Gaussian Process Behaviour in Wide Deep Neural Networks

Alexander G. de G. Matthews AM554@CAM.AC.UK
Department of Engineering
Trumpington Street
University of Cambridge, UK

Mark Rowland MR504@CAM.AC.UK
Department of Pure Mathematics and Mathematical Statistics
Wilberforce Road
University of Cambridge, UK

Jiri Hron JH2084@CAM.AC.UK
Department of Engineering
Trumpington Street
University of Cambridge, UK

Richard E. Turner RET26@CAM.AC.UK
Department of Engineering
Trumpington Street
University of Cambridge, UK

Zoubin Ghahramani ZOUBIN@ENG.CAM.AC.UK
Department of Engineering
Trumpington Street
University of Cambridge, UK
Uber AI Labs

Abstract

Whilst deep neural networks have shown great empirical success, there is still much work to be done to understand their theoretical properties. In this paper, we study the relationship between random, wide, fully connected, feedforward networks with more than one hidden layer and Gaussian processes with a recursive kernel definition. We show that, under broad conditions, as we make the architecture increasingly wide, the implied random function converges in distribution to a Gaussian process, formalising and extending existing results by Neal (1996) to deep networks. To evaluate convergence rates empirically, we use maximum mean discrepancy. We then compare finite Bayesian deep networks from the literature to Gaussian processes in terms of the key predictive quantities of interest, finding that in some cases the agreement can be very close. We discuss the desirability of Gaussian process behaviour and review non-Gaussian alternative models from the literature.[1]

1. Introduction

This work substantially extends the work of Matthews et al. (2018) published at ICLR 2018. Deep feedforward neural networks have emerged as an essential component of modern machine learning. As such there has been significant research effort in trying to understand the theoretical properties of such models. One important branch of this research is the study of random networks. By assuming a probability distribution on the network param-

[1] Code for the experiments in the paper can be found at `https://github.com/widedeepnetworks/widedeepnetworks`

eters, a distribution is induced on the input to output function that the networks encode. This has proved important in the study of initialisation and learning dynamics (Schoenholz et al., 2017) and expressivity (Poole et al., 2016). It is, of course, essential in the study of Bayesian priors on networks (Neal, 1996). The Bayesian approach makes little sense if prior assumptions are not understood, and distributional knowledge can be essential in finding good posterior approximations.

Since we typically want our networks to have high modelling capacity, it is natural to consider limit distributions of networks as they become large. Whilst distributions on deep networks are generally challenging to work with exactly, the limiting behaviour can lead to more insight. Further, as we shall see, finite networks used in the literature may be very close to this behaviour.

The seminal work in this area is that of Neal (1996), which showed that under certain conditions random neural networks with one hidden layer converge to a Gaussian process. The question of the type of convergence is non-trivial and part of our discussion. Historically this result was significant because it provided a connection between flexible Bayesian neural networks and Gaussian processes (Williams, 1998; Rasmussen and Williams, 2006)

1.1 Our contributions

We extend the theoretical understanding of random fully connected networks and their relationship to Gaussian processes. In particular, we prove a rigorous result (Theorem 4) on the convergence of certain sequences of finite fully connected networks with more than one hidden layer to Gaussian processes. The number of hidden layers can be any fixed number. The sizes of the hidden layers must strictly increase for each network in the sequence although the different hidden layers are allowed to grow at different rates. The weights are assumed to be independent normally distributed with their variances sensibly scaled as the network grows following the prescription of Neal (1996). The nonlinearities are assumed to obey the 'linear envelope' Condition 1 which all commonly used nonlinearities do in fact obey. Since these are the only assumptions on the sequence of networks it will be seen that the result is a meaningfully general one.

MCMC：https://zhuanlan.zhihu.com/p/75617364

Further, we empirically study the distance between finite networks and their Gaussian process analogues by using maximum mean discrepancy (MMD, Gretton et al., 2012) as a distance measure. We then systematically compare exact Gaussian process inference with 'gold standard' MCMC inference for finite Bayesian neural networks. Of the six datasets we consider, five show close agreement between the two models. Owing to the computational burden of the MCMC algorithms, the problems we can study by this method are constrained in terms of their network size, the data dimensionality and the number of data points. Nevertheless our results suggest that some experiments in the literature studied under the banner of Bayesian deep learning would have given very similar results to a Gaussian process with the appropriate kernel. A practical recommendation following from our study is that the Bayesian deep learning community should routinely compare their results to Gaussian processes with the kernels studied in this paper.

Our work is of relevance to the theoretical understanding of neural network initialisation and dynamics. It is also important in the area of Bayesian deep networks because it demonstrates that Gaussian process behaviour can arise in more situations of practical interest than previously thought. If this behaviour is desired then Gaussian process inference (exact and approximate) should also be considered in addition to standard techniques for inference in Bayesian deep learning. In some scenarios, the behaviour may not be desired because it implies a lack of a hierarchical representation and a Gaussian statistical assumption. We therefore highlight promising ideas from the literature to prevent such behaviour.

1.2 Related work

The case of random neural networks with one hidden layer was studied by Neal (1996). Cho and Saul (2009) provided analytic expressions for single layer kernels including those corresponding to a rectified linear unit (ReLU). They also studied recursive kernels designed to ‘mimic computation in large, multilayer neural nets’. As discussed in Section 3 they arrived at the correct kernel recursion through an erroneous argument. Such recursive kernels were later used with empirical success in the Gaussian process literature (Krauth et al., 2017), with a similar justification to that of Cho and Saul. The first case we are aware of using a Gaussian process construction with more than one hidden layer is the work of Hazan and Jaakkola (2015). Their contribution is similar in content to Lemma 2 discussed here, and the work has had increasing interest from the kernel community (Mitrovic et al., 2017). Recent work from Daniely et al. (2016) uses the concept of ‘computational skeletons’ to give concentration bounds on the difference in the second order moments of large finite networks and their kernel analogue, with strong assumptions on the inputs. The Gaussian process view given here, without strong input assumptions, is related but concerns not just the first two moments of a random network but the full distribution. As such the theorems we obtain are distinct. A less obvious connection is to the recent series of papers studying deep networks using a mean field approximation (Poole et al., 2016; Schoenholz et al., 2017). In those papers a second order approximation gives equivalent behaviour to the kernel recursion. By contrast, in this paper the claim is that the behaviour emerges as a consequence of increasing width and is therefore something that needs to be proved. Another surprising connection is to the analysis of self-normalizing neural networks (Klambauer et al., 2017). In their analysis the authors assume that the hidden layers are wide in order to invoke the central limit theorem. The premise of the central limit theorem will only hold approximately in layers after the first one and this theoretical barrier is something we discuss here. An area that is less related than might be expected is that of ‘Deep Gaussian Processes’ (DGPs) (Damianou and Lawrence, 2013). As will be discussed in Section 7, narrow intermediate representations mean that the marginal behaviour of DGPs is not close to that of a Gaussian process. Duvenaud et al. (2014) offer an analysis that largely applies to DGPs though they also study the Cho and Saul recursion with the motivating argument from the original paper.

Simultaneously with the submission of the previous version of our paper to ICLR 2018, at the same conference venue, Lee et al. (2018) released a paper that has overlap with our own. There are however some important differences. Empirically, whilst we compare finite Bayesian neural networks, using ‘gold standard’, asymptotically exact, sampling and MMD, to their Gaussian process analogues, Lee et al. compare finite neural networks trained with stochastic gradient descent (SGD) to Gaussian processes instead. The latter comparison to SGD is suggestive that this optimization method mimics Bayesian inference – an idea that has been receiving increasing attention (Welling and Teh, 2011; Mandt et al., 2017; Smith and Le, 2018). This is of particular importance because typically SGD is still more scalable than traditional Markov Chain based methods, enabling Lee et al. to consider some relatively large datasets. The empirical comparison of Lee et al. is therefore particularly intriguing and we hope it will lead to follow up work. Therefore, whilst there is overlap, the two papers also have independent value going forward. Theoretically, there are also important differences. Lee et al. give an argument for the Gaussian process limit, although importantly this depends on sequentially taking the number of units in each successive layer to be infinite. The proof we give here concerns the case where the layers grow simultaneously, which is arguably more relevant in practice. Note that we show a precise type of convergence – namely 'weak convergence' also called 'convergence in distribution'. Owing to the challenging nature of obtaining a full rigorous proof, the earlier version of this paper (Matthews et al., 2018) did not achieve full generality either. We needed to assume specific growth rates for the sizes of the hidden layers, and the ReLU nonlinearity. What follows here removes these assumptions and thus resolves the conjecture made in the earlier version of this work in the affirmative. The new proof method, placing a particular emphasis on exchangeability, may well be of use more generally.

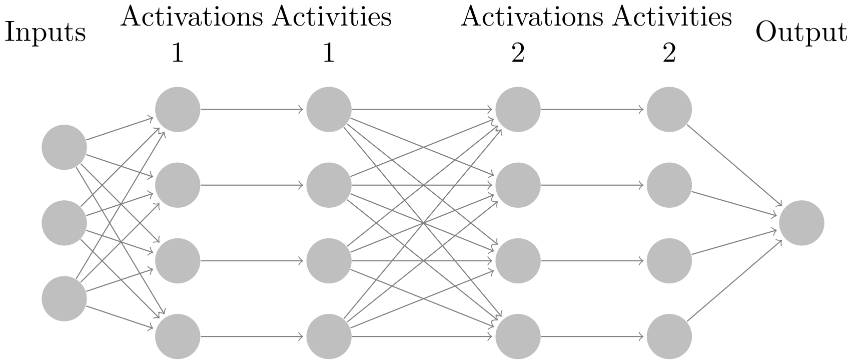

Figure 1: In this paper we consider fully connected feedforward networks with more than one hidden layer. We call the pre-nonlinearity an *activation* and post-nonlinearity an *activity.* As the network becomes increasingly wide the distribution of the marginal distributions of the activations at each layer and of the output will become close to a Gaussian process in a sense described in the text.

2. The deep wide limit

2.1 The result for one hidden layer

We consider a fully connected network as shown in Figure 1. The inputs and outputs will be real-valued vectors of dimension M and L respectively. The network is fully connected. The initial step and recursion are standard. The initial step is:

$$f_i^{(1)}(x) = \sum_{j=1}^{M} w_{i,j}^{(1)} x_j + b_i^{(1)} \,. \tag{1}$$

We make the functional dependence on x explicit in our notation as it will help clarify what follows. For a network with D hidden layers the recursion is, for each $\mu = 1, \dots, D$,

$$g_i^{(\mu)}(x) = \phi(f_i^{(\mu)}(x)) \,, \tag{2}$$

$$f_i^{(\mu+1)}(x) = \sum_{j=1}^{H_\mu} w_{i,j}^{(\mu+1)} g_j^{(\mu)}(x) + b_i^{(\mu+1)} \,, \tag{3}$$

so that $f^{(D+1)}(x)$ is the output of the network given input x. ϕ denotes the nonlinearity. In all cases the equations hold for each value of i; i ranges between 1 and H_μ in Equation (2), and between 1 and $H_{\mu+1}$ in Equation (3) except in the case of the final activation where the top value is L. The network could of course be modified to be probability simplex-valued by adding a softmax at the end.

A distribution on the parameters of the network will be assumed. Conditional on the inputs, this induces a distribution on the activations and activities. In particular we will assume independent normal distributions on the weights and biases

$$w_{i,j}^{(\mu)} \sim \mathcal{N}(0, C_w^{(\mu)}) \text{ i.i.d.} \tag{4}$$

$$b_i^{(\mu)} \sim \mathcal{N}(0, C_b^{(\mu)}) \text{ i.i.d..} \tag{5}$$

We will be interested in the behaviour of this network as the widths H_μ becomes large. The weight variances for $\mu \geq 2$ will be scaled according to the width of the network to avoid a divergence in the variance of the activities in this limit. As will become apparent, the appropriate scaling is

$$C_w^{(\mu)} = \frac{\hat{C}_w^{(\mu)}}{H_{\mu-1}} \,, \quad \mu \geq 2 \,. \tag{6}$$

The assumption is that $\hat{C}_w^{(\mu)}$ will remain fixed as we take the limit. Neal (1996) analysed this problem for $D = 1$, showing that as $H_1 \to \infty$, the values of $f_i^{(2)}(x)$, the output of the network in this case, converge to a certain multi-output Gaussian process if the activities have bounded variance.

Since our approach relies on the multivariate central limit theorem, we will arrange the relevant terms into (column) vectors to make the linear algebra clearer. Consider any two inputs x and x' and all output functions ranging over the index i. We define the vector $f^{(2)}(x)$ of length L whose elements are the numbers $f_i^{(2)}(x)$. We define $f^{(2)}(x')$ similarly. For the weight matrices defined by $w_{i,j}^{(\mu)}$ for fixed μ we use a 'placeholder' index $\bullet$ to return column and row vectors from the weight matrices. In particular $w^{(1)}_{j,\bullet}$ denotes row j of the weight matrix at depth 1. Similarly, $w^{(2)}_{\bullet,j}$ denotes column j at depth 2. The biases are given as column vectors $b^{(1)}$ and $b^{(2)}$. Finally we concatenate the two vectors $f^{(2)}(x)$ and $f^{(2)}(x')$ into a single column vector $F^{(2)}$ of size $2L$. The vector in question takes the form

$$F^{(2)} = \begin{pmatrix} f^{(2)}(x) \\ f^{(2)}(x') \end{pmatrix} = \begin{pmatrix} b^{(2)} \\ b^{(2)} \end{pmatrix} + \sum_{j=1}^{H_1} \begin{pmatrix} w^{(2)}_{\bullet,j}\phi(w^{(1)}_{j,\bullet}x + b^{(1)}_j) \\ w^{(2)}_{\bullet,j}\phi(w^{(1)}_{j,\bullet}x' + b^{(1)}_j) \end{pmatrix} . \tag{7}$$

The benefit of writing the relation in this form is that the applicability of the multivariate central limit theorem is immediately apparent. Each of the vector terms on this right hand side is independent and identically distributed conditional on the inputs x and x'. By assumption, the activities have bounded variance. The scaling we have chosen on the variances is precisely that required to ensure the applicability of the theorem, and is also in line with most commonly used initialisation strategies in practice. Therefore as H becomes large $F^{(2)}$ converges in distribution to a multivariate normal distribution. The limiting normal distribution is fully specified by its first two moments. Defining $\gamma \sim \mathcal{N}(0, C^{(1)}_b), \epsilon \sim \mathcal{N}(0, C^{(1)}_w I_M)$, the moments in question are:

$$\mathbb{E}\left[f^{(2)}_i(x)\right] = 0 \tag{8}$$

$$\mathbb{E}\left[f^{(2)}_i(x) f^{(2)}_j(x')\right] = \delta_{i,j}\left[\hat{C}^{(2)}_w \mathbb{E}_{\epsilon,\gamma}\left[\phi(\epsilon^T x + \gamma)\phi(\epsilon^T x' + \gamma)\right] + C^{(2)}_b\right] . \tag{9}$$

Note that we could have taken a larger set of input points to give a larger vector F and again we would conclude that this vector converged in distribution to a multivariate normal distribution. More formally, we can consider the set of possible inputs as an *index set*. What we have shown is that for any finite index set the distribution over functions converges to a multivariate normal. If we consider these limiting multivariate normals they obey a consistency property under marginalization. This means that the limiting distributions can be used to define a Gaussian process by the Kolmogorov extension theorem.

2.2 Definition of weak convergence of random functions

There are some important technical issues here that are not discussed in the original work of Neal (1996). In some sense, the convergence of the finite-dimensional distributions is enough if we want to answer questions about finite events, just as many of the uses of Gaussian processes within machine learning (Rasmussen and Williams, 2006) can be expressed in terms of finite-dimensional multivariate normal distributions. The reader who is content with restricting their attention to such a case may safely omit the rest of this subsection.

Given a consistent set of finite-dimensional marginals, the Kolmogorov extension theorem ensures the existence of an underlying infinite-dimensional object – a distribution over functions. If we want to make precise mathematical statements about convergence to this object some care is needed.

Firstly, the Kolmogorov theorem ensures the existence of a distribution which is uniquely defined on a specific σ-algebra, namely the *product* σ-algebra. The σ-algebra defines which events we can assign probabilities to. If we try to consider events outside the σ-algebra then the rules governing probability distributions (c.f. measures) can break down. Secondly, in abstract spaces, the definition of convergence in distribution is necessarily with respect to some *topology*. In everything that follows we will assume that this topology is generated by a metric. We also assume that the index set of the stochastic process is countably infinite. We use the metric ρ :

$$\rho(v, v') = \sum_{i=1}^{\infty} 2^{-i} \min(1, |v_i - v'_i|) \qquad \forall v, v' \in \mathbb{R}^{\mathbb{N}} \, , \tag{10}$$

This metric metrises the product topology of the product of countably many copies of $\mathbb{R}$ with the usual Euclidean topology (Dashti and Stuart, 2013). For such a countable index set, it is sufficient (Billingsley, 1999, p. 19) to prove weak convergence of the finite-dimensional marginals of the process to the corresponding multivariate Gaussian random variables. This is not generally the case if we remove the assumption of a countable index set (Billingsley, 1999, p. 19).

The restriction to countably infinite index sets means that phenomena that depend on uncountably many indices such as continuity, boundedness and differentiability are not covered by our theory. There is literature extending measures on the product σ-algebra of an uncountable index set using, for instance, the Kolmogorov continuity theorem. One could then consider proving convergence with respect to the topology in question. We do not do this in this paper but it could certainly be of interest.

2.3 The recursion lemma and the linear envelope property

In the case of a multivariate normal distribution a set of variables having a covariance of zero implies that the variables are mutually independent. Looking at Equation (9), we see that the limiting distribution has independence between different components i, j of the output. Combining this with the recursion (2), we might intuitively suggest that the next layer also converges to a multivariate normal distribution in the limit of large H_μ.

This will indeed be the case assuming that the nonlinearity does not induce heavy tail behaviour. We give an assumption on the nonlinearity that will be used throughout the sequel:

Definition 1 (Linear envelope property for nonlinearities) *A nonlinearity $\phi : \mathbb{R} \mapsto \mathbb{R}$ is said to obey the the linear envelope property if there exist $c, m \geq 0$ such that the following inequality holds*

$$|\phi(u)| \leq c + m|u| \;\; \forall u \in \mathbb{R} \, . \tag{11}$$

The majority of commonly used nonlinearities, including the sigmoid, ReLU, ELU, and SeLU nonlinearities have the linear envelope property. Intuitively the linear bounds on the nonlinearity stop it from inducing heavy tail behaviour when a random variable is passed through it. An exponential nonlinearity would not have this property. We could indeed craft a nonlinearity that is designed to violate the linear envelope property and give heavy tail behaviour. Consider, for example, the composition of the Gaussian cumulative density function (CDF) followed by the Cauchy inverse CDF. Passing a standard normal variate through such a function would, by construction, give a Cauchy distributed variable, which has an undefined mean. Whilst it may not be the most general assumption possible for what will follow, the linear envelope assumption rules in most practically used nonlinearities and, as we shall see rules out all nonlinearities for which our theory does not hold.

Next we state the following lemma, which we attribute to Hazan and Jaakkola (2015):

Lemma 2 (Normal recursion) *If the activations of a previous layer are normally distributed with moments:*

$$\mathbb{E}\left[f_i^{(\mu-1)}(x)\right] = 0 \tag{12}$$

$$\mathbb{E}\left[f_i^{(\mu-1)}(x) f_j^{(\mu-1)}(x')\right] = \delta_{i,j} K(x,x'), \tag{13}$$

Then under the recursion (2) *and as* $H \to \infty$ *the activations of the next layer converge in distribution to a normal distribution with moments*

$$\mathbb{E}\left[f_i^{(\mu)}(x)\right] = 0 \tag{14}$$

$$\mathbb{E}\left[f_i^{(\mu)}(x) f_j^{(\mu)}(x')\right] = \delta_{i,j}\left[\hat{C}_w^{(\mu)} \mathbb{E}_{(\epsilon_1,\epsilon_2)\sim\mathcal{N}(0,K)}[\phi(\epsilon_1)\phi(\epsilon_2)] + C_b^{(\mu)}\right], \tag{15}$$

where K *is a* 2×2 *matrix containing the input covariances.*

Unfortunately the lemma is not sufficient to show that the joint distribution of the activations of higher layers converge in distribution to a multivariate normal. This is because for finite H the input activations do not have a multivariate normal distribution - this is only attained (weakly or in distribution) in the limit. It could be the case that the *rate* at which the limit distribution is attained affects the distribution in subsequent layers.

Therefore the proof of our main result will require considerably more technical machinery then would be suggested by the recursion in Lemma 2. We discuss the more general result in the next section.

2.4 Convergence for more than one hidden layer

In order to state our theorem we will need one more definition, namely that of a width function:

Definition 3 (Width functions) *For a given fixed input* $n \in \mathbb{N}$*, a width function* $h_\mu : \mathbb{N} \mapsto \mathbb{N}$ *at depth* μ *specifies the number of hidden units* H_μ *at depth* μ*.*

For a given fixed input $n \in \mathbb{N}$, the set of width functions together fully specify a shape for a fully connected network. In this way, the countable sequence of natural numbers specifies a countable sequence of fully connected networks. We will be interested in the case where each of the width functions tends to infinity. Note that this includes the case of taking the width functions to be the identity, which gives the case where each hidden layer has the same number of hidden units H and H tends *jointly* to infinity rather than taking the limit in sequence. We are now ready to state the main theorem.

Theorem 4 *Consider a random deep neural network of the form in Equations* (1) *and* (2) *with a continuous nonlinearity obeying the linear envelope condition 1. Then for all sets of strictly increasing width functions h_μ and for any countable input set $(x[i])_{i=1}^{\infty}$, the distribution of the output of the network converges in distribution to a Gaussian process as $n \to \infty$. The Gaussian process has mean function zero and covariance function is given by the recursion Lemma 2.*

The convergence in distribution in the statement of the theorem is to be understood in relation to the topology induced by the metric ρ described in Expression (10). Note the generality allowed by the statement in terms of width functions. We could for instance have width functions growing at very different rates, such as $h_\mu(n) = n^\mu$. The special case in which all width functions are the identity is most common in other papers on fully connected networks and is used in the majority of our experiments. It is sufficiently important that we state it as a corollary.

Corollary 5 *Consider a random deep neural network of the form in Equations* (1) *and* (2) *with a continuous nonlinearity obeying the linear envelope condition 1 and with common number of hidden units $H_\mu = H$ for each hidden layer μ. Then for any countable input set $(x[i])_{i=1}^{\infty}$, the distribution of the output of the network converges in distribution to a Gaussian process as $H \to \infty$. The Gaussian process has mean function zero and covariance function is as in the recursion Lemma 2.*

We postpone the proof of the main theorem until Section 6. We next look at specific instances of the implied covariance function.

3. Specific kernels under recursion

Cho and Saul (2009) suggest a family of kernels based on a recurrence designed to 'mimic computation in large, multilayer neural nets'. It is therefore of interest to see how this relates to deep wide Gaussian processes. A kernel may be associated with a feature mapping $\Phi(x)$ such that $K(x, x') = \Phi(x) \cdot \Phi(x')$. Cho and Saul define a recursive kernel through a new feature mapping by compositions such as $\Phi(\Phi(x))$. However this cannot be a legitimate way to create a kernel because such a composition represents a type error. There is no reason to think the output dimension of the function Φ matches the input dimension and indeed the output dimension may well be infinite. Nevertheless, the paper provides an elegant solution to a different task: it derives closed form solution to the recursion from Lemma 2 (Hazan and Jaakkola, 2015) for the special case

$$\phi(u) = \Theta(u) u^r \text{ for } r = 0, 1, 2, 3\,, \tag{16}$$

where Θ is the Heaviside step function. Specifically, the recursive approach of Cho and Saul (2009) can be adapted by using the fact that $u^\top z$ for $z \sim \mathcal{N}(0, LL^\top)$ is equivalent in distribution to $(L^\top u)^\top \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, and by optionally augmenting u to incorporate the bias. Since $r = 1$ corresponds to rectified linear units, we apply this analytic kernel recursion in all of our experiments.

4. Measuring convergence using maximum mean discrepancy

In this section we use the kernel based two sample tests of Gretton et al. (2012) to empirically measure the similarity of finite random neural networks to their Gaussian process analogues. The maximum mean discrepancy (MMD) between two distributions $\mathcal{P}$ and $\mathcal{Q}$ is defined as:

$$\mathcal{MMD}(\mathcal{P}, \mathcal{Q}, \mathcal{H}) := \sup_{||h||_{\mathcal{H}} \leq 1} \Big[\mathbb{E}_{\mathcal{P}}[h] - \mathbb{E}_{\mathcal{Q}}[h] \Big] , \tag{17}$$

where $\mathcal{H}$ denotes a reproducing kernel Hilbert space and $|| \bullet ||_{\mathcal{H}}$ denotes the corresponding norm. It gives the biggest possible difference between expectations of a function under the two distributions under the constraint that the function has Hilbert space norm less than or equal to one. We used the unbiased estimator of squared MMD given in Equation (3) of Gretton et al. (2012).

In this experiment and where required in what follows, we take weight variance parameters $\hat{C}_w^{(\mu)} = 0.8$ and bias variance $C_b = 0.2$. We took 10 standard normal input points in 4 dimensions and pass them through 2000 independent random neural networks drawn from the distribution discussed in this paper. This was then compared to 2000 samples drawn from the corresponding Gaussian process marginal distribution. The experiment was performed with different numbers of hidden layers, different choices of monotonic width functions (which will be described in the sequel), and network sequence index $n \in \mathbb{N}$ as described in Definition 3. We repeated each experiment 20 times which allows us to reduce variance in our results and give a simple estimate of measurement error. The experiments use an RBF kernel for the MMD estimate with lengthscale 1/2. In order to help give an intuitive sense of the distances involved we also include a comparison between two Gaussian processes with isotropic RBF kernels using the same MMD distance measure. The kernel length scales for this pair of 'calibration' Gaussian processes are taken to be l and $2l$, where the characteristic length scale $l = \sqrt{8}$ is chosen to be sensible for the standard normal input distribution on the four dimensional space. Note that there are multiple different uses for kernels in this experiment. The first use is to estimate MMD, the second is for the covariance function of the calibration Gaussian processes and the third use is for the covariance function of the limit Gaussian process. The first and second cases both happen to use the RBF kernel with various length scales, but they should not be confused.

We investigated three choices of strictly increasing width functions, all of which meet the assumptions required by Theorem 4 for convergence in distribution to the corresponding Gaussian process. The identity width function $h_\mu(n) = n$ corresponds to the case where all hidden layers are the same size and n may be directly identified with the width of the network. To test a broader variety of the predictions made by the theory we introduced two other width function specifications. What we call the *largest last* width function is given by:

$$h_\mu(n) = n\mu. \tag{18}$$

For example, in a three hidden layer neural network, with $n = 50$, starting from the layer closest to the inputs, we would have hidden layer sizes $50, 100, 150$. The *largest first* width function is given by:

$$h_{\mu}(n) = n(D - \mu + 1) \tag{19}$$

For example in a three hidden layer neural network, with $n = 50$, starting from the layers closest to the inputs, we would have have hidden layer sizes $150, 100, 50$. For both the largest first and largest last width functions the sequence index n may be identified with the width of the narrowest hidden layer.

The results of the experiment are shown in Figure 2. We see that for each fixed depth the network converges towards the corresponding Gaussian process as the width increases. For the same number of hidden units per layer, the MMD distance between the networks and their Gaussian process analogue becomes higher as depth increases. The rate of convergence to the Gaussian process is slower as the number of hidden layers is increased. Unsurprisingly, since the corresponding networks will have strictly more units, both the largest last and largest first width functions converge faster than the identity width function. The largest last width function seems to converge slightly faster than the largest last width function with respect to this metric. The comparison is more interesting in this case since these two width functions have similar numbers of units. All of the results are consistent with the predictions of Theorem 4.

5. Empirical Comparison of Bayesian Deep Networks to Gaussian Processes

In this section we compare the behaviour of finite *Bayesian* deep networks of the form considered in this paper with their Gaussian process analogues. For expectations of bounded continuous functions, if we make the networks wide enough the agreement will be very close. It is also of interest, however, to consider the behaviour of networks actually used in the literature. Fully connected Bayesian deep networks with finite variance priors on the weights have been considered in several recent works (Graves, 2011; Hernández-Lobato and Adams, 2015; Blundell et al., 2015; Hernández-Lobato et al., 2016), though the specific details vary. From a Bayesian perspective, the previous section could be interpreted as using MMD as a similarity metric between priors. By constrast, in this section we will compare data dependent quantities that are typically used in Bayesian modelling practice.

We use rectified linear units and correct the variances to avoid a loss of prior variance as depth is increased. Our general strategy was to compare exact Gaussian process inference against expensive, 'gold standard', Markov Chain Monte Carlo (MCMC) methods. We choose the latter because used correctly it works well enough to largely remove questions of posterior approximation quality from the calculus of comparison. It does mean however that our empirical study does not extend to datasets which are large in terms of number of data points or dimensionality, where such inference is challenging. We therefore sound a note of caution about extrapolating our empirical finite network conclusions too confidently to this domain. On the other hand, lower dimensional, prior-dominated problems are generally regarded as an area of strength for Bayesian approaches and in this context our results are directly relevant.

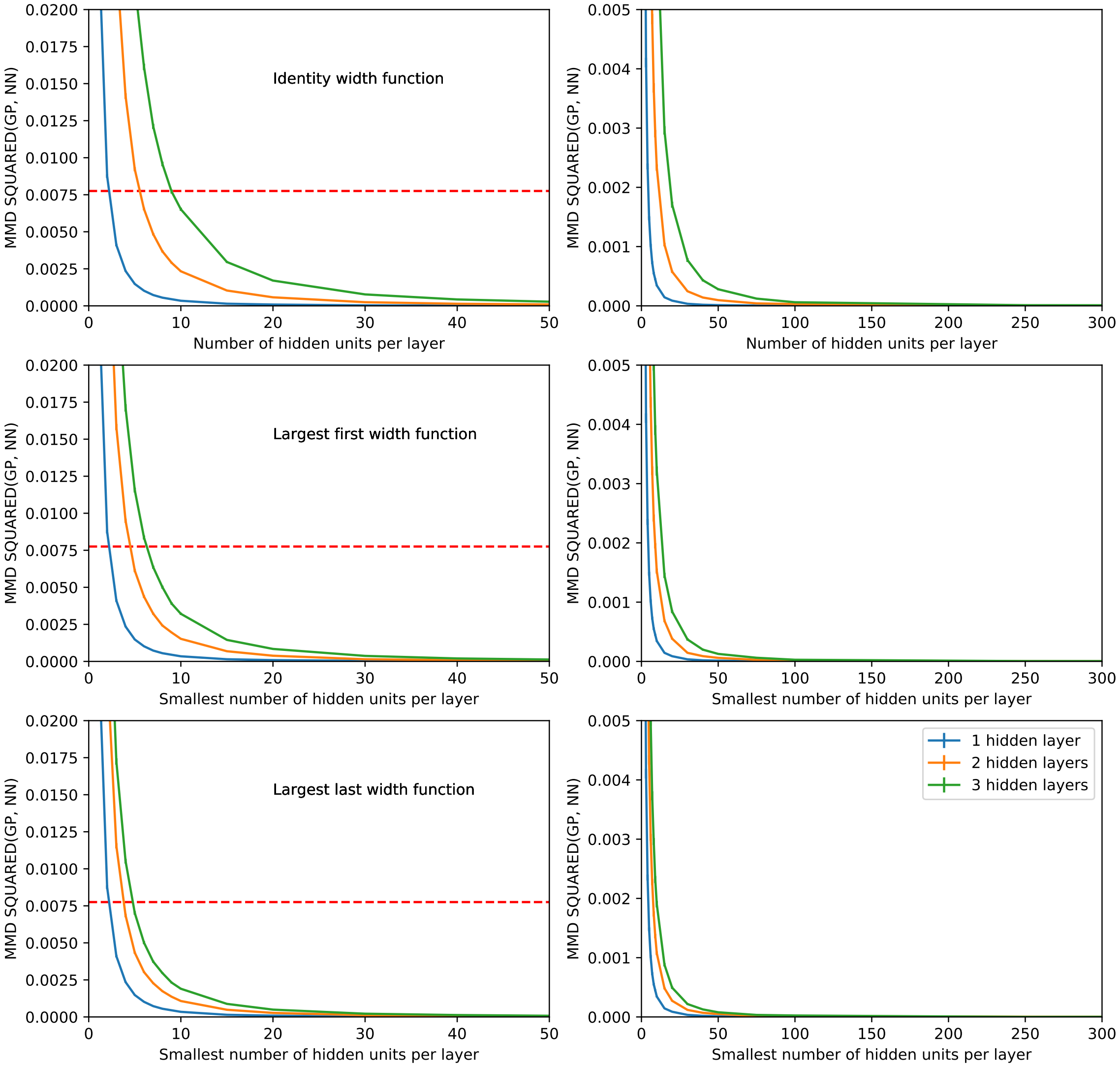

Figure 2: A comparison of finite random neural networks with ReLU nonlinearity to their corresponding Gaussian process analogue using an (RBF) kernel estimator of the squared maximum mean discrepancy (MMD). The results are consistent with the emergence of Gaussian process behaviour as the networks become wide. The red dashed line is for calibration and denotes the squared MMD between two Gaussian processes with isotropic RBF kernels and length scales l and $2l$ where $l = \sqrt{8}$ is the characteristic length scale of the input space. Different columns are different scalings of the same row plots. The rows correspond to different choices of width function. The most standard choice of the same number of hidden units per layer corresponds to the identity width function. The other width functions are described in the text. Assuming all layer sizes are strictly increasing, the independence of the choice of width function is a prediction of the theory, and is consistent with these results.

We use 3 hidden layers and 50 hidden units which is typical of the smaller Bayesian neural networks used by Hernández-Lobato and Adams (2015). Hernández-Lobato and Adams (2015) also use the variance scaling of Neal (1996) on their normally distributed weights and give a hierarchical treatment of the hyperparameters. Note that much larger networks have been used in the literature. For example, Blundell et al. (2015) use as many as 1200 units per layer, though they use a two component scale mixture of Gaussians for the weight prior. This would require an extension of our theory to non-Gaussian weight distributions for our results to be strictly applicable. Our modest choice of 50 hidden units per layer is partly also motivated by necessity. For larger networks the MCMC would be prohibitively slow.

The experiments are divided into those with fixed hyperparameters and those where the hyperparameters are learnt. The hyperparameters are specifically the noise variance, the raw weight variance $\hat{C}_w$ and the bias variance C_b. The latter two hyperparameters are shared across layers. The fixed hyperparameter experiments are the comparison most directly relevant to the theory presented here. However we found that as we moved to larger datasets both the neural network prior and the Gaussian process prior were often misspecified to an extent that made the results practically uninteresting. Since we were already computationally constrained by the neural network MCMC, we adopted the pragmatic solution of using the type II maximum likelihood parameter estimate of the Gaussian process model for both the neural network and Gaussian process priors. Although the number of hyperparameters is small, this technically adds dependency, so the fixed-hyperparameter experiments are complementary.

5.1 Experiments with fixed hyperparameters

We computed the posterior moments by the two different methods on some example datasets. For the MCMC we used Hamiltonian Monte Carlo (HMC) (Neal, 2010) updates interleaved with elliptical slice sampling (Murray et al., 2010). We considered a simple one-dimensional regression problem and a two dimensional real-valued embedding of the four data point XOR problem. To distinguish this from a later larger embedding we term this the *small XOR* dataset. We see in Figures 3 and 4 (left) that the agreement in the posterior moments between the Gaussian process and the Bayesian deep network is very close.

A key quantity of interest in Bayesian machine learning is the marginal likelihood. It is the normalising constant of the posterior distribution and gives a measure of the model fit to the data. For a Bayesian neural network, it is generally very difficult to compute, but with care and computational time it can be approximated using Hamiltonian annealed importance sampling (Sohl-Dickstein and Culpepper, 2012). The log-importance weights attained in this way constitute a stochastic lower bound on the marginal likelihood (Grosse et al., 2015). Figure 4 (right) shows the result of such an experiment compared against the (extremely cheap) Gaussian process marginal likelihood computation on the small XOR problem. The value of the log-marginal likelihood computed in the two different ways agree to within a single nat which is negligible from a model selection perspective (Grosse et al., 2015).

Predictive log-likelihood is a measure of the quality of probabilistic predictions given by a Bayesian regression method on a test point. To compare the two models we sampled 10 standard normal train and test points in 4 dimensions and passed them through a random network of the type under study to get regression targets. We then discarded the true network parameters and compared the predictions of posterior inference between the two methods. We also compared the marginal predictive distributions of a latent function value. Figure 5 shows the results. We see that the correspondence in predictive log-likelihood is close but not exact. Similarly the marginal function values are close to those of a Gaussian process but are slightly more concentrated.

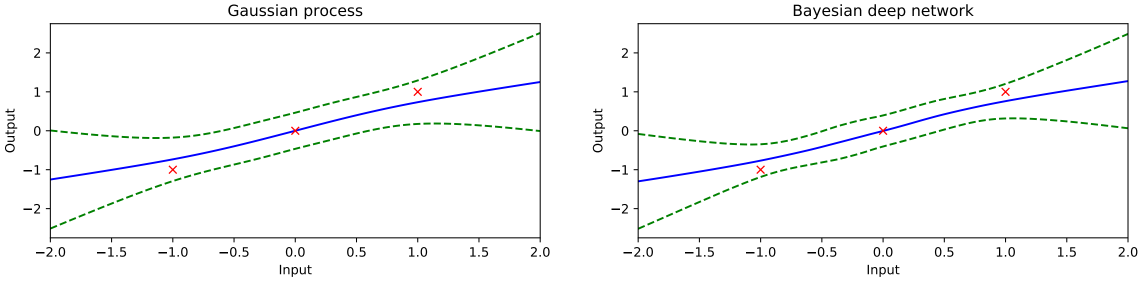

Figure 3: A comparison between Bayesian posterior inference in a Bayesian deep neural network and posterior inference in the analogous Gaussian process. The neural network has 3 hidden layers and 50 units per layer. The lines show the posterior mean and two σ credible intervals.

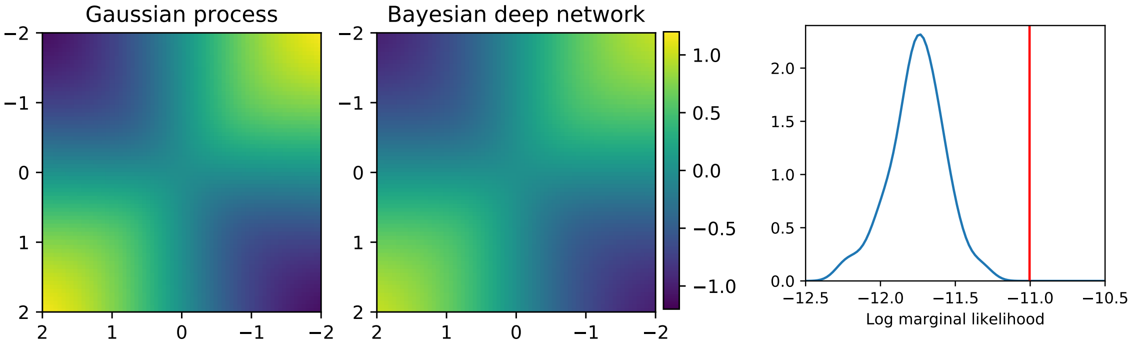

Figure 4: A comparison between posterior inference for a Gaussian process and a Bayesian deep network for the small XOR dataset, a four point real embedding of the XOR function. Left and centre: The two posterior means. The mean absolute different between the two posterior estimate grids is 0.064. Right: Kernel density estimate of the log weights from annealed importance sampling on a Bayesian deep network compared to the analogous Gaussian process marginal likelihood shown by the vertical line. The neural network has 3 hidden layers and 50 units per layer.

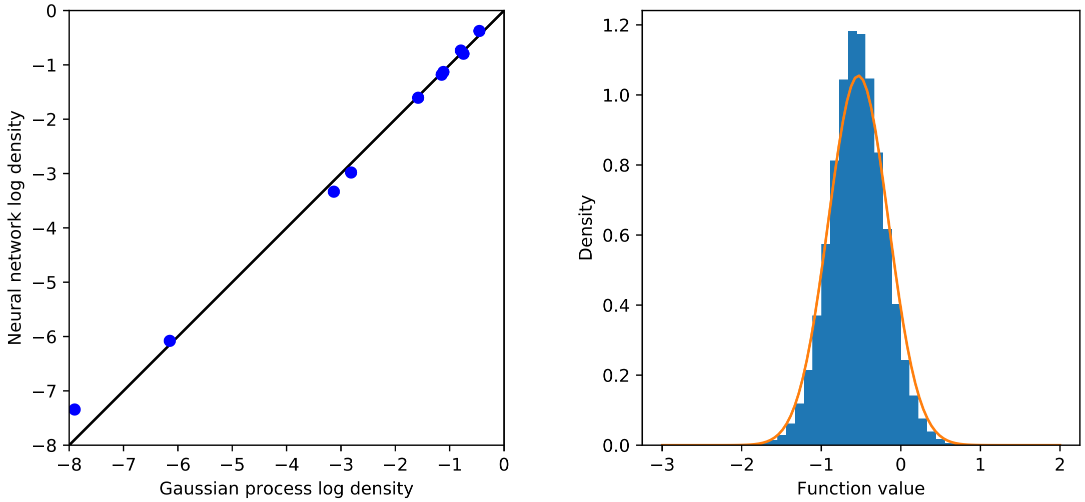

Figure 5: A comparison of the predictive distributions of a Bayesian deep network and a Gaussian process on a randomly generated test case. Left: the per-point log-densities of the two models. Right: predictive marginal distribution for the latent function on a randomly selected test point.

5.2 Experiments with learnt hyperparameters

As described above, in this section we compare neural networks and the corresponding Gaussian process on larger datasets using hyperparameters for both models that are taken from the learnt Gaussian process kernel, estimated using type II maximum likelihood.

We made a comparison for the 100 data point Snelson dataset, a regression benchmark commonly used in the sparse Gaussian process literature (Snelson and Ghahramani, 2005). Figure 6 shows that the agreement is very close.

Next we made a comparison for a larger embedding of the real valued XOR function which we term the *smooth XOR dataset*, to distinguish it from the small XOR dataset above. In detail, we have:

$$f(x_1, x_2) = -\gamma x_1 x_2 \exp\left\{-\frac{(x_1^2 + x_2^2)}{\beta}\right\} \tag{20}$$

where γ and β are chosen so that $f(-1,-1) = f(1,1) = -1$ and $f(1,-1) = f(-1,1) = 1$. One hundred input points (x_1, x_2) are sampled from a standard normal distribution and Gaussian noise of variance 0.01 is added to the outputs. In order to allow better visualisation of the posterior we take test points along two linear cross sections as shown in Figure 7. This allows us to plot the two posteriors along the cross-sections in a manner similar to a one dimensional regression problem. Figure 7 shows the results. We can see that there is again close agreement between the Bayesian neural network posterior and that of the Gaussian process.

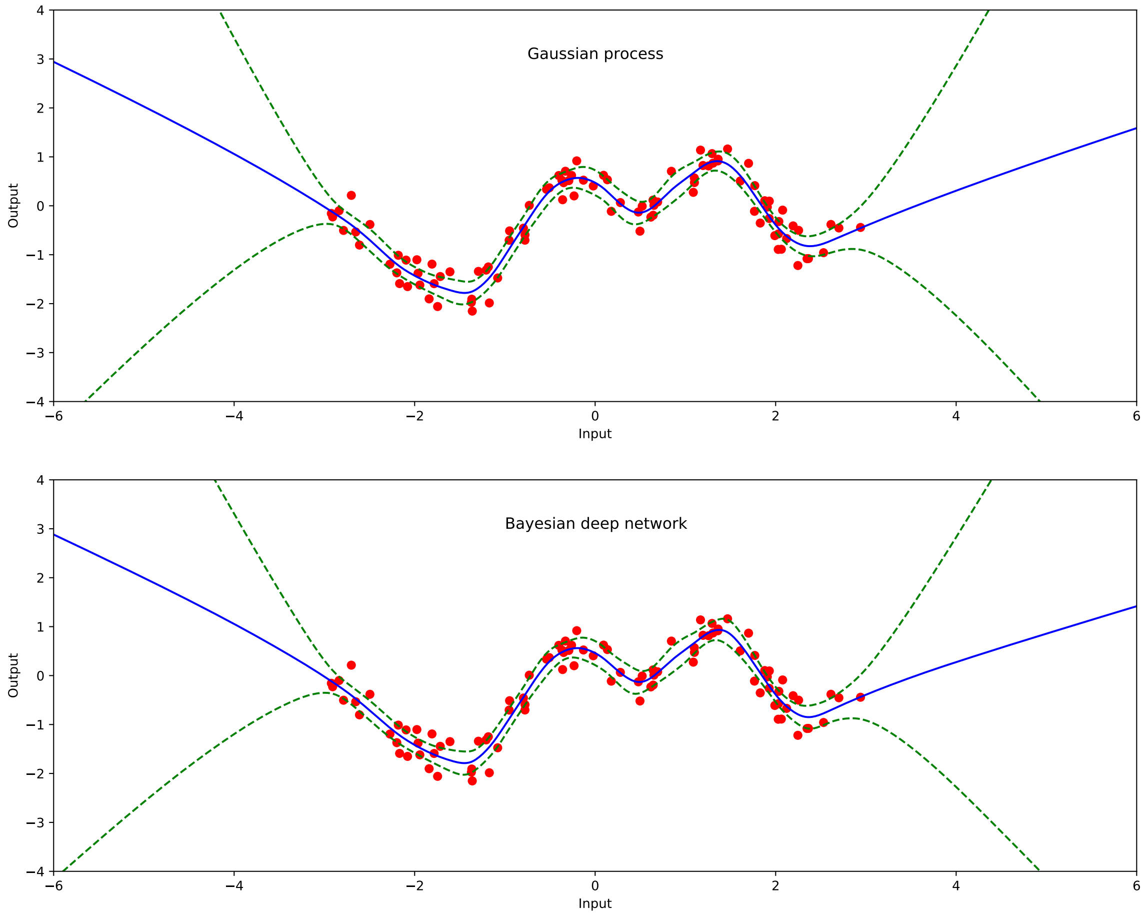

Figure 6: A comparison between Bayesian posterior inference in a Bayesian deep neural network and posterior inference in the analogous Gaussian process for the Snelson dataset. The neural network has 3 hidden layers and 50 units per layer. The lines show the posterior mean and two σ credible intervals.

Finally, we make a comparison on the Delft yacht hydrodynamics dataset. The task is to predict the residuary resistance per unit weight of displacement for a yacht hull based on six relevant attributes. We randomly partition the data into 100 training instances and 208 test instances. The data has very low noise. To make it a more challenging task for probabilistic modelling we add Gaussian noise of variance 0.01. We evaluate per test data point hold out log likelihood for both the Gaussian process and the neural network and the marginal posterior on a randomly selected test function value. The results are shown in Figure 8. The results indicate that on this dataset the Bayesian deep network and the Gaussian process do not make similar predictions. Of the two, the Bayesian neural network achieves significantly better log likelihoods on average, indicating that a finite network performs better than its infinite analogue in this case.

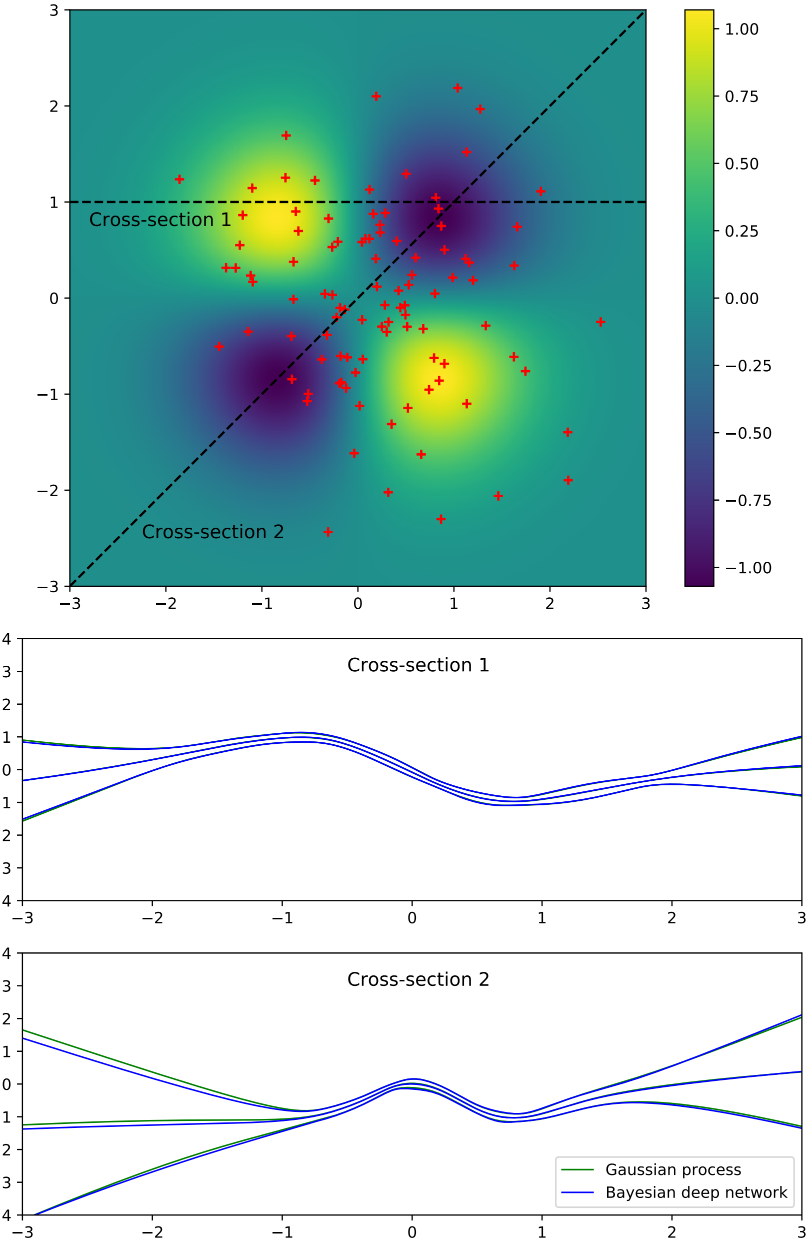

Figure 7: A comparison between the Bayesian posterior in a deep neural network and the analogous Gaussian process for the smooth XOR dataset. Top: A visualization of the smooth XOR dataset. The heat plot shows the smooth XOR function. The red crosses show the position of the training inputs. The black dashed lines show linear cross sections of the space along which we study the two posteriors. Middle and bottom: The two posteriors along the linear cross sections. In each case, the middle line is the posterior mean and the other lines represent the two σ credible intervals.

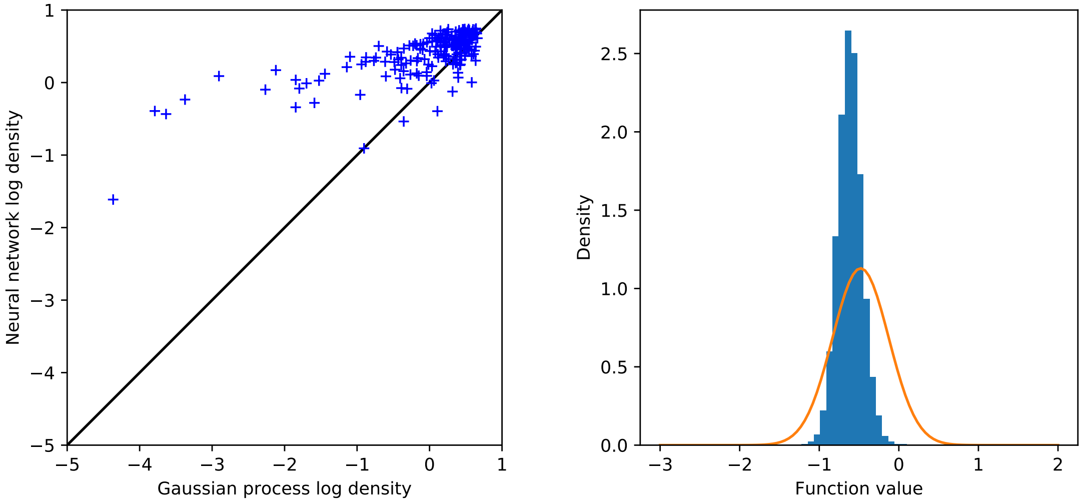

Figure 8: A comparison of the predictive distributions of a Bayesian deep network and a Gaussian process on the yacht hydrodynamics dataset. Left: the per-point log-densities of the two models. Right: predictive marginal distribution for the latent function on a randomly selected test point.

5.3 Summary and discussion of Bayesian posterior comparison

A summary of the datasets studied for comparing posteriors is given in Table 1. Of the datasets studied, the Bayesian neural network showed close agreement with the Gaussian process on five of the six datasets according to the various metrics used, the exception being the yacht dataset. It is notable that the yacht dataset has the highest dimensionality of those considered.

As already noted, our comparison method is computationally expensive as a result of the gold-standard MCMC algorithms used for Bayesian neural network inference. This means we are restricted to relatively small, low dimensional datasets. This caveat is particularly important in light of the yacht data results. On the other hand, we were also limited in the size of finite network we could consider for the same computational reason. As already discussed, the 50 hidden unit networks we use are on the small end of the range of networks that have been studied in the literature (Hernández-Lobato and Adams, 2015), compared to values as high as 1200 used in other works (Blundell et al., 2015). We would, of course, expect that where the model matches the assumption of our theory the agreement would become closer as the number of hidden units increases. As a result of the empirical analysis in Section 4, we would predict more difference if the number of hidden layers was substantially increased, though this has been relatively rare in the existing Bayesian literature thus far.

Bringing these considerations together, it seems likely that some experiments in the literature studied under the banner of Bayesian deep learning would have given very similar results to a Gaussian process with the correct kernel. In the case where the two true poste-

Dataset	Training points	Dimensionality	Learnt hyperparams	Figure
Small regression	3	1	✗	3
Small XOR	4	2	✗	4
Random	10	4	✗	5
Snelson	100	1	✓	6
Smooth XOR	100	2	✓	7
Yacht	100	6	✓	8

Table 1: Summary of datasets used for Bayesian posterior comparison.

riors are close, but the posterior approximation for the neural network is significantly worse than any approximation required for the Gaussian process, it would be expected that the Gaussian process would perform better. It should again be noted that the Bayesian neural network experiments were significantly slower than those conducted using the Gaussian process. The Snelson example took 44 hours on ten 3.2 GHz I7 CPU cores to obtain the two million samples required for the Bayesian neural network, where the Gaussian process took a matter of seconds.

Practically, we suggest that the Bayesian deep learning community routinely compare their results to Gaussian processes with the kernels studied here. This will be facilitated by the release of our covariance function code built on GPflow (Matthews et al., 2017). Such a convention would significantly increase our empirical knowledge of the phenomenon studied in this paper.

6. Proof of the main theorem

Let us first sketch the proof we will follow in this section. We first show that, for a countable set of inputs, the infinite-dimensional convergence problem can be reduced to a set of one-dimensional problems based on finite linear projections. When we examine these one-dimensional projections, we find their structure involves a sum of terms we refer to as *summands*. For fixed width functions the summands are exchangeable, which leads us to consider central limit theorems for exchangeable arrays. A result of Blum et al. (1958) plays an essential role and requires certain moment conditions, that we show by induction through the layers of the network, starting nearest the input. There is a slight complication around the correct scaling of the summands to map onto the exchangeable central limit theorem, but this can be resolved with care.

We already pointed out in Section 2.2 that with a countable index set convergence with respect to the metric ρ is equivalent to convergence of each finite-dimensional marginal. The Cramér-Wold device (Cramér and Wold, 1936) (Billingsley, 1986, p. 383) states that convergence of a sequence of finite-dimensional vectors to some limit is equivalent to convergence on all possible linear projections to the corresponding real-valued random variable. Putting these two results together we obtain the following lemma.

Lemma 6 (Convergence of finite linear projections) *Consider a sequence of random functions U_j taking values in $\mathbb{R}^Q$ each defined on a countable input set Q, with the sequence* *of functions indexed by j. Let $\mathcal{L} \subseteq Q$ be a finite subset of the input set. Further, let $\alpha \in \mathbb{R}^{\mathcal{L}}$. Then convergence in distribution of the sequence of random functions U_j taking values in $\mathbb{R}^Q$ to a limiting random function U_* with respect to the metric ρ is equivalent to weak convergence of $\sum_{u \in \mathcal{L}} U_j(u)\alpha_u$ to the corresponding finite linear projection $\sum_{u \in \mathcal{L}} U_*(u)\alpha_u$ for every such $\mathcal{L}$ and α.*

Therefore our task is reduced to that of proving convergence of a sequence of real-valued random variables to another real-valued random variable – a considerable simplification. In particular, we will leverage a theorem of Blum et al. (1958) on central limit theorems for exchangeable sequences.

It will be convenient to consider a sequence of 'infinite width, finite fan-out, networks'. By this we mean that the indices i in the recursion (2) can be thought of as running over all natural numbers instead of just up to H_μ (hence infinite width). The limits of the sums in the recursion will retain the same finite values, which depend on the width functions evaluated at some n (hence finite fan-out). This makes only a superficial change because it adds extra copies of the same variables at each depth. For fixed n, these extra variables will not effect the downstream distribution of the network. The change is however useful in the book-keeping needed to prove convergence. We have defined a countable sequence of such networks because n is a natural number.

It will also be useful to slightly rewrite the defining initialisation and recursion (2) from the more familiar form to one which is easier to manipulate:

$$f_i^{(1)}(x) = \sum_{j=1}^{M} \epsilon_{i,j}^{(1)} x_j \sqrt{\hat{C}_w^{(1)}} + b_i^{(1)}\,, \;\; i \in \mathbb{N}\,, \tag{21}$$

and:

$$g_i^{(\mu)}(x) = \phi(f_i^{(\mu)}(x))\,, \tag{22}$$

$$f_i^{(\mu+1)}(x) = \frac{1}{\sqrt{h_\mu(n)}} \sum_{j=1}^{h_\mu(n)} \epsilon_{i,j}^{(\mu+1)} g_j^{(\mu)}(x) \sqrt{\hat{C}_w^{(\mu+1)}} + b_i^{(\mu+1)}\,, \;\; i \in \mathbb{N}\,, \tag{23}$$

where:

$$\epsilon_{i,j}^{(\mu)} \sim \mathcal{N}(0,1) \;\; \text{i.i.d} \; \forall \mu, i, j\,. \tag{24}$$

This amounts to reparameterising the weights in terms of standard normals and making the previously mentioned infinite extension of the width variable i. We re-emphasize that neither step changes the distribution over final function values. With the aim of mapping onto Lemma 6 we make the following definitions:

Definition 7 (Projections and summands) *The projections are defined in terms of a finite linear projection of the function values without biases:*

$$\mathcal{T}^{(\mu)}(\mathcal{L},\alpha)[n] = \sum_{(x,i)\in\mathcal{L}} \alpha^{(x,i)} \left[f_i^{(\mu)}(x)[n] - b_i^{(\mu)} \right]. \tag{25}$$

where $\mathcal{L} \subset \mathcal{X} \times \mathbb{N}$ is a finite set of tuples of data points and indices of pre-nonlinearities, with $\mathcal{X} = (x[i])_{i=1}^{\infty}$. $\alpha \in \mathbb{R}^{|\mathcal{L}|}$ is a vector parameterising the linear projection. The suffix $[n]$ indicates that the corresponding width functions are instantiated with input n.

The summands are defined as:

$$\gamma_j^{(\mu)}(\mathcal{L},\alpha)[n] := \sum_{(x,i)\in\mathcal{L}} \alpha^{(x,i)} \epsilon_{i,j}^{(\mu)} g_j^{(\mu-1)}(x)[n] \sqrt{\hat{C}_w^{(\mu)}}\,, \tag{26}$$

in order to ensure the summation relation

$$\mathcal{T}^{(\mu)}(\mathcal{L},\alpha)[n] := \frac{1}{\sqrt{h_{\mu-1}(n)}} \sum_{j=1}^{h_{\mu-1}(n)} \gamma_j^{(\mu)}(\mathcal{L},\alpha)[n]\,. \tag{27}$$

The last relation follows from applying the definitions and re-arranging the order of summation. Note the similarity between the definition of projections used here and in Lemma 6. We next show that the summands are exchangeable.

Lemma 8 (Exchangeability of summands) *For each fixed n and $\mu \in \{2,\ldots,D+1\}$, the countable sequence of summands $\gamma_j^{(\mu)}(\mathcal{L},\alpha)[n]$ are an exchangeable sequence with respect to the index j.*

Proof To prove the lemma we use de Finetti's theorem, which states that a sequence of random variables is exchangeable if and only if they are i.i.d. conditional on some set of random variables. It is therefore sufficient to exhibit such as set of random variables. To do this we apply the recursion. Removing some multiplicative constants we have:

$$\gamma_j^{(\mu)}(\mathcal{L},\alpha)[n] \propto \sum_{(x,i)\in\mathcal{L}} \alpha^{(x,i)} \epsilon_{i,j}^{(\mu)} g_j^{(\mu-1)}(x)[n] \tag{28}$$

$$= \sum_{(x,i)\in\mathcal{L}} \alpha^{(x,i)} \epsilon_{i,j}^{(\mu)} \phi \left(\frac{1}{\sqrt{h_{\mu-2}(n)}} \sum_{j=1}^{h_{\mu-2}(n)} \epsilon_{j,k}^{(\mu-1)} g_k^{(\mu-2)}(x)[n] \sqrt{\hat{C}_w^{(\mu-1)}} + b_j^{(\mu-1)} \right), \tag{29}$$

with the convention that $h_0(n) = M$ and $g_k^{(0)}(x) = x_k$ for $k = 1,\ldots,M$. Conditional on the finite set of random variables $\left\{ g_k^{(\mu-2)}(x)[n] : k = 1,...,H_{\mu-2}, x \in \mathcal{L}_{\mathcal{X}} \right\}$ (where $\mathcal{L}_{\mathcal{X}}$ is the set of inputs points in $\mathcal{L}$), the summands are independent and identically distributed.

■

Thus we are led to consider central limit theorems for sequences of exchangeable sequences. The work of Blum et al. (1958) will provide our starting point.

Theorem 9 (CLT for exchangeable sequences (Blum et al., 1958)) *For each positive integer n let $(X_{n,i}; i = 1, 2, ...)$ be an infinitely exchangeable process with mean zero, variance one, and finite absolute third moment. Define*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{n,i}. \tag{30}$$

Then if the following conditions hold:

1. $\mathbb{E}_n[X_{n,1} X_{n,2}] = \mathrm{o}(\frac{1}{n})$
2. $\lim_{n \to \infty} \mathbb{E}_n \left[X_{n,1}^2 X_{n,2}^2 \right] = 1$
3. $\mathbb{E}_n \left[|X_{n,1}|^3 \right] = \mathrm{o}(\sqrt{n})$

Then S_n converges in distribution to a standard normal.

This is effectively a generalisation of the classical CLT from independent identically distributed variables to the more general class of exchangeable ones. We will need to address the fact that the theorem applies to unit variance variables and that we have non-identity width functions. The next lemma adapts the work of Blum et al. to our specific requirements.

Lemma 10 (Adapted CLT for sequences of exchangeable sequences) *For each positive integer n let $(X_{n,i}; i = 1, 2, ...)$ be an infinitely exchangeable process with mean zero, finite variance σ_n^2, and finite absolute third moment. Suppose also that the variance has a limit $\lim_{n \to \infty} \sigma_n^2 = \sigma_*^2$. Define*

$$S_n = \frac{1}{\sqrt{h(n)}} \sum_{i=1}^{h(n)} X_{n,i}\,, \tag{31}$$

where $h : \mathbb{N} \mapsto \mathbb{N}$ is a strictly increasing function. Then if the following conditions hold:

a) $\mathbb{E}_n[X_{n,1} X_{n,2}] = 0$

b) $\lim_{n \to \infty} \mathbb{E}_n \left[X_{n,1}^2 X_{n,2}^2 \right] = \sigma_*^4$

c) $\mathbb{E}_n \left[|X_{n,1}|^3 \right] = \mathrm{o}(\sqrt{h(n)})$

Then S_n converges in distribution to $\mathcal{N}(0, \sigma_*^2)$, where $\mathcal{N}(0, 0)$ is interpreted as converging to 0.

We postpone the proof of Lemma 10 until Appendix A. Our next step will be to apply Lemma 10 to the projections and summands by showing they meet each condition. We first establish the existence of a limiting variance.

Lemma 11 (Limiting variance) *The limiting variance, defined as*

$$\sigma^2(\mu, \mathcal{L}, \alpha)[*] := \lim_{n \to \infty} \sigma^2(\mu, \mathcal{L}, \alpha)[n] \,, \tag{32}$$

exists, where $\sigma^2(\mu, \mathcal{L}, \alpha)[n]$ is the variance of the random variables $\gamma_j^{(\mu)}(\mathcal{L}, \alpha)[n]$, and has the value

$$\sigma^2(\mu, \mathcal{L}, \alpha)[*] = \alpha^T K(\mathcal{L}) \alpha \,, \tag{33}$$

where $K \in \mathbb{R}^{\mathcal{L} \times \mathcal{L}}$ is the Gram matrix implied by the recursion 2 without a bias correction on the final layer.

The proof of this Lemma can be found in Appendix B.1.

Lemma 12 (Convergence in distribution of projections) *As $n \to \infty$ the projection $\mathcal{T}^{(\mu)}(\mathcal{L}, \alpha)[n]$ converges in distribution to $\mathcal{N}(0, \sigma^2(\mu, \mathcal{L}, \alpha)[*])$.*

The full details of Lemma 12 are explained in Appendix B.1. Here we outline the key points of the approach. We apply Lemma 10 to the projections, using the fact that the summands are exchangeable for each n and with the limiting variance $\sigma^2(\mu, \mathcal{L}, \alpha)[*]$ derived in Lemma 11. Condition a) of Lemma 10 follows straightforwardly from the fact that the summands are uncorrelated. That Condition c) is fulfilled is intuitively reasonable given that we in fact expect this absolute third moment to tend to a constant. Condition c) will however still need to be shown carefully. This leaves Condition b). Convergence of the expectation of a sequence of random variables can be ensured if the sequence is uniformly integrable and the sequence converges in distribution (Billingsley, 1999). Thus the main work of Appendix B.1 is to prove these conditions in our case, by induction forwards through the network.

Lemma 12 shows consistency of convergence of the finite linear projections of the pre-bias function distribution with the stated Gaussian process. By Lemma 6, this is sufficient for convergence in distribution to the Gaussian process. As the biases are normally distributed it is straightforward to add them and get the final result. Therefore we are done.

7. Desirability of Gaussian process behaviour and methods to avoid it

When using deep Bayesian neural networks as priors, the emergence of Gaussian priors raises important questions in the cases where it is applicable, even if one sets aside questions of computational tractability. The kernels considered in this paper have not been commonly used in the Gaussian process literature and warrant further analysis. It has been argued by previous authors that there are important cases where kernel machines with *local* kernels will perform badly (Bengio et al., 2005). The analysis applies to the posterior mean of a Gaussian process. The kernels considered in this paper do not meet the strict definition of what could be considered local, though the Euclidean inner product between two points is sufficient to compute the corresponding covariance. In any case, the fact remains that a Gaussian process with a fixed kernel does not use a learnt hierarchical representation. Such representations are widely regarded to be essential to the success of deep learning. A complication to consider is when a hierarchical treatment of the model is taken, learning model hyperparameters. Typically only a few such hyperparameters are used and it seems unlikely this could offer the same benefits as full representation learning. Using significantly more hyperparameters would move the model beyond the scope of this paper. MacKay (2002, p. 547) famously reflected on what is lost when taking the Gaussian process limit of a single hidden layer network, remarking that Gaussian processes will not learn hidden features. Neal (1996, p. 43) makes similar comments and also expresses the hope that Bayesian neural networks could expand the range of probabilistic models beyond Gaussian processes. In light of the results in this paper for networks with more than one hidden layer these considerations are of considerable importance going forward.

There is literature on learning the representation of a standard, usually structured, network composed with a Gaussian process (Wilson et al., 2016a,b; Al-Shedivat et al., 2017). This differs from the assumed paradigm of this paper, where all model complexity is specified probabilistically and we do not assume convolutional, recurrent or other problem specific structure.

Within the paradigm considered here, the question therefore arises as to what can be done to avoid marginal Gaussian process behaviour if it is not desired. Speaking loosely, to stop the onset of the central limit theorem and the approximate analogues discussed in this paper one needs to make sure that one or more of its conditions is far from being met. Since the chief conditions on the summands are independence, bounded variance and many terms, violating these assumptions will remove Gaussian process behaviour. Deep Gaussian processes (Damianou and Lawrence, 2013) are not close to standard Gaussian processes marginally because they are typically used with narrow intermediate layers. It can be challenging to choose the precise nature of these narrow layers a priori. Neal (1996) suggests using networks with infinite variance in the activities. With a single hidden layer and correctly scaled, these networks become alpha stable processes in the wide limit. Neal also discusses variants that destroy independence by coupling weights. These alternatives each arguably have a mechanism to discover hierarchies of features. Again, given the convergence results for multiple hidden layer networks from this paper, there is now further motivation to study the non-Gaussian alternatives as well.

8. Conclusions

Studying the limiting behaviour of distributions on feedforward neural networks has been a fruitful avenue for understanding these models historically. In this paper we have formalised and extended prior results by Neal (1996) to deep networks. In particular, we have shown that, under broad conditions, as we make the architecture increasingly wide, the implied random function converges in distribution to a Gaussian process. Our empirical study using MMD suggests that this behaviour is exhibited in a variety of models of size comparable to networks used in the literature. This led us to juxtapose finite Bayesian neural networks with their Gaussian process analogues. In several cases there was close agreement, leading us to conclude that it is likely some results from the existing Bayesian deep learning literature would be very similar to those obtained with the corresponding Gaussian process model.

We recommend that empirical investigation of Bayesian neural networks should routinely include comparison to their Gaussian process analogue. If Gaussian process behaviour is desired then exact and approximate inference using the analytic properties of Gaussian processes should be considered as an alternative to neural network inference. Since Gaussian processes have an equivalent flat representation then in the context of deep learning there may well be cases where the behaviour is not desired and steps should be taken to avoid it.

We view these results as a new opportunity to further the understanding of neural networks in the work that follows. Initialisation and learning dynamics are crucial topics of study in modern deep learning which require that we understand random networks. Bayesian neural networks should offer a principled approach to generalisation but this relies on successfully approximating a clearly understood prior. In illustrating the continued importance of Gaussian processes as limit distributions, we hope that our results will further research in these broader areas.

9. Acknowledgements

We wish to thank Neil Lawrence and Jascha Sohl-Dickstein for helpful conversations. We also thank anonymous reviewers of previous versions for their insights. Alexander Matthews and Zoubin Ghahramani acknowledge the support of EPSRC Grant EP/N014162/1 and EPSRC Grant EP/N510129/1 (The Alan Turing Institute). Jiri Hron holds a Nokia CASE Studentship. Mark Rowland acknowledges support by EPSRC grant EP/L016516/1 for the Cambridge Centre for Analysis. Richard E. Turner is supported by Google as well as EPSRC grants EP/M0269571 and EP/L000776/1.

References

Maruan Al-Shedivat, Andrew G. Wilson, Yunus Saatchi, Zhiting Hu, and Eric P. Xing. Learning Scalable Deep Kernels with Recurrent Structure. *Journal of Machine Learning Research (JMLR)*, 2017.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. The Curse of Dimensionality for Local Kernel Machines. Technical Report 1258, Département d'informatique et recherche opérationnelle, Université de Montréal, 2005.

Patrick Billingsley. *Probability and Measure.* John Wiley and Sons, second edition, 1986.

Patrick Billingsley. *Convergence of Probability Measures.* John Wiley & Sons Inc., Second edition, 1999.

J. R. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher. Central limit theorems for interchangeable processes. *Canadian Journal of Mathematics*, 10:222–229, 1958.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *International Conference on Machine Learning (ICML)*, 2015.

Youngmin Cho and Lawrence K. Saul. Kernel Methods for Deep Learning. *Advances in Neural Information Processing Systems (NIPS)*, 2009.

H. Cramér and H. Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, s1-11(4):290–294, 1936.

Andreas C. Damianou and Neil D. Lawrence. Deep Gaussian Processes. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. *Advances in Neural Information Processing Systems (NIPS)*, 2016.

M. Dashti and A. M. Stuart. The Bayesian Approach To Inverse Problems. *ArXiv e-prints*, February 2013.

David Duvenaud, Oren Rippel, Ryan P. Adams, and Zoubin Ghahramani. Avoiding Pathologies in very Deep Networks. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

Alex Graves. Practical Variational Inference for Neural Networks. *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-sample test. *Journal of Machine Learning Research (JMLR)*, 2012.

Roger B. Grosse, Zoubin Ghahramani, and Ryan P. Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *ArXiv e-prints*, November 2015.

Tamir Hazan and Tommi Jaakkola. Steps Toward Deep Kernel Methods from Infinite Neural Networks. *ArXiv e-prints*, August 2015.

José M. Hernández-Lobato and Ryan P. Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. *International Conference on Machine Learning (ICML)*, 2015.

José M. Hernández-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard E. Turner. Black-box alpha divergence minimization. *International Conference on Machine Learning (ICML)*, 2016.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. 2017.

Karl Krauth, Edwin V. Bonilla, Kurt Cutajar, and Maurizio Filippone. AutoGP: Exploring the capabilities and limitations of Gaussian Process models. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. *International Conference on Learning Representations (ICLR)*, 2018.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms.* Cambridge University Press, 2002.

S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *ArXiv e-prints*, April 2017.

Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian Process Library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian Process Behaviour in Wide Deep Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Deep Kernel Machines via the Kernel Reparametrization Trick. In *International Conference on Learning Representations (ICLR) Workshop Track*, 2017.

Iain Murray, Ryan P. Adams, and David J. C. MacKay. Elliptical Slice Sampling. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Radford M. Neal. *Bayesian Learning for Neural Networks.* Springer, 1996.

Radford M. Neal. MCMC using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*, 2010.

Ben Poole, Subhaneil Lahiri, Maithreyi Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in Deep Neural Networks through Transient Chaos. *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2006.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep Information Propagation. *International Conference on Learning Representations (ICLR)*, 2017.

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. *International Conference on Learning Representations (ICLR)*, 2018.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Jascha Sohl-Dickstein and Benjamin J. Culpepper. Hamiltonian Annealed Importance Sampling for partition function estimation. *CoRR*, abs/1205.1925, 2012.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. *International Conference on Machine Learning (ICML)*, 2011.

Christopher K. I. Williams. Computing with Infinite Networks. *Advances in Neural Information Processing Systems (NIPS)*, 1998.

Andrew G. Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep Kernel Learning. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016a.

Andrew G. Wilson, Zhiting Hu, Ruslan R. Salakhutdinov, and Eric P. Xing. Stochastic Variational Deep Kernel Learning. *Advances in Neural Information Processing Systems (NIPS)*, 2016b.

Appendix A. Adapting the exchangeable CLT of Blum et al. 1958.

This section gives further detail on our adaption of Theorem 9 to our specific needs. It states and proves an intermediate Lemma 13 and then using that lemma gives the postponed proof of Lemma 10.

Lemma 13 (Variance adapted CLT for sequences of exchangeable sequences) *For each positive integer n let $(X_{n,i}; i = 1, 2, ...)$ be an infinitely exchangeable process with mean zero, finite variance σ_n^2, and finite absolute third moment. Suppose also that the variance has a limit $\lim_{n\to\infty} \sigma_n^2 = \sigma_*^2$. Define*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{n,i}. \tag{34}$$

Then if the following conditions hold:

i. $\mathbb{E}_n[X_{n,1}X_{n,2}] = 0$

ii. $\lim_{n\to\infty} \mathbb{E}_n\left[X_{n,1}^2 X_{n,2}^2\right] = \sigma_*^4$

iii. $\mathbb{E}_n\left[|X_{n,1}|^3\right] = \mathrm{o}(\sqrt{n})$

Then S_n converges in distribution to $\mathcal{N}(0, \sigma_*^2)$, where $\mathcal{N}(0, 0)$ is interpreted as converging to 0.

Proof [Proof of Lemma 10] Either $\sigma_*^2 = 0$ or it does not. We deal with each case separately.

In the case where $\sigma_*^2 = 0$, we have:

$$\begin{aligned}
\mathrm{Var}[S_n] &= \frac{1}{n}\mathrm{Var}\left[\sum_{i=1}^{n} X_{n,i}\right] &(35)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n} \mathrm{Var}[X_{n,i}] + \sum_{i\neq i'} \mathrm{Cov}[X_{n,i}, X_{n,i'}]\right) &(36)\\
&= \mathrm{Var}[X_{n,1}] = \sigma_n^2, &(37)
\end{aligned}$$

where we have used property (i) that the distinct elements in a row are uncorrelated. Now the proof can take a similar route to that used in proving the weak law of large numbers with finite variance. Chebyshev's inequality we have:

$$\Pr(|S_n| \leq \beta) \leq \frac{\sigma_n^2}{\beta^2}\,, \tag{38}$$

for all $\beta > 0$, So that:

$$\Pr(|S_n| > \beta) \leq 1 - \frac{\sigma_n^2}{\beta^2}\,. \tag{39}$$

That is to say S_n converges in probability to 0. In such a case of a constant target convergence in probability is equivalent to convergence in distribution.

In the case where $\sigma_*^2 \neq 0$ there is always some M such that for $n \geq M$ $\sigma_n^2 > 0$ by the definition of a limit. Let us assume we are in this range of n. Then the standardised values $\frac{X_{n,i}}{\sigma_n}$ will obey the conditions of Theorem 9, which we now show term by term. We clearly have mean zero, unit variance and finite third moment, and conditions 1) and i) are identical. This leaves us to validate conditions 2) and 3). Starting from ii) we have:

$$\lim_{n\to\infty} \mathbb{E}_n\left[X_{n1}^2 X_{n2}^2\right] = \sigma_*^4 \tag{40}$$

$$\lim_{n\to\infty} \mathbb{E}_n\left[X_{n1}^2 X_{n2}^2\right] \lim_{n\to\infty} \mathbb{E}_n\left[\frac{1}{\sigma_n^4}\right] = 1 \tag{41}$$

$$\lim_{n\to\infty} \mathbb{E}_n\left[\frac{X_{n1}^2 X_{n2}^2}{\sigma_n^4}\right] = 1 \tag{42}$$

which clearly implies condition 2). Starting from condition iii) we have:

$$\lim_{n\to\infty}\left[\frac{|X_{n1}|^3}{\sqrt{n}}\right] = 0 \tag{43}$$

$$\lim_{n\to\infty}\left[\frac{|X_{n1}|^3}{\sigma_n^3\sqrt{n}}\right] = \frac{1}{\sigma_*^3}\lim_{n\to\infty}\left[\frac{|X_{n1}|^3}{\sqrt{n}}\right] = 0 \tag{44}$$

which implies condition 3). Since multiplication by a constant is a continuous function we have therefore showed that $S_n\frac{\sigma_*}{\sigma_n}$ converges in distribution to $\mathcal{N}(0, \sigma^2(\mu, \mathcal{L}, \alpha)[*])$.

Note that the sequence $|S_n\frac{\sigma_*}{\sigma_n} - S_n|$ converges *surely* to 0. This certainly implies convergence in probability of the same sequence to zero. We can therefore invoke a general result on convergence of sequences that says if a sequence of random variables X_i converges to X_* and $|X_i - Y_i|$ converges in probability to zero, then Y_i converges in distribution to X_* (Vaart, 1998).

■

Proof [Proof of Lemma 10] Lemma 13 applies to what are known as triangular arrays in the literature. This lemma is the generalisation to arrays that are not strictly triangular. To do this we embed the non-triangular array in a large triangular one. We fill the extra spaces with standard normal random variables. This gives an interleaved sequence. The terms we actually care about will obey the necessary conditions if conditions 1) 2) and 3) if a) b) and c) hold. The conditions 1) 2) and 3) will hold trivially for the standard normal rows. Thus the whole sequence converges in distribution. But since any subsequence also converges in distribution we get our required result. ■

Appendix B. Details of the proof of Theorem 4

Here, we summarise the high-level structure of the proof of Theorem 4. The argument is inductive, showing sequentially that the hidden units in each layer of the network converge in distribution; to avoid repetition, all mentions of convergence in distribution of infinite-dimensional random variables in what follows are specifically with respect to the topology generated by the metric ρ introduced in Section 2.2. The main part of the inductive argument is summarised in the following proposition.

Proposition 14 *For any $\mu \in \{2, \dots, \mu + 1\}$, suppose that the collection of random variables $\{f_i^{(\mu-1)}(x)[n]\}_{i\in\mathbb{N},x\in\mathcal{X}}$ converges in distribution as $n \to \infty$ to a centred Gaussian with covariance function of the form given in Lemma 2. Then any finite linear combination $\mathcal{T}^{(\mu)}(\mathcal{L},\alpha)[n]$ (with $\mathcal{L} \subset \mathcal{X} \times \mathbb{N}$ finite and $\alpha \in \mathbb{R}^{\mathcal{L}}$) of pre-nonlinearities at the next layer also converges in distribution to a centred Gaussian of the form described in Lemma 11.*

Note that the conclusion of Proposition 14 leads to the statement of Lemma 12. By the Cramér-Wold device discussed in Section 6, the convergence of the finite linear projections established in Proposition 14 guarantees convergence of all finite-dimensional marginal distributions. Adding in the independent bias terms yields convergence of finite-dimensional marginals of the pre-activations at layer μ; this may be demonstrated via a standard argument using characteristic functions. Due to the remarks on weak convergence in Section 2.2, convergence in distribution of all finite-dimensional marginals guarantees convergence in distribution of the full collection of random variables $\{f_i^{(\mu)}(x)[n]\}_{i\in\mathbb{N},x\in\mathcal{X}}$ in the next layer, completing the inductive step.

The proof of Theorem 4 is then concluded by observing that the pre-nonlinearities in the first hidden layer, $\{f_i^{(1)}(x)[n]\}_{i\in\mathbb{N},x\in\mathcal{X}}$, have a fixed Gaussian distribution that does not depend on n.

We thus turn our attention to proving Proposition 14. The main idea is to use Lemma 10, taking each of the random variables $X_{n,i}$ (for $i \in \mathbb{N}, n \in \mathbb{N}$) appearing in the statement of the Lemma to be the summands appearing in the finite linear projections $\mathcal{T}^{(\mu)}(\mathcal{L},\alpha)[n]$:

$$X_{n,i} = \gamma_i^{(\mu)}(\mathcal{L},\alpha)[n]\,. \tag{45}$$

Addressing the conditions of Lemma 10, we note that the exchangeability condition is provided by Lemma 8, the mean-zero condition is immediate, the limiting variance condition is dealt with by Lemma 11. Condition a) of Lemma 10 holds trivially as the random variables X_{n1} and X_{n2} are mean-zero and uncorrelated. The remaining conditions of Lemma 10 are dealt with through the following results; Lemma 15 deals with Condition b), whilst Lemma 16 deals with the growth of third absolute moments as required by Condition c).

Lemma 15 (Convergence of $\mathbb{E}\left[|X_{n,1}X_{n,2}|^2\right]$) *Consider arbitrary $\mu \in \{2, \dots, D+1\}$ and the corresponding set of random variables $\{f_i^{(\mu)}(x)[n]\}_{(i,x)\in\mathcal{L}}$. Assume that the countably infinite vector of random variables $\{f_i^{(\mu-1)}(x)[n]\}_{i\in\mathbb{N},x\in\mathcal{X}}$ converges in distribution to a centred Gaussian process with covariance specified by the recursion in Lemma 2 as $n \to \infty$. Then*

$$\lim_{n\to\infty} \mathbb{E}\left[|X_{n,1}X_{n,2}|^2\right] = \sigma_*^4 .$$

Lemma 16 (Bound on $\mathbb{E}\left[|X_{n,1}|^3\right]$) *For arbitrary given α, $\mathcal{L}$, and $\mu \in \{1, 2, \dots, D+1\}$, $\mathbb{E}\left[|X_{n,1}|^3\right] < c < \infty$ with c independent of n. Thus $\mathbb{E}\left[|X_{n,1}|^3\right] = \mathrm{o}(\sqrt{h(n)})$.*

Thus, all that remains to establish Theorem 4 is the proof of these intermediate lemmas; the proofs are given in the sections that follow.

B.1 Proofs of main lemmas and corollaries

Throughout this section, we simplify the notation by defining

$$\begin{aligned} \gamma_j^{(\mu)}(\mathcal{L},\alpha)[n] &:= \alpha^\top \tilde{g}_j^{(\mu)}[n] && j \in \mathbb{N}\,, \\ \tilde{g}_j^{(\mu)}[n]_i &:= \epsilon_{(i),j}^{(\mu)} g_j^{(\mu-1)}(x_{(i)})[n] && i \in \{1, \dots, |\mathcal{L}|\}\,, \end{aligned}$$

where $((i), x_{(i)})$ is the i^{th} member of the set $\mathcal{L}$. Without loss of generality, in what follows we will take $\hat{C}_w^{(\mu)} = 1$ to lighten notation.

To prove Lemma 15, we need to know the value of σ_*^4 where $\sigma_*^2 = \sigma^2(\mu, \mathcal{L}, \alpha)[*]$ as defined in Lemma 11. Lemma 17 combined with the inductive propagation of convergence in distribution verifies Lemma 11 and thus yields σ_*^4.

Lemma 17 *Consider arbitrary $\mu \in \{2, \dots, D+1\}$. Assume that the countably infinite vector of random variables $\{f_i^{(\mu-1)}(x)[n]\}_{i\in\mathbb{N},x\in\mathcal{X}}$ converges in distribution to a centred Gaussian process with covariance specified by the recursion in Lemma 2 as $n \to \infty$. Then*

$$\sigma^2(\mu,\mathcal{L},\alpha)[*] = \lim_{n\to\infty} \sigma^2(\mu,\mathcal{L},\alpha)[n] = \alpha^T K(\mathcal{L})\alpha\,.$$

Proof Lemma 11 introduces $K(\mathcal{L})$ which is the marginal covariance of the limiting Gaussian process without the bias term (c.f. the recursion in Lemma 2).

We use exchangeability of $\gamma_j^{(\mu)}(\mathcal{L},\alpha)[n]$ over the index j to obtain

$$\sigma^2(\mu,\mathcal{L},\alpha)[n] = \mathbb{E}\left[(\gamma_1^{(\mu)}(\mathcal{L},\alpha)[n])^2\right] = \alpha^\top \mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]\tilde{g}_1^{(\mu)}[n]^\top\right]\alpha\,.$$

Hence the limit of $\sigma^2(\mu,\mathcal{L},\alpha)[n]$ is fully determined by the behaviour of $\tilde{g}_1^{(\mu)}[n]$ as $n \to \infty$.

We can thus focus on individual entries of the expectation on the RHS of the above equation. For entry (i, j) with $i, j \in \{1, \ldots, |\mathcal{L}|\}$, we have

$$\mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]_i \, \tilde{g}_1^{(\mu)}[n]_j\right] = \delta_{(i)=(j)} \mathbb{E}\left[g_1^{(\mu-1)}(x_{(i)})[n] g_1^{(\mu-1)}(x_{(j)})[n]\right].$$

Since $g_1^{(\mu-1)}(x_{(k)})[n] = \phi(f_1^{(\mu-1)}(x_{(k)})[n]), k \in \{i, j\}$ and the collection $\{f_1^{(\mu-1)}(x)[n]\}_{x \in \mathcal{X}}$ converges in distribution as $n \to \infty$ by assumption, we can use the continuity of ϕ and the continuous mapping theorem to deduce that the post-nonlinearities are converging in distribution. Because the function $h(x_1, x_2) = x_1 x_2$ is continuous, we can apply the continuous mapping theorem again to deduce that the two-way products of post-nonlinearities are converging in distribution to the limit specified by the pushforward of the limiting multivariate normal distribution.

Theorem 3.5 in (Billingsley, 1999) tells us that the expectation

$$\lim_{n \to \infty} \mathbb{E}\left[g_1^{(\mu-1)}(x_{(i)})[n] g_1^{(\mu-1)}(x_{(j)})[n]\right] = \mathbb{E}\left[g_1^{(\mu-1)}(x_{(i)})[*] g_1^{(\mu-1)}(x_{(j)})[*]\right],$$

if the family of random variables indexed by n is uniformly integrable. Uniform integrability is a corollary of Lemma 21. Inspection of the recursion in Lemma 11 finishes the proof. ■

Proof [Proof of Lemma 15] Substituting for $X_{n,1}$ and $X_{n,2}$, we have

$$\mathbb{E}\left[\left|\gamma_1^{(\mu)}(\mathcal{L}, \alpha)[n] \gamma_2^{(\mu)}(\mathcal{L}, \alpha)[n]\right|^2\right] = \alpha^\top \mathbb{E}\left[\tilde{g}_1^{(\mu)}[n] \tilde{g}_1^{(\mu)}[n]^\top \alpha \alpha^\top \tilde{g}_2^{(\mu)}[n] \tilde{g}_2^{(\mu)}[n]^\top\right] \alpha. \tag{46}$$

The expectation on the RHS can be rewritten as

$$\mathbb{E}\left[\tilde{g}_1^{(\mu)}[n] \tilde{g}_1^{(\mu)}[n]^\top \alpha \alpha^\top \tilde{g}_2^{(\mu)}[n] \tilde{g}_2^{(\mu)}[n]^\top\right] = \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \alpha_i \alpha_j \mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]_i \, \tilde{g}_2^{(\mu)}[n]_j \, \tilde{g}_1^{(\mu)}[n] \, \tilde{g}_2^{(\mu)}[n]^\top\right],$$

Hence the limit of the LHS of Equation (46) is fully determined by the behaviour of $\tilde{g}_t^{(\mu)}[n], t = 1, 2$, as $n \to \infty$. We can thus focus on individual entries of the expectation on the RHS of the above equation. For entry (k, l) with $k, l \in \{1, \ldots, |\mathcal{L}|\}$, we have

$$\begin{aligned} &\mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]_i \, \tilde{g}_2^{(\mu)}[n]_j \, \tilde{g}_1^{(\mu)}[n]_k \, \tilde{g}_2^{(\mu)}[n]_l\right] \\ &= \delta_{(i)=(k)} \delta_{(j)=(l)} \mathbb{E}\left[g_1^{(\mu-1)}(x_{(i)})[n] g_2^{(\mu-1)}(x_{(j)})[n] g_1^{(\mu-1)}(x_{(k)})[n] g_2^{(\mu-1)}(x_{(l)})[n]\right]. \end{aligned} \tag{47}$$

In analogy with the proof of Lemma 17, we can establish convergence in distribution of the four-way product inside the RHS expectation, and combine Lemma 21 with Theorem 3.5 in (Billingsley, 1999) to get convergence in distribution as $n \to \infty$. Hence $\mathbb{E}\left[|\gamma_1^{(\mu)}(\mathcal{L}, \alpha)[n] \gamma_2^{(\mu)}(\mathcal{L}, \alpha)[n]|^2\right]$ converges to a limit which is a function of terms

$$\begin{aligned} &\mathbb{E}\left[g_1^{(\mu-1)}(x_{(i)})[*] g_2^{(\mu-1)}(x_{(j)})[*] g_1^{(\mu-1)}(x_{(k)})[*] g_2^{(\mu-1)}(x_{(l)})[*]\right] \\ &= \mathbb{E}\left[g_1^{(\mu-1)}(x_{(i)})[*] g_1^{(\mu-1)}(x_{(k)})[*]\right] \mathbb{E}\left[g_2^{(\mu-1)}(x_{(j)})[*] g_2^{(\mu-1)}(x_{(l)})[*]\right]. \end{aligned}$$

Substituting back and inspecting the recursion in Lemma 11 concludes the proof. ■

Proof [Proof of Lemma 16] By Hölder's inequality, it is sufficient to exhibit a bound on the sequence of fourth moments, which are algebraically convenient to work with. Hence it is sufficient to prove that

$$\mathbb{E}\left[\left|\gamma_1^{(\mu)}(\mathcal{L},\alpha)[n]\right|^4\right] = \alpha^\top \mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]\tilde{g}_1^{(\mu)}[n]^\top \alpha\alpha^\top \tilde{g}_1^{(\mu)}[n]\tilde{g}_1^{(\mu)}[n]^\top\right]\alpha\,,$$

is bounded by a constant independent of n. A way to obtain such constant is to bound each term inside the RHS expectation. We rearrange

$$\mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]\tilde{g}_1^{(\mu)}[n]^\top \alpha\alpha^\top \tilde{g}_1^{(\mu)}[n]\tilde{g}_1^{(\mu)}[n]^\top\right] = \sum_{i=1}^{|\mathcal{L}|}\sum_{j=1}^{|\mathcal{L}|} \alpha_i\alpha_j \mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]_i\, \tilde{g}_1^{(\mu)}[n]_j\, \tilde{g}_1^{(\mu)}[n]\, \tilde{g}_1^{(\mu)}[n]^\top\right].$$

Hence it is sufficient to ensure that the expectations

$$\mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]_i\, \tilde{g}_1^{(\mu)}[n]_j\, \tilde{g}_1^{(\mu)}[n]_k\, \tilde{g}_1^{(\mu)}[n]_l\right],$$

are bounded by a constant independent of n for any combination of $i, j, k, l \in \{1, \ldots, |\mathcal{L}|\}$. Substituting back for $\tilde{g}_1^{(\mu)}[n]$

$$\begin{aligned}&\mathbb{E}\left[\tilde{g}_1^{(\mu)}[n]_i\, \tilde{g}_1^{(\mu)}[n]_j\, \tilde{g}_1^{(\mu)}[n]_k\, \tilde{g}_1^{(\mu)}[n]_l\right]\\ &\quad = \delta_{(i)=(j)=(k)=(l)}\mathbb{E}\left[g_1^{(\mu-1)}(x_{(i)})[n]g_1^{(\mu-1)}(x_{(j)})[n]g_1^{(\mu-1)}(x_{(k)})[n]g_1^{(\mu-1)}(x_{(l)})[n]\right],\end{aligned}$$

We thus only need to bound the second factor on the RHS. After upper bounding by the absolute value, we can use Lemma 18 to conclude it is sufficient to bound the fourth moment of $g_1^{(\mu-1)}(x_{(t)})[n]$ for $t = 1, \ldots, |\mathcal{L}|$.[2] Using the linear envelope condition, we see

$$\mathbb{E}\left[\left|g_1^{(\mu-1)}(x_{(t)})[n]\right|^4\right] \le 2^{4-1}\mathbb{E}\left[c^4 + m^4\left|f_1^{(\mu-1)}(x_{(t)})[n]\right|^4\right].$$

By Lemma 20 and a simple application of Hölder's inequality, we know that the fourth moment above is bounded by a constant independent of n. Because we are only considering a finite set of inputs, we can bound the fourth moments for all $f_1^{(\mu-1)}(x_{(t)})[n]$ by a shared constant, namely the maximum over $t \in \{1, \ldots, |\mathcal{L}|\}$. This constant is independent of n which concludes the proof. ■

B.2 Proofs of auxiliary results

The following results are useful in proving Lemmas 15 and 16.

[2] This result can be obtained by allowing only $p_i = 0, 1$ in Lemma 18.

Lemma 18 *Suppose X_1, X_2, X_3, and X_4 are random variables on $\mathbb{R}$ with the usual Borel σ-algebra. Assume that $\mathbb{E}\left[|X_i|^8\right] < \infty$ for all $i = 1, 2, 3, 4$. Then for any choice of $p_i = 0, 1, 2$ (where $i = 1, 2, 3, 4$), the expectations $\mathbb{E}[\prod_{i=1}^4 |X_i|^{p_i}]$ are uniformly bounded by a polynomial in the 8^{th} moments $\mathbb{E}\left[|X_i|^8\right] < \infty$ for $i = 1, \ldots, 4$.*

Proof Throughout this proof, we will be using the following inequality

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|]\mathbb{E}[|Y|] + \{\mathrm{Var}(|X|)\mathrm{Var}(|Y|)\}^{1/2} \ ,$$

which can be derived from the boundedness of Pearson correlation coefficient.

Using the above inequality, we have

$$\mathbb{E}\left[\prod_{i=1}^4 |X_i|^{p_i}\right] \leq \mathbb{E}[|X_1|^{p_1}|X_2|^{p_2}]\mathbb{E}[|X_3|^{p_3}|X_4|^{p_4}] + \{\mathrm{Var}[|X_1|^{p_1}|X_2|^{p_2}]\mathrm{Var}[|X_3|^{p_3}|X_4|^{p_4}]\}^{1/2} \ .$$

The expectations in the first term on the RHS can then be again upper bounded, for example

$$\mathbb{E}[|X_1|^{p_1}|X_2|^{p_2}] \leq \mathbb{E}[|X_1|^{p_1}]\mathbb{E}[|X_2|^{p_2}] + \{\mathrm{Var}[|X_1|^{p_1}]\mathrm{Var}[|X_2|^{p_2}]\}^{1/2} \ ,$$

which is bounded if $\mathbb{E}\left[|X_i|^4\right] < \infty$ for $i = 1, 2$. Similarly for $\mathbb{E}[|X_3|^{p_3}|X_4|^{p_4}]$.

The second term of the first upper bound can be upper bounded in similar way

$$\mathrm{Var}[|X_1|^{p_1}|X_2|^{p_2}] \leq \mathbb{E}\left[|X_1|^{2p_1}|X_2|^{2p_2}\right] ,$$

where $\mathbb{E}\left[|X_1|^{2p_1}|X_2|^{2p_2}\right]$ can again be upper bounded by argument similar to the above, yielding an upper bound that may be expressed as a fixed polynomial in $\mathbb{E}\left[|X_i|^8\right]$ for $i = 1, 2, 3, 4$, as $p_i \leq 2$ and the lower order absolute moments may be bounded by exponents of the higher ones via Hölder's inequality. ■

Lemma 19 *Assume $w_1, \ldots, w_k \in \mathbb{R}$ are arbitrary constants, and ε_i, $i = 1, \ldots, k$, are i.i.d. standard normal variables. Define the vector $w = (w_i)_{i=1}^k$. Then for $p \geq 0$*

$$\mathbb{E}\left[\left|\sum_{i=1}^k w_i \varepsilon_i\right|^p\right] = \|w\|_2^p \frac{2^{\frac{p}{2}}\Gamma(\frac{p+1}{2})}{\Gamma(\frac{1}{2})} \, .$$

Proof Use the linearity of the dot product and Gaussianity of ε_i's to obtain

$$\mathbb{E}\left[\left|\sum_{i=1}^k w_i \varepsilon_i\right|^p\right] = \mathbb{E}[|\|w\|_2 \tilde{\varepsilon}|^p] = \|w\|_2^p \, \mathbb{E}[|\tilde{\varepsilon}|^p] \, ,$$

where $\tilde{\varepsilon}$ is a standard normal random variable. The result is then obtained by realising that powers of standard normal are distributed according to Generalised Gamma variable for which the expectation is known. ■

Lemma 20 *For any given $\mu \in \{1, 2, \ldots, D+1\}$, and input $x \in \mathcal{X}$, the eighth moments of the random variables $f_i^{(\mu)}(x)[n]$ are bounded by a finite constant independent of $n \in \mathbb{N}$ and $i \in \mathbb{N}$.*

Proof The statement is trivially true for $\mu = 1$: the law of $f_i^{(1)}(x)[n]$ for any $(i, x) \in \mathbb{N} \times \mathcal{X}$ is a normal distribution by the Gaussianity of the weights and biases, $f_i^{(1)}(x)[n]$ is equal in law to $f_i^{(1)}(x)[m]$, $\forall (m, n) \in \mathbb{N} \times \mathbb{N}$, implying that the moments are bounded by a constant independent of n, and independence of the constant of index i is obtained by exchangeability.

We can thus proceed by induction. We assume that the condition holds for all $\mu = 1, 2, \ldots, t-1$ (for some $t \in \{2, \ldots, D+1\}$), and prove that it must then also necessarily hold for $\mu = t$. First we obtain the following upper bound

$$\mathbb{E}\left[|f_i^{(t)}(x)[n]|^8\right] \leq 2^{8-1}\mathbb{E}\left[|b_i^{(t)}|^8 + \left|\sum_{j=1}^{h_{t-1}(n)} w_{i,j}^{(t)} g_j^{(t-1)}(x)[n]\right|^8\right],$$

noting that the expectation of the first term is uniformly bounded in i by properties of the Gaussian distribution.

Hence we focus on the second term. We use Lemma 19 to obtain

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{j=1}^{h_{t-1}(n)} w_{i,j}^{(t)} g_j^{(t-1)}(x)[n]\right|^8\right] &= \mathbb{E}\left[\mathbb{E}\left[\left|\sum_{j=1}^{h_{t-1}(n)} w_{i,j}^{(t)} g_j^{(t-1)}(x)[n]\right|^8 \;\middle|\; g_{1:h_{t-1}(n)}^{(t-1)}(x)[n]\right]\right] \\ &= \frac{2^4\Gamma(4+1/2)}{\Gamma(1/2)}\mathbb{E}\left[\left|\frac{\hat{C}_w^{(t-1)}}{h_{t-1}(n)}||g_{1:h_{t-1}(n)}^{(t-1)}(x)[n]||_2^2\right|^4\right], \qquad (48)\end{aligned}$$

where $g_{1:h_{t-1}(n)}^{(t-1)}(x)[n]$ is the set of post-nonlinearities corresponding to $j = 1, 2, \ldots, h_{t-1}(n)$. Observe that

$$\begin{aligned}\frac{1}{h_{t-1}(n)}||g_{1:h_{t-1}(n)}^{(t-1)}(x)[n]||_2^2 &= \frac{1}{h_{t-1}(n)}\sum_{j=1}^{h_{t-1}(n)} (g_j^{(t-1)}(x)[n])^2 \\ &\leq \frac{1}{h_{t-1}(n)}\sum_{j=1}^{h_{t-1}(n)} (c + m|f_j^{(t-1)}(x)[n]|)^2,\end{aligned}$$

by the linear envelope property. Suppressing a multiplicative constant independent of x and n and substituting this bound back into Expression (48) yields

$$\begin{aligned}&\mathbb{E}\left[\left|\sum_{j=1}^{h_{t-1}(n)} w_{i,j}^{(t)} g_j^{(t-1)}(x)[n]\right|^8\right] \\ \leq&\frac{1}{h_{t-1}(n)^4}\mathbb{E}\left[\left|\sum_{j=1}^{h_{t-1}(n)} c^2 + 2cm|f_j^{(t-1)}(x)[n]| + m^2|f_j^{(t-1)}(x)[n]|^2\right|^4\right].\end{aligned}$$

The above can be simply multiplied out, yielding a weighted sum of expectations of the form

$$\mathbb{E}\left[|f_k^{(t-1)}(x)[n]|^{p_1}\,|f_l^{(t-1)}(x)[n]|^{p_2}\,|f_r^{(t-1)}(x)[n]|^{p_3}\,|f_q^{(t-1)}(x)[n]|^{p_4}\right],$$

with $p_i \in \{0,1,2\}$ for $i = 1,2,3,4$, and $k,l,r,q \in \{1,\dots,h_{t-1}(n)\}$, and where the weights of these terms are independent of n. Using Lemma 18, each of these terms is bounded if the eighth moments of $f_k^{(t-1)}(x)[n]$ are bounded which is our inductive hypothesis. The number of terms in the expanded sum is upper bounded by $(3h_{t-1}(n))^4$ and thus we can use the same constant for any $n \in \mathbb{N}$ due to the $1/h_\mu(n)^4$ scaling. Noticing that $f_j^{(t-1)}(x)[n]$ are exchangeable over the index j for any fixed x and n concludes the proof. ■

Lemma 21 *Consider a collection of random variables $g_i^{(\mu)}(x_1)[n]$, $g_j^{(\mu)}(x_2)[n]$, $g_k^{(\mu)}(x_3)[n]$, and $g_l^{(\mu)}(x_4)[n]$ with any $i,j,k,l \in \mathbb{N}$, $x_1,x_2,x_3,x_4 \in \mathcal{X}$, neither necessarily distinct. Then the family of random variables*

$$g_i^{(\mu)}(x_1)[n]g_j^{(\mu)}(x_2)[n]g_k^{(\mu)}(x_3)[n]g_l^{(\mu)}(x_4)[n]\,, \tag{49}$$

indexed by n is uniformly integrable for any $\mu = 1,2,\dots,D+1$.

Proof A simple way to prove uniform integrability of an arbitrary family of real-valued random variables $\{X_n\}_{n\in\mathbb{N}}$ is to show $\sup_n \mathbb{E}\left[|X_n|^{1+\epsilon}\right] < \infty$ for some $\epsilon > 0$. We use $\epsilon = 1$, i.e. the second moment of the four-way product from Equation (49) will be bounded by a constant independent of n.

Write

$$\mathbb{E}\left[\left|g_i^{(\mu)}(x_1)[n]g_j^{(\mu)}(x_2)[n]g_k^{(\mu)}(x_3)[n]g_l^{(\mu)}(x_4)[n]\right|^2\right],$$

and recall that by Lemma 18, we only need to bound the eighth moments of $g_i^{(\mu)}(x)[n]$ by a constant independent of n for i,j,k and l, and $\{x_t\}_{t=1}^4$. Using the linear envelope property

$$\mathbb{E}\left[\left|g_i^{(\mu)}(x)[n]\right|^8\right] \le 2^{8-1}\mathbb{E}\left[c^8 + m^8\left|f_i^{(\mu)}(x)[n]\right|^8\right].$$

The result in Lemma 20 gives us a constant upper bounding the left hand side which depends on x but not n. Because we are considering only a fixed finite set of inputs, we can take the maximum of these constants to conclude the proof. ■