boldsymbol{\theta})] - \mathbb{E}_{\widetilde{P}}[\mathcal{L}(\boldsymbol{\theta})] + \text{KL}(\widetilde{P}, P) \tag{518}$$

$$\leq -\mathbb{E}_{\widetilde{P}}[\mathcal{L}(\boldsymbol{\theta})] + \text{KL}(\mathcal{N}(\mu_b, \sigma_b^2), \mathcal{N}(0, 1)) \tag{519}$$

$$\leq \frac{1}{2\sigma^2}\sum_{n=1}^{N}\mathbb{E}_{\widetilde{P}}[(y_n - f(\mathbf{x}_n))^2] + \frac{1}{2}\left(\mu_b^2 + \sigma_b^2 - \log(\sigma_b^2)\right). \tag{520}$$

For our experiments we compute the expectation by Monte Carlo sampling. Using $\widetilde{P}$ instead of P lowers the upper bound on $\text{KL}(Q^*, P)$ for any dataset for which the increase in the log likelihood from using the optimal bias more than offsets the increase in the KL divergence to the prior (e.g., datasets that are shifted by a constant from the prior mean of zero, as in the counterexample dataset). In the case of a one-hidden layer network, we can evaluate the expectation in Equation (520) either using properties of the activation in closed form or up to special function, or via one-dimensional Gaussian quadrature more generally. Additionally, the expectation is independent of M.

Figures. Here we explain a few details specific to each figure

- *Figure 1*: The shaded region represent ± 1 standard deviation.
- *Figure 2*: We train on the "2 points" dataset. We use 1,000 samples to estimate the posterior predictive mean on a grid of 25 inputs spaced uniformly over $[-1, 1]$. To reduce Monte-Carlo error, we also estimate the predictive mean under the prior with the same random seed. For each M, we use the same random seed, so the shaded regions reflect the randomness in the variational parameter initialization only (we use 10 random initializations). We then plot the largest absolute difference from the prior mean of zero. To compute the theoretical bound we use Equation (6), with the KL divergence estimated as in Equation (520) and $\|\mathbf{x}\|_2^2 \leq 1$.

- *Figures 3 and 5* : We use 5 different train/test splits (or 5 different random datasets in the case of the synthetic datasets). For each dataset and each M, we select the the model with the highest ELBO among two random restarts of the variational parameters. The shaded regions represent 95% confidence intervals estimated by boostrapping. To compute the RMSE to the prior, we use 1,000 samples of the posterior predictive evaluated at 100 input points drawn randomly from a uniform distribution over $[-1, 1]$ in each input dimension.
- *Figure 4*: We train single-layer networks of width 4,096,000.