Avoiding Pathologies in Very Deep Networks

David Duvenaud, Oren Rippel, Ryan Adams, Zoubin Ghahramani

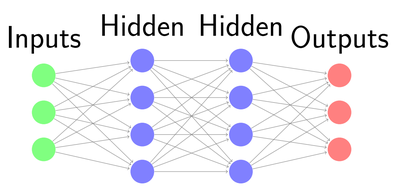
UNIVERSITY OF CAMBRIDGE

HARVARD

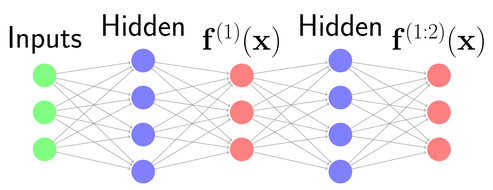
Abstract

- We analyze deep Gaussian processes, a type of infinitely-wide, deep neural net.
- We study distributions of deep GPs and a find a pathology, then show a simple fix.
- We also derive kernels corresponding to infinitely deep nets.

Deep Nets and Deep Gaussian processes



Two-hidden layer MLP



Two-layer GP: $\mathbf{y} = \mathbf{f}^{(2)}(\mathbf{f}^{(1)}(\mathbf{x}))$

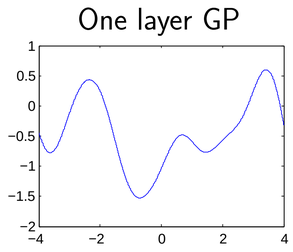
Deep GPs are priors on compositions of functions:

$$\mathbf{f}^{(1:L)}(\mathbf{x}) = \mathbf{f}^{(L)}(\mathbf{f}^{(L-1)}(\dots \mathbf{f}^{(2)}(\mathbf{f}^{(1)}(\mathbf{x}))\dots))$$

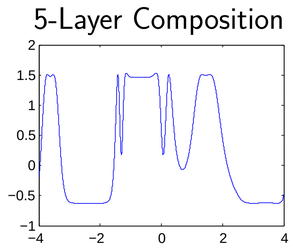
where each $\mathbf{f}^{(\ell)} \overset{\text{ind}}{\sim} \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$. Can be viewed as either

1. MLPs with nonparametric activation functions
2. MLPs with infintely-many parametric hidden nodes

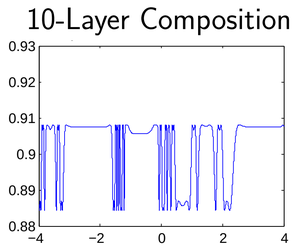
One layer GP



5-Layer Composition



10-Layer Composition

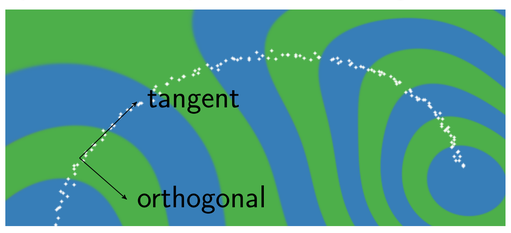


One-dimensional asymptotics

- Can examine deep GPs through distribution of derivatives.
- We use squared-exp kernel $k(x, x') = \sigma_f^2 \exp\left(-(x-x')^2/2\ell^2\right)$
- By the chain rule, the derivative of a deep GP is a product of independent derivatives, each normally distributed.
- The absolute value of derivative is a product of half-normals, each with mean $\sqrt{2\sigma_f^2/\pi\ell^2}$.
- By Central Limit Theorem, size of derivative has log-normal limiting distribution.
- Derivative becomes almost zero everywhere, with large jumps.

Degrees of Freedom of a Neural Network

- Neural net computes a representation $\mathbf{y} = f(\mathbf{x})$ of data $\mathbf{x}$.
- $\mathbf{y}$ needs to capture relevant degrees of freedom of $\mathbf{x}$.



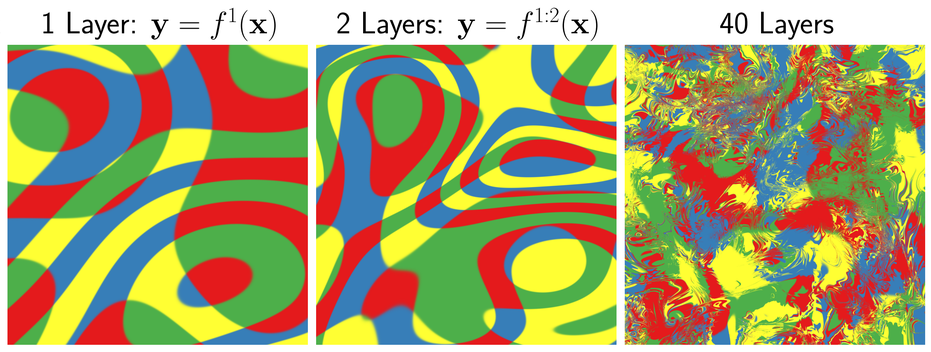
Contours show $f(\mathbf{x})$,
a noise-tolerant representation
of a data manifold (white)

- (Rifai et. al., 2011) argue that a good latent representation is invariant in directions orthogonal to the manifold on which the data lie.
- Conversely, a good latent representation must also change in directions tangent to the data manifold, in order to preserve relevant information.

Random deep nets have few degrees of freedom

We visualize random mappings to show properties of prior on functions:

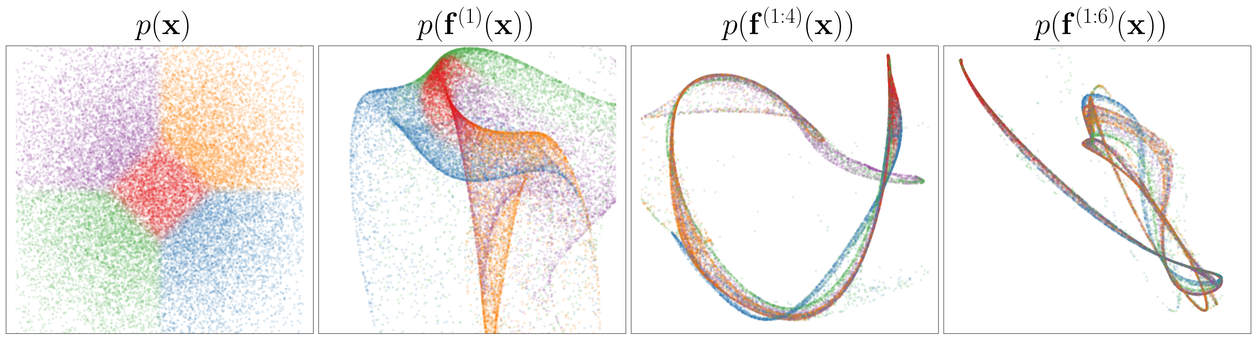
Identity Map: $\mathbf{y} = \mathbf{x}$ | 1 Layer: $\mathbf{y} = f^1(\mathbf{x})$ | 2 Layers: $\mathbf{y} = f^{1:2}(\mathbf{x})$ | 40 Layers



As depth increases, there is usually only one direction we can move $\mathbf{x}$ to change $\mathbf{y}$.

We also visualize a distribution warped by successive functions drawn from a GP prior:

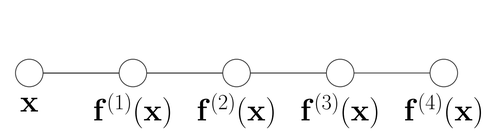
$p(\mathbf{x})$ | $p(\mathbf{f}^{(1)}(\mathbf{x}))$ | $p(\mathbf{f}^{(1:4)}(\mathbf{x}))$ | $p(\mathbf{f}^{(1:6)}(\mathbf{x}))$



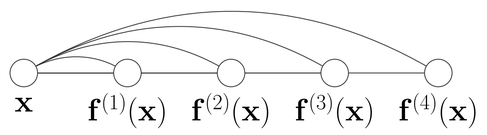
As depth increases, the density concentrates along one-dimensional filaments.

Fixing the pathology

- Following a suggestion from (Neal, 1995), connect input to every layer:



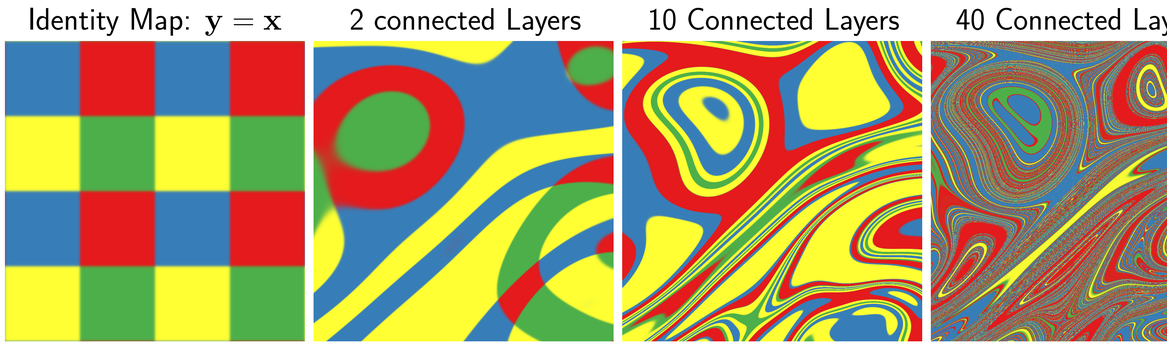
Standard MLP architecture.



Input-connected architecture.

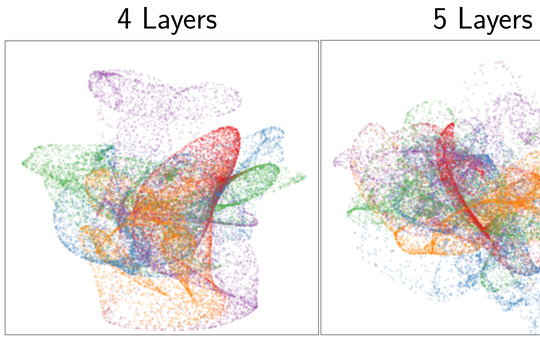
- This fixes the problem: Locally there are usually D degrees of freedom, at any depth:

Identity Map: $\mathbf{y} = \mathbf{x}$ | 2 connected Layers | 10 Connected Layers | 40 Connected Lay



Pathology is also resolved in deep density models: Density does not concentrate along filaments when input connects to all layers.

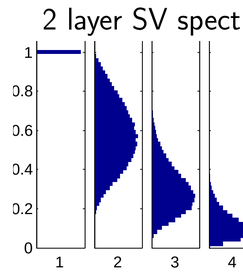
4 Layers | 5 Layers



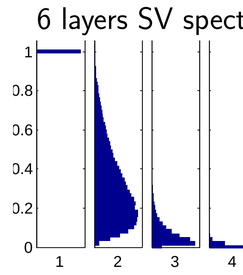
Explaining the Pathology

- Jacobian of a deep GP is a product of independent Gaussian matrices.
- Singular values spectrum of Jacobian quantifies relative size of derivatives.
- As the net gets deeper, distribution of SVs becomes heavy-tailed, and the largest singular value dominates.
- Eventually, there is only one direction we can move $\mathbf{x}$, in order to change $\mathbf{y}$.

2 layer SV spect



6 layers SV spect



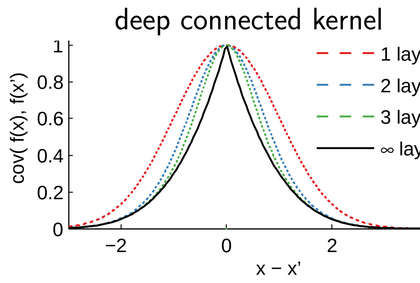
Infinitely Deep Kernels

- Can also analyze fixed deep feature mappings:
- (Cho, 2012) built kernels from multiple layers of feature mappings:
- If $k_1(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\mathsf{T}\Phi(\mathbf{x}')$, we can also build kernel $k_2(\mathbf{x}, \mathbf{x}') = k_2(\Phi(\mathbf{x}), \Phi(\mathbf{x}')) = \Phi(\Phi(\mathbf{x}))^\mathsf{T}\Phi(\Phi(\mathbf{x}'))$.
- For the squared-exp kernel, this composition operation has a closed form:

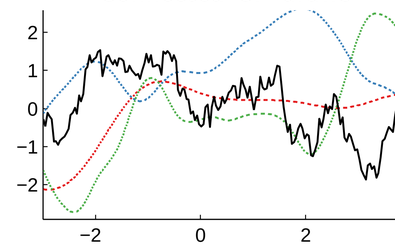
$$k_{n+1}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\left|\left|\begin{bmatrix}\Phi_n(\mathbf{x}) \\ \mathbf{x}\end{bmatrix} - \begin{bmatrix}\Phi_n(\mathbf{x}') \\ \mathbf{x}'\end{bmatrix}\right|\right|_2^2\right)$$
$$= \exp\left(k_n(\mathbf{x}, \mathbf{x}') - 1 - \frac{1}{2}||\mathbf{x} - \mathbf{x}'||_2^2\right)$$

- This kernel satisfies $k_\infty - \log(k_\infty) = 1 + \frac{1}{2}||\mathbf{x} - \mathbf{x}'||_2^2$
- No closed form, but it is continuous and differentiable everywhere except at $\mathbf{x} = \mathbf{x}'$.

deep connected kernel



Connected GP draws



Conclusions

- Random networks capture fewer degrees of freedom as they get deeper
- Connecting the input to each layer resolves this pathology
- Deep Gaussian processes are a data-independent way to characterize neural networks
- Deep One-Dimensional GPs have a log-normal distribution on the magintude of their derivativ
- Can build "deep net" kernels