Gaussian Process Behaviour in Wide Deep Neural Networks

Alexander G. de G. Matthews

DeepMind

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani.
Gaussian Process Behaviour in Wide Deep Neural Networks
In 6th International Conference on Learning Representations (ICLR), Vancouver, Canada, April 2018.

Extended version on arXiv. Includes:

1) More general theory and better proof method.
2) More extensive experiments.

Code to reproduce all experiments is at: https://github.com/widedeepnetworks/widedeepnetworks

Authors

Alex Matthews

Mark Rowland

Jiri Hron

Richard Turner

Zoubin Ghahramani

UNIVERSITY OF CAMBRIDGE

DeepMind

UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF CAMBRIDGE

Google AI

Microsoft

Uber

UNIVERSITY OF
CAMBRIDGE

Potential of Bayesian neural networks

Data efficiency is a serious problem for instance in deep RL.

Generalization in deep learning is (still) poorly understood.

Can reveal and critique the true model assumptions of deep learning?

Priors on weights are difficult to interpret.

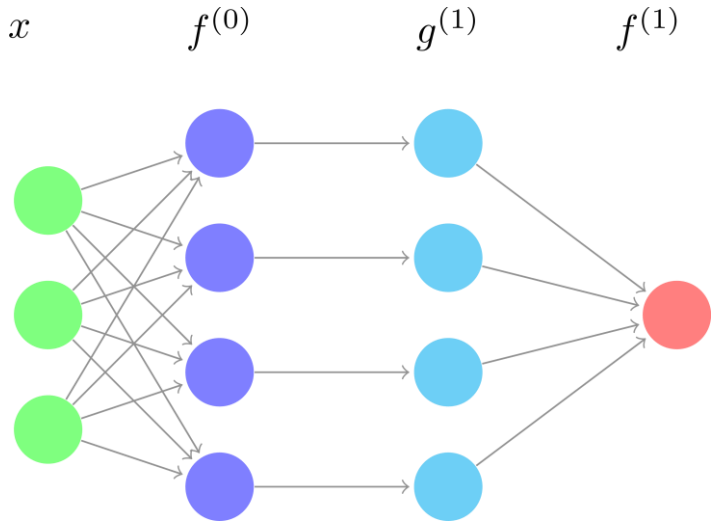
If we do not understand the prior then why do we expect good performance?

Possible e.g that we are doing good inference with a terrible prior.

Increasing width, single hidden layer (Neal 1994)

$$f^{(0)} = W^{(0)} x$$
$$g^{(1)} = \phi(f^{(0)})$$
$$f^{(1)} = W^{(1)} g^{(1)}$$

$$x \in \mathbb{R}^{D \times N}$$
$$W^{(0)} \in \mathbb{R}^{K \times D}$$
$$W^{(1)} \in \mathbb{R}^{1 \times K}$$
$$f^{(1)} \in \mathbb{R}^{1 \times N}$$



Carefully scaled prior

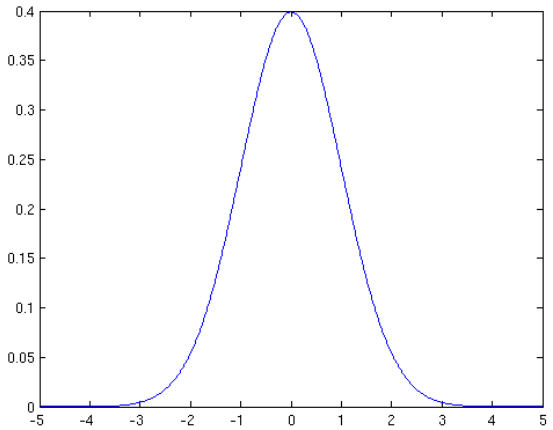
$$W^{(0)}_{i,j} \sim \mathcal{N}(0, \frac{1}{D}) \text{ Indep.}$$
$$W^{(1)}_{i,j} \sim \mathcal{N}(0, \frac{1}{K}) \text{ Indep.}$$

$$[f^{(1)}]^T \xrightarrow[K \to \infty]{\mathcal{D}} \mathcal{N}(0, C)$$
$$C_{i,j} = \mathbb{E}_{w \sim \mathcal{N}(0, \frac{\mathbb{I}}{D})}[\phi(x_{*,i} w)\phi(x_{*,j} w)]$$

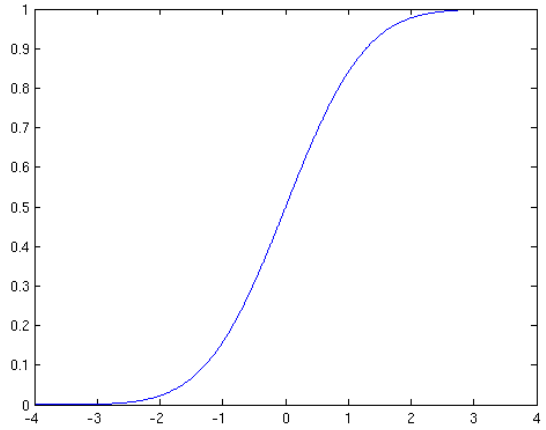
Proof:
Standard Multivariate CLT

The Central Limit Theorem (CLT)

1D Convergence in distribution ↔ Convergence of CDF at all continuity points



$$\int_{-\infty}^{u} p(u')du'$$



Consider a sequence of i.i.d random variables $(u_1, u_2, .., u_n)$. With mean 0 and finite variance σ^2.

Define the standardized sum: $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i$.

Then: $S_n \xrightarrow{D} N(0, \sigma^2)$

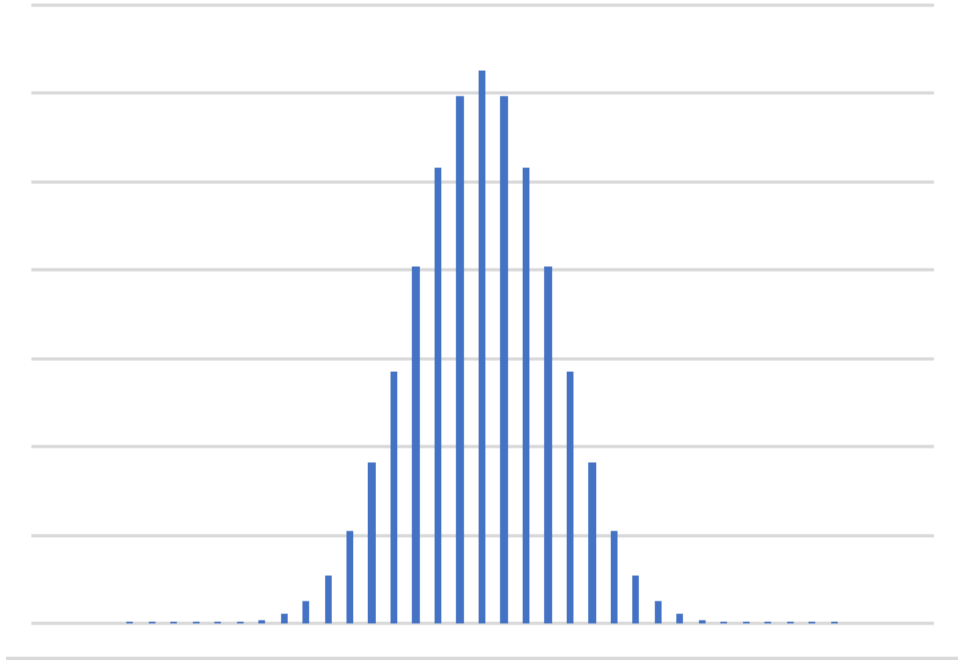
Subtleties of convergence in distribution: a simple example

Consider an i.i.d sequence of Rademacher RVs.

$p(x_i = -1) = p(x_i = 1) = 1/2$

Define $S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_n$

Convergence to $N(0, 1)$ in distribution as $n \to \infty$



Consider the set $A = \{\frac{a}{\sqrt{b}} : a \in \mathbb{Z},\ b \in \mathbb{N}\}$

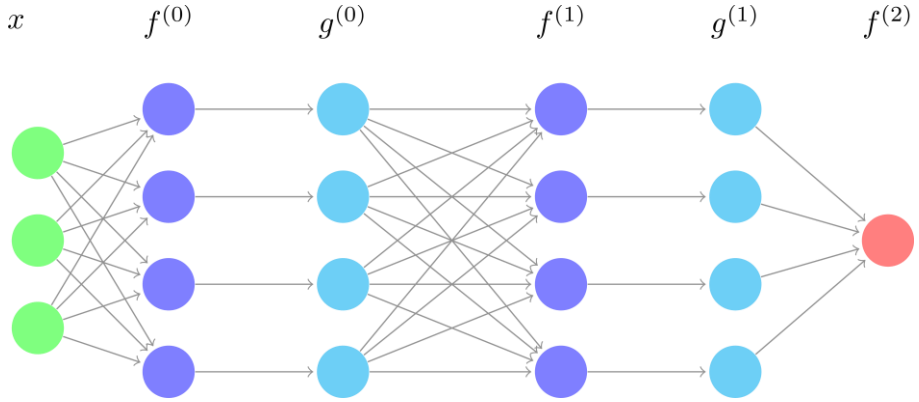
Then $Pr(S_n \in A) = 1$ for all n whereas A has probability zero under $N(0, 1)$.

Question: What does it mean for a stochastic process to converge in distribution?

One answer: All finite dimensional distributions converge in distribution.

Increasing width, multiple hidden layers

$$
\begin{aligned}
f^{(0)} &= W^{(0)} x \\
g^{(1)} &= \phi(f^{(0)}) \\
f^{(1)} &= W^{(1)} g^{(1)} \\
g^{(2)} &= \phi(f^{(1)}) \\
f^{(2)} &= W^{(2)} g^{(2)}
\end{aligned}
\qquad
\begin{aligned}
x &\in \mathbb{R}^{D \times N} \\
W^{(0)} &\in \mathbb{R}^{K^{(1)} \times D} \\
W^{(1)} &\in \mathbb{R}^{K^{(2)} \times K^{(1)}} \\
W^{(2)} &\in \mathbb{R}^{1 \times K^{(2)}} \\
f^{(2)} &\in \mathbb{R}^{1 \times N}
\end{aligned}
$$



Carefully scaled prior

$$
\begin{aligned}
W^{(0)}_{i,j} &\sim \mathcal{N}(0, \frac{1}{D}) \text{ Indep.} \\
W^{(1)}_{i,j} &\sim \mathcal{N}(0, \frac{1}{K^{(1)}}) \text{ Indep.} \\
W^{(2)}_{i,j} &\sim \mathcal{N}(0, \frac{1}{K^{(2)}}) \text{ Indep.}
\end{aligned}
$$

$$
[f^{(2)}]^T \xrightarrow[K^{(1)}, K^{(2)} \to \infty]{\mathcal{D}} \mathcal{N}(0, C^{(2)})
$$

$$
C^{(1)}_{i,j} = \mathbb{E}_{w \sim \mathcal{N}(0, \frac{\mathbb{I}}{D})}[\phi(x_{*,i} w)\phi(x_{*,j} w)]
$$

$$
C^{(2)}_{i,j} = \mathbb{E}_{\begin{pmatrix}\epsilon_i \\ \epsilon_j\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}C^{(1)}_{i,i} & C^{(1)}_{i,j} \\ C^{(1)}_{j,i} & C^{(1)}_{j,j}\end{pmatrix}\right)}[\phi(\epsilon_i)\phi(\epsilon_j)]
$$

Lee, Bahri, Novak, Schoenholz, Pennington and Sohl-Dickstein
Deep Neural Networks as Gaussian Processes
International Conference on Learning Representations (ICLR), 2018.

Publicly available on the same day.
Accepted at the same conference.

Schoenholz, Gilmer, Ganguli, and Sohl-Dickstein.
Deep Information Propagation.
International Conference on Learning Representations (ICLR), 2017.

Daniely, Frostig, and 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.

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

Duvenaud, Rippel, Adams, and Ghahramani.
Avoiding Pathologies in very Deep Networks.
International Conference on Artificial Intelligence and Statistics (AISTATS), 2014

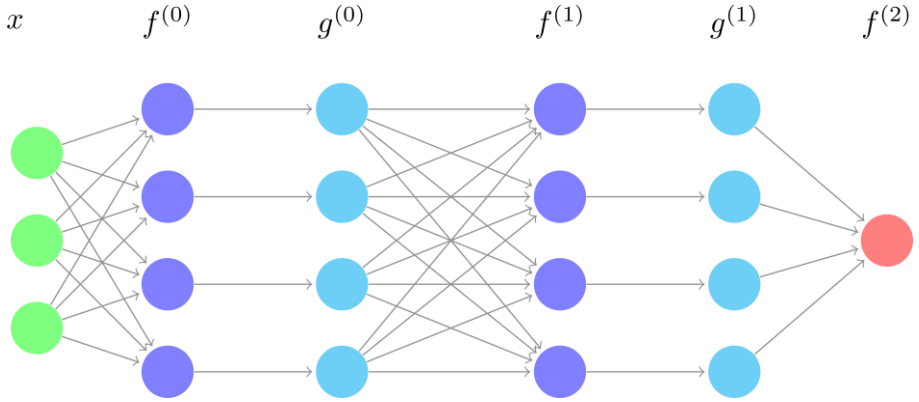
Cho and Saul.
Kernel Methods for Deep Learning.
Advances in Neural Information Processing Systems (NIPS), 2009.

Our contributions

1) Rigorous, general, proof of CLT for networks with more than one hidden layer.

2) Empirical comparison to finite but wide Bayesian neural networks from the literature.

Multiple hidden layers: A first intuition



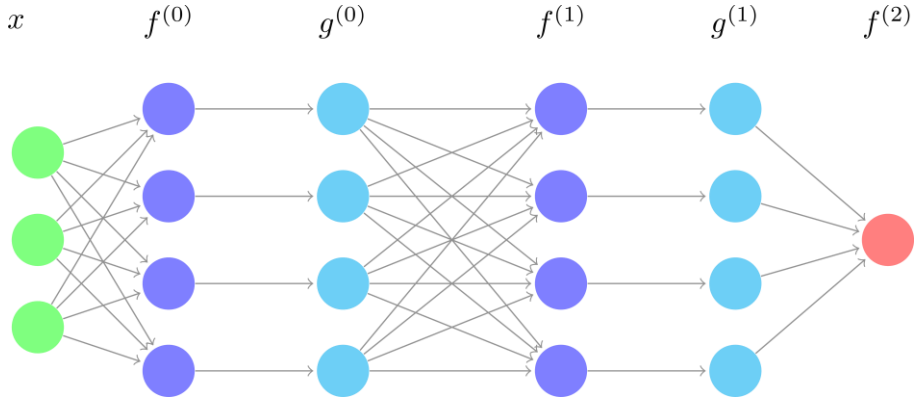
Single input data point: Neal (1995) showed that the $f^{(1)}$ units will converge to independent normal variables. But since $g^{(1)}$ becoming increasingly independent this suggests $f^{(2)}$ will also converge to a normal distribution.

Multiple input data points: There is now a correlated normal vector at each $f^{(1)}$ unit with elements corresponding to the different input points. The different $g^{(1)}$ units will still become increasingly independent. This suggests $f^{(2)}$ converges to a correlated normal vector.

Problem with argument: The $f^{(1)}$ units are only independent asymptotically. Convergence may depend on the rate which this limit is achieved.

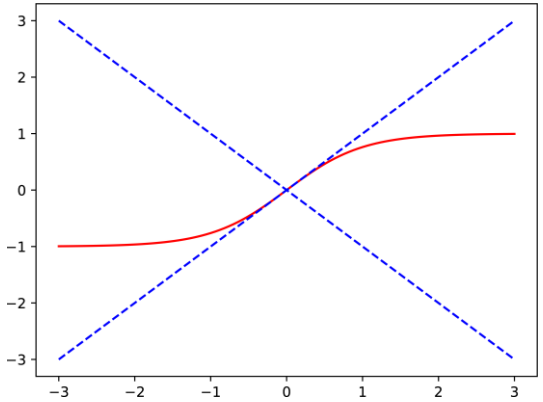
Careful treatment: Preliminaries

$$x \in \mathbb{R}^{D \times N}$$
$$W^{(0)} \in \mathbb{R}^{K^{(1)} \times D}$$
$$W^{(1)} \in \mathbb{R}^{K^{(2)} \times K^{(1)}}$$
$$W^{(2)} \in \mathbb{R}^{1 \times K^{(2)}}$$
$$f^{(2)} \in \mathbb{R}^{1 \times N}$$



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

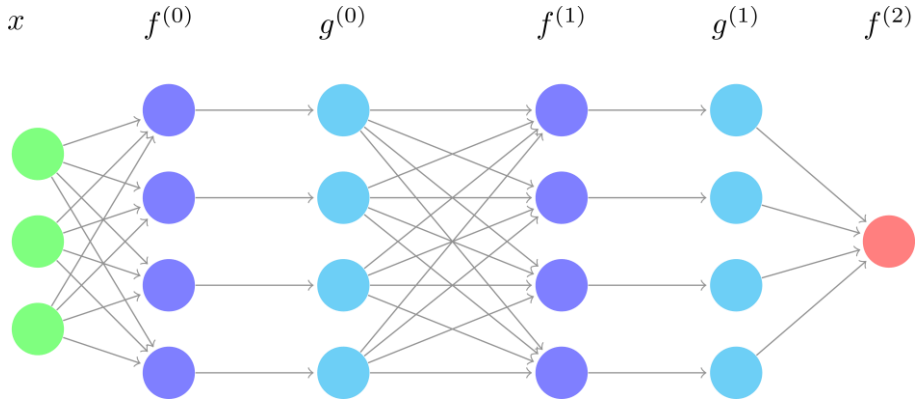
$$|\phi(u)| \leq c + m|u| \quad \forall u \in \mathbb{R}\,.$$



For a given network sequence index $\kappa \in \mathbb{N}$, a *width function* $h_d : \mathbb{N} \mapsto \mathbb{N}$ at depth d specifies the number of hidden units $K^{(d)}$ at depth d.

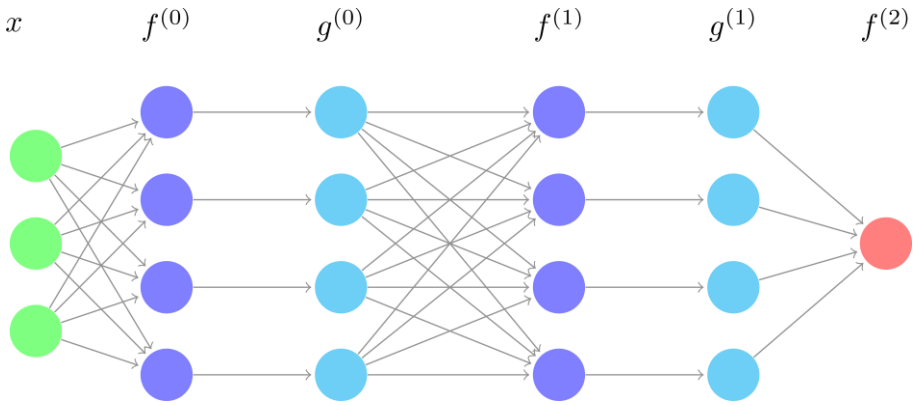
Careful treatment

$$
\begin{aligned}
f^{(0)} &= W^{(0)}x \\
g^{(1)} &= \phi(f^{(0)}) \\
f^{(1)} &= W^{(1)}g^{(1)} \\
g^{(2)} &= \phi(f^{(1)}) \\
f^{(2)} &= W^{(2)}g^{(2)}
\end{aligned}
\qquad
\begin{aligned}
x &\in \mathbb{R}^{D\times N} \\
W^{(0)} &\in \mathbb{R}^{K^{(1)}\times D} \\
W^{(1)} &\in \mathbb{R}^{K^{(2)}\times K^{(1)}} \\
W^{(2)} &\in \mathbb{R}^{1\times K^{(2)}} \\
f^{(2)} &\in \mathbb{R}^{1\times N}
\end{aligned}
$$



Consider a random deep neural network of the type shown above with a continuous nonlinearity obeying the linear envelope condition. Let the weights distributions be appropriately scaled independent normals. Then for all sets of strictly increasing width functions h_d 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 $\kappa \to \infty$. The Gaussian process has mean function zero and the covariance function is given by the recursion already shown.

Proof sketch



1. Proceed through the network by induction starting closest to data.
2. At each layer, reduce the problem to the convergence of any finite linear projection of data and units.
3. Prove convergence in distribution of projections and certain moment bounds at each layer.

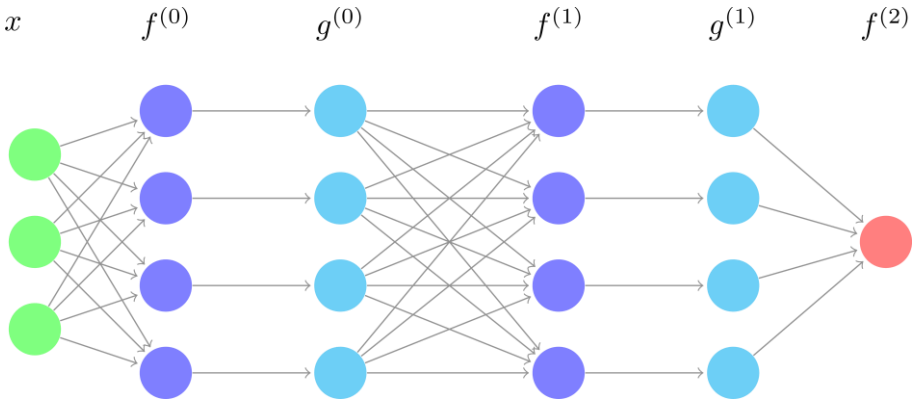
The proof of convergence in distribution for step 3) makes heavy use of the exchangeable central limit theorem of Blum et al 1958, for triangular arrays.

Exchangeability

An infinite sequence of random variables is *exchangeable* if any finite permutation leaves its distribution invariant.

de Finetti's theorem:

An infinite sequence of random variables is exchangeable if and only if it is i.i.d conditional on some random variable.



If we condition on $g^{(0)}$ then the different paths through the second hidden layer to $f^{(2)}$ are independent.

Exchangeable central limit theorem
Blum et al 1958

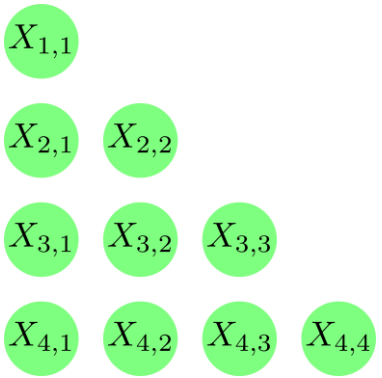
For each positive integer n let $(X_{n,i}; i = 1, 2, ...)$ be an infinitely exchangeable process with mean zero, variance one, and finite absolute third moment. Define

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

Then if the following conditions hold:

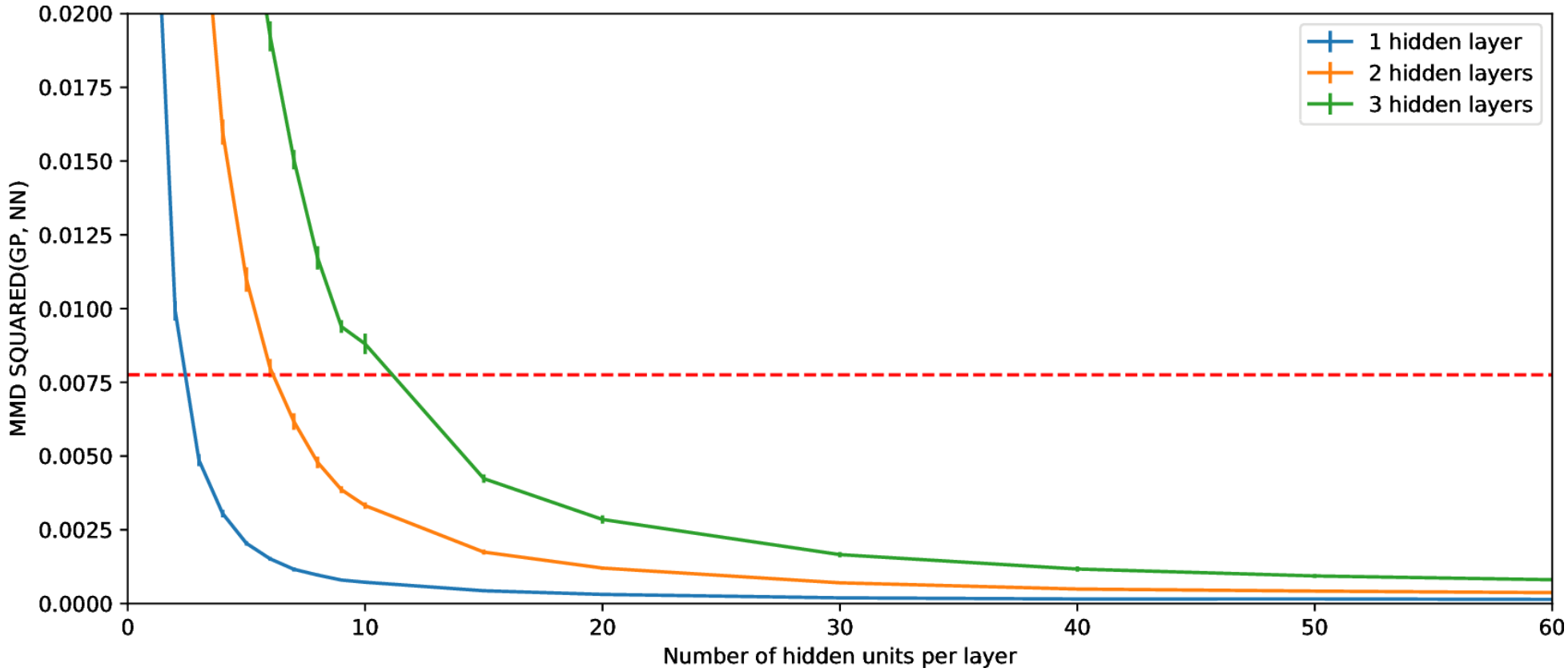
1. $\mathbb{E}_n[X_{n,1} X_{n,2}] = \mathrm{o}(\frac{1}{n})$
2. $\lim_{n\to\infty} \mathbb{E}_n\left[X_{n,1}^2 X_{n,2}^2\right] = 1$
3. $\mathbb{E}_n\left[|X_{n,1}|^3\right] = \mathrm{o}(\sqrt{n})$

Then S_n converges in distribution to a standard normal.



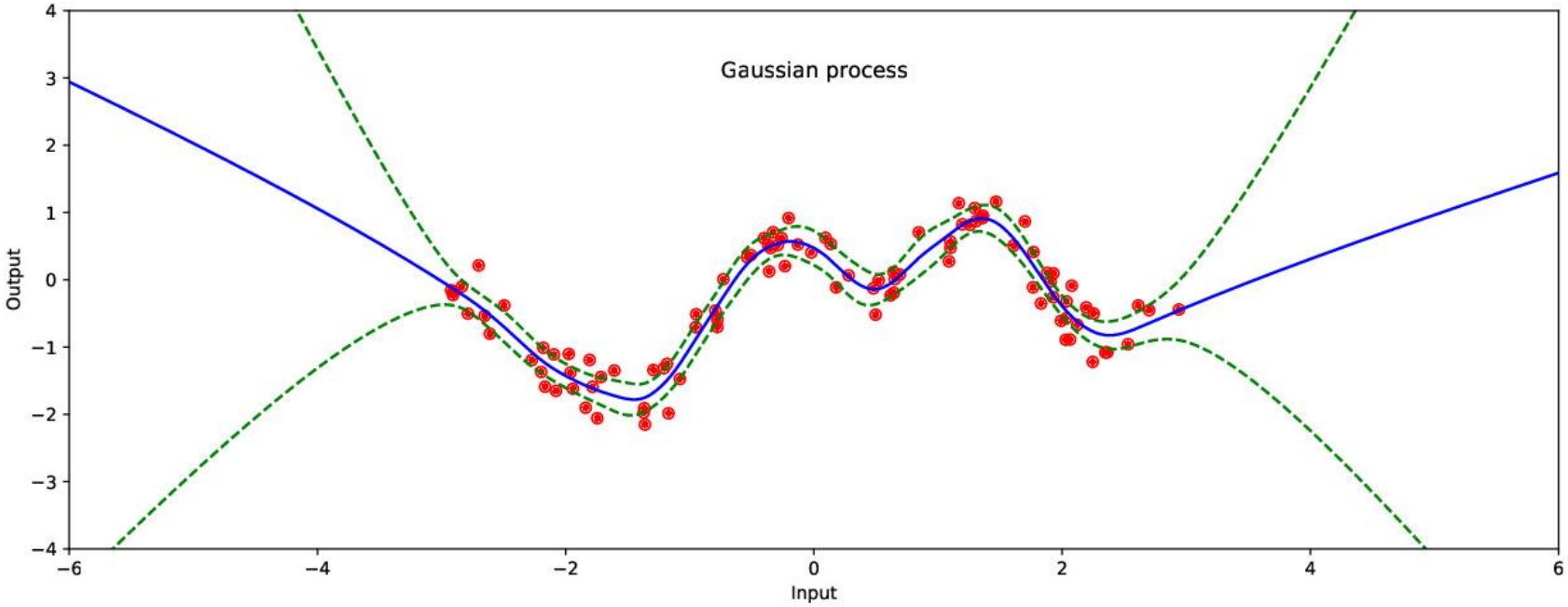
Triangular array:
Allows for the definition of the RVs to change as well as the number.

Empirical rate of convergence

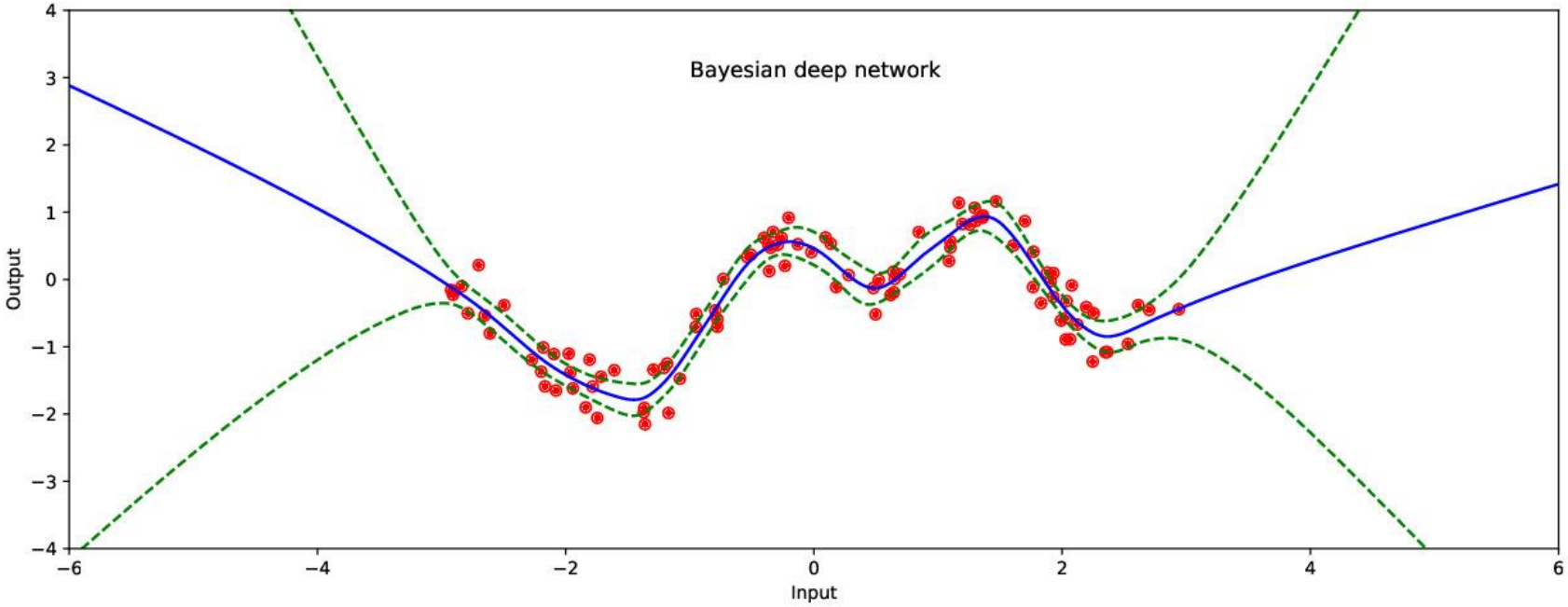


Compare:

1) Exact posterior inference in Gaussian process with the limit kernel (*Fast for this data).*



2) Three hidden layer network with 50 units per hidden layer with gold-standard HMC (*Slow for this data).*



Limitations of kernel methods

A general regression algorithm can be written in the following way:

$$\bar{f}^* = A(x^*, X, y)$$

X training inputs, y training outputs, x^* current test input and $\bar{f}^*$ the prediction.

Kernel methods (including the GP posterior mean) can be written as

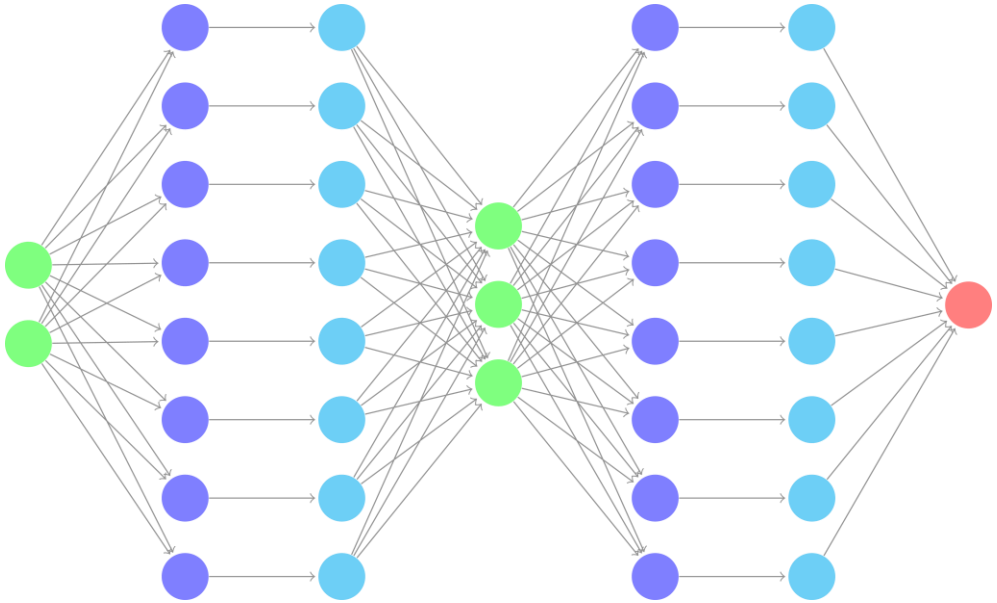
$$\bar{f}^* = \beta(X, x^*)^T y + c(X, x^*)$$

i.e they are an affine transformation of the training outputs.

Deep Gaussian Processes

Can view (some of) these models as taking the limit of some layers but keeping others narrow.

This prevents the onset of the central limit theorem.



Damianou and Lawrence. 2013

A subset of subsequent work

With apologies to many excellent omissions…

Subsequent work: convolutional neural networks and NTK

Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, Jascha Sohl-Dickstein.

ICLR 2019

Deep Convolutional Networks as shallow Gaussian Processes

Adrià Garriga-Alonso, Carl Edward Rasmussen, Laurence Aitchison

ICLR 2019

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot, Franck Gabriel, Clement Hongler

NeurIPS 2018