Avoiding Pathologies in Very Deep Networks

David Duvenaud, Oren Rippel, Ryan Adams, Zoubin Ghahramani

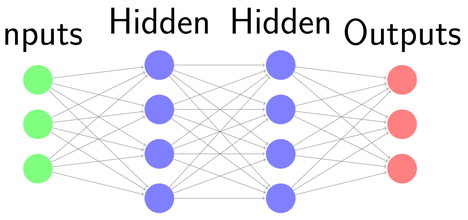
UNIVERSITY OF CAMBRIDGE

HARVARD

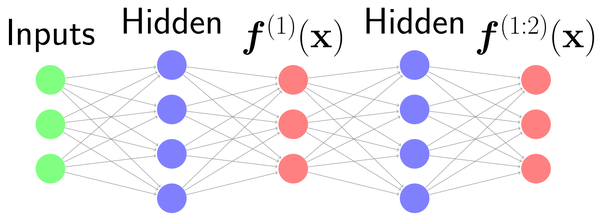
Abstract

- 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 Gaussian Processes as Neural Networks



Two-hidden layer MLP



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

Deep GPs are compositions of functions, each $f^{(\ell)} \overset{\text{ind}}{\sim} \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

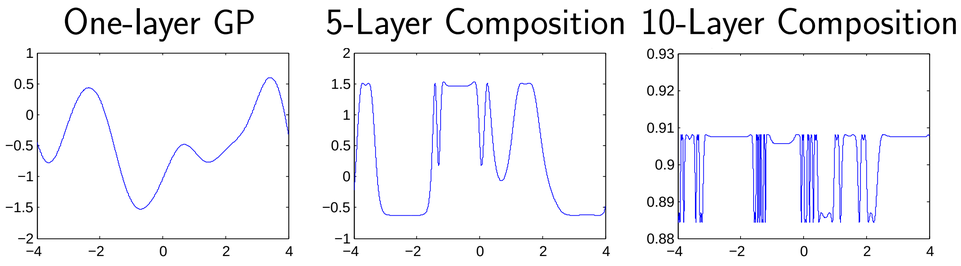
$$\boldsymbol{f}^{(1:L)}(\mathbf{x}) = \boldsymbol{f}^{(L)}(\boldsymbol{f}^{(L-1)}(\dots \boldsymbol{f}^{(2)}(\boldsymbol{f}^{(1)}(\mathbf{x}))\dots))$$

Can be derived as either

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

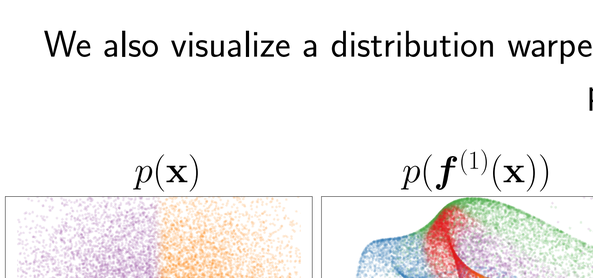
One-dimensional asymptotics

- Size of derivative has log-normal limiting distribution: small almost everywhere, with large jumps.



Degrees of Freedom of a Neural Network

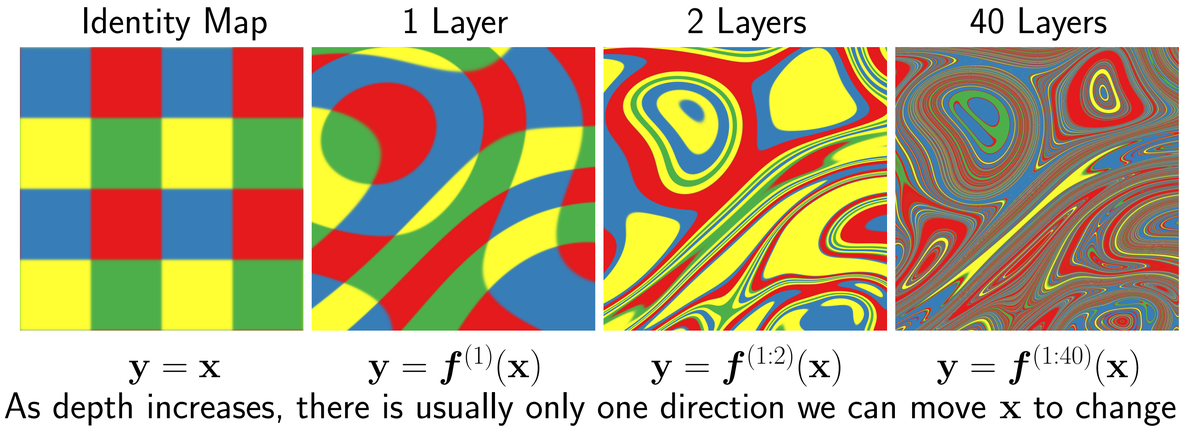
Contours of a good representation



- Representation $\mathbf{y} = \boldsymbol{f}(\mathbf{x})$ should capture relevant degrees of freedom of $\mathbf{x}$.
- Representation must change in directions tangent to the data manifold, to preserve information.

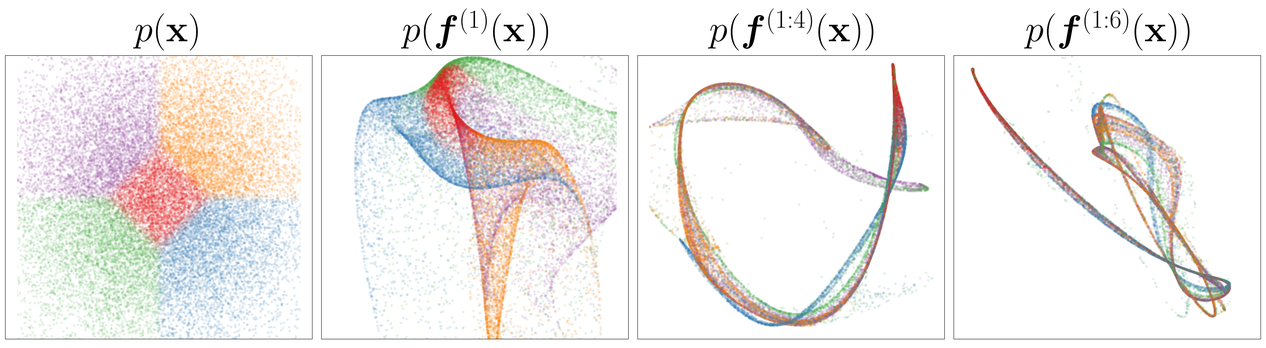
Random deep nets have few degrees of freedom

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



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

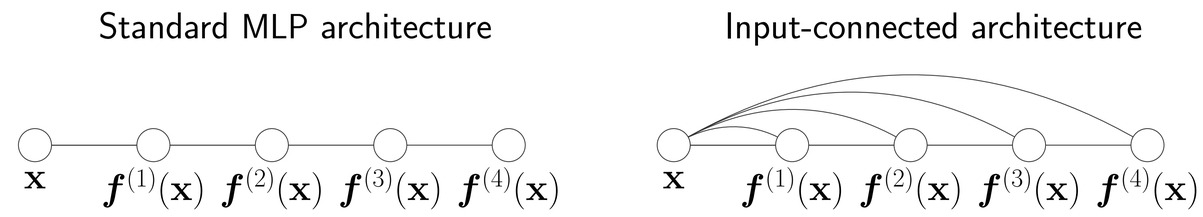
We also visualize a distribution warped by successive functions drawn from a GP prior:



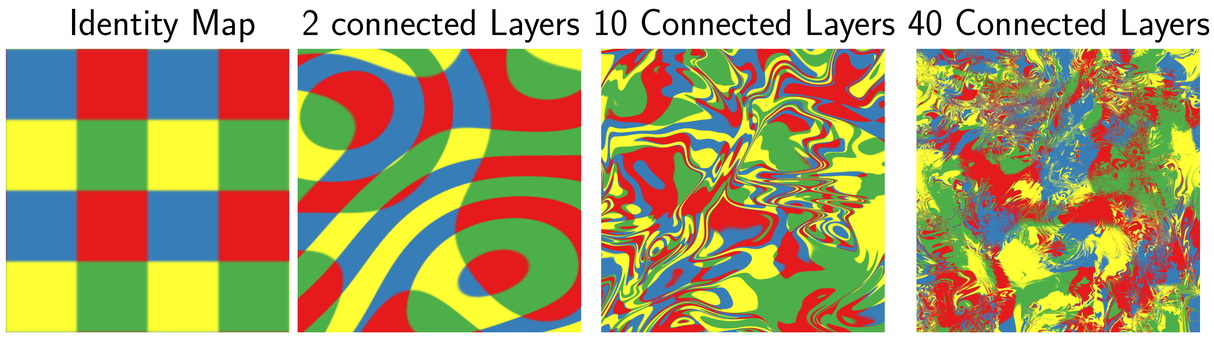
As depth increases, the density concentrates along one-dimensional filaments.

Fixing the pathology

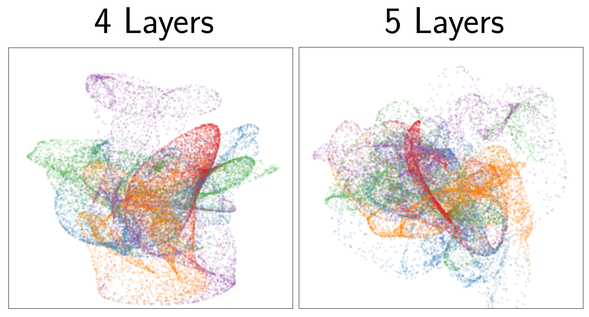
- Following (Neal, 1995), connect input to every layer:



- This fixes the problem: Locally up to D degrees of freedom, at any depth:

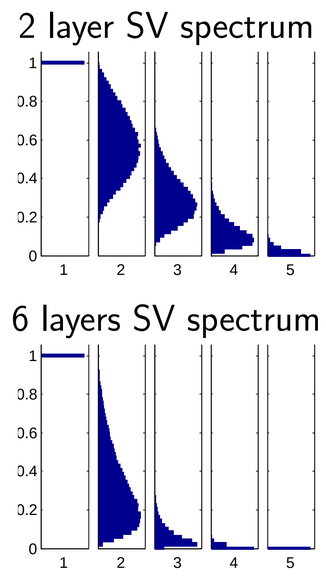


Pathology is also resolved in deep density models: Density does not concentrate along filaments when input connects to all layers.



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 $\mathbf{y}$.

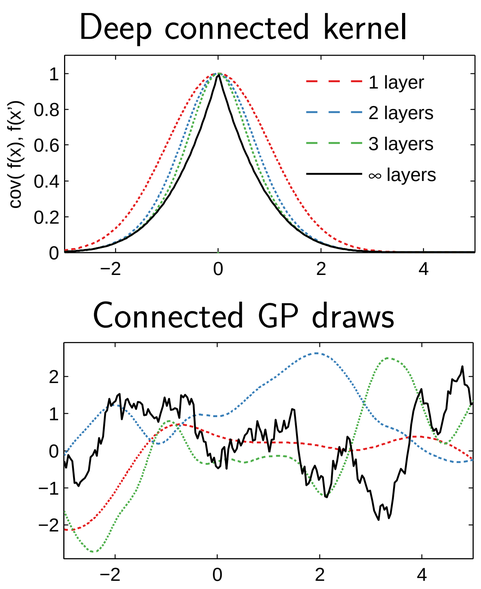


Infinitely Deep Kernels

- If $k_1(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\mathsf{T}\mathbf{h}(\mathbf{x}')$, $k_2(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{h}(\mathbf{x}))^\mathsf{T}\mathbf{h}(\mathbf{h}(\mathbf{x}'))$.
- Closed form for squared-exp kernel:

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

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



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 derivatives.
- Can build "deep net" kernels