Avoiding Pathologies in Very Deep Networks

David Duvenaud, Oren Rippel, Ryan P. Adams, Zoubin Ghahramani

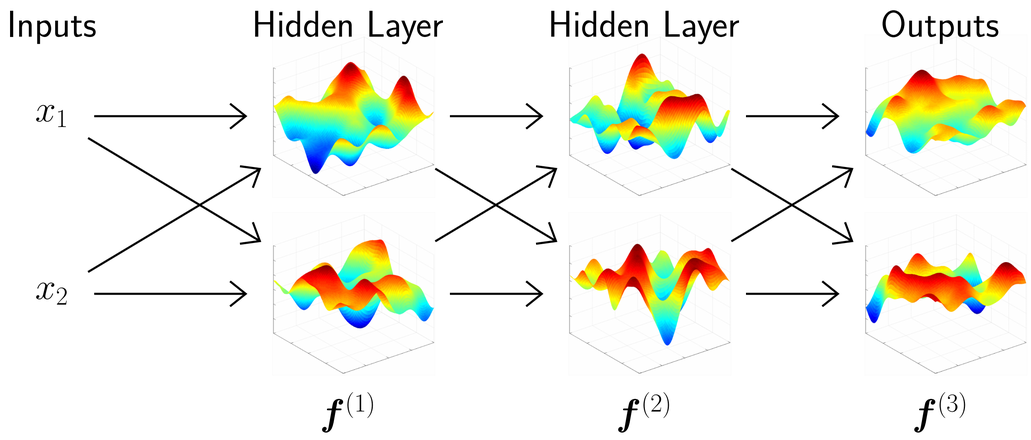
UNIVERSITY OF CAMBRIDGE

HARVARD

Abstract

- We compare architectures by building priors over deep nets
- Characterize a pathology in standard architecture
- Show a simple alternative architecture which fixes the problem.

Nonparametric Priors on Deep Neural Networks

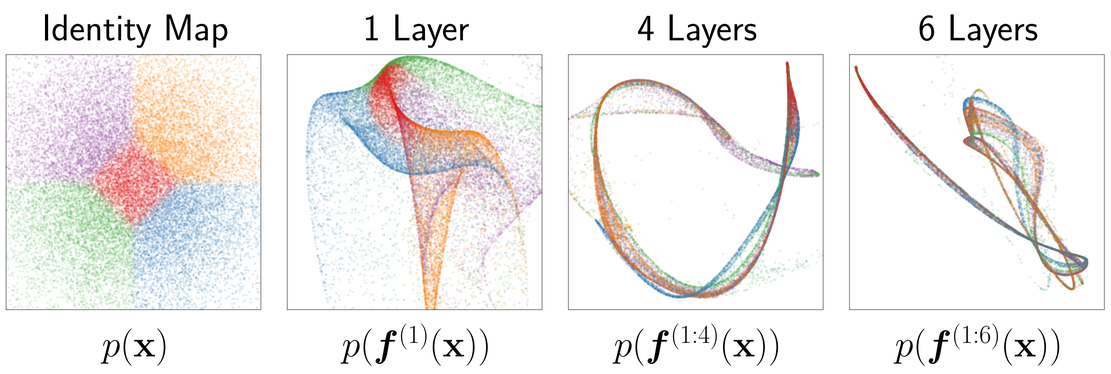


Deep GPs are compositions of functions, each $f^{(\ell)} \overset{\text{ind}}{\sim} \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

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

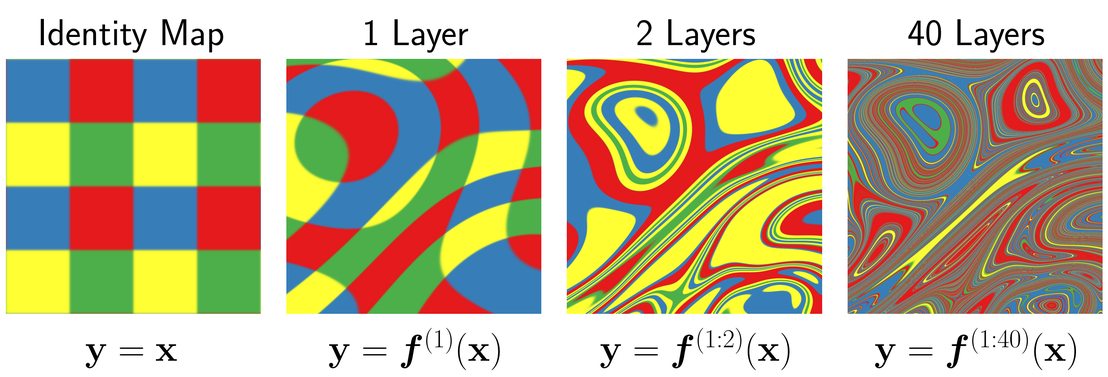
Random Deep Nets Capture Few Degrees of Freedom

A distribution warped by a function drawn from a deep GP prior:



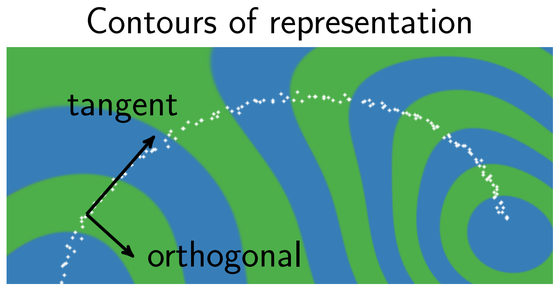
As depth increases, the density concentrates along one-dimensional filaments.

Sampled mappings illustrate properties of this prior on functions:



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

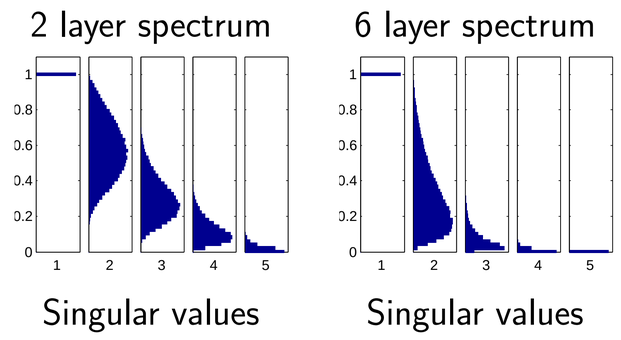
Good Representations Change Along All Tangents



Representation $\mathbf{y} = \boldsymbol{f}(\mathbf{x})$ must change in directions tangent to the data manifold, to preserve information. (Rifai et. al., 2011)

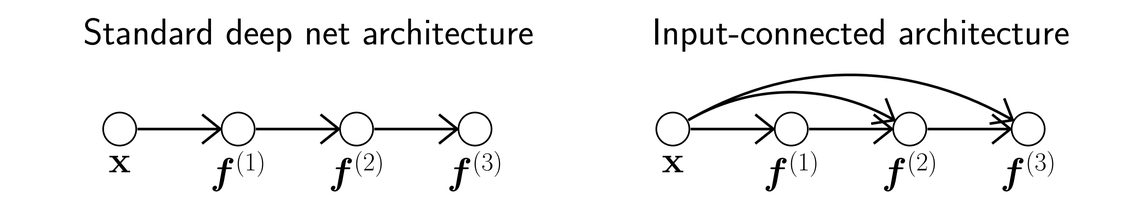
Explaining the Pathology

- Jacobian of a deep GP is a product of independent Gaussian matrices.
- Singular value spectrum shows relative size of derivatives.
- As net deepens, one direction has much larger derivative than others.

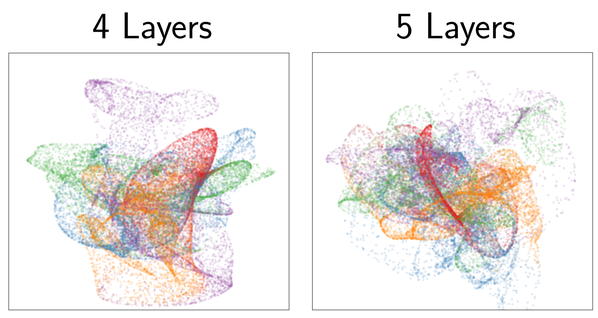


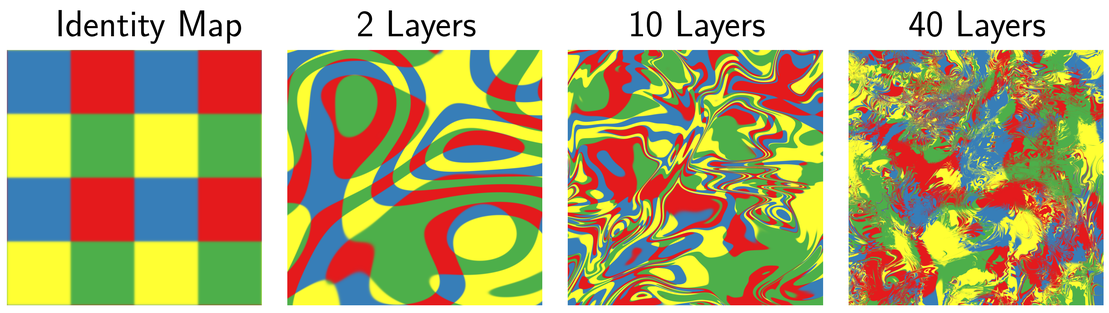
Fixing the pathology

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



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





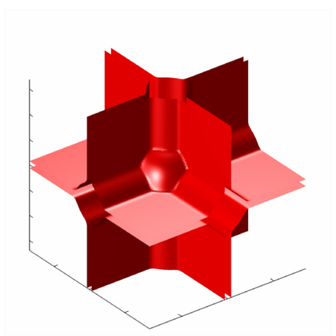
Locally up to D degrees of freedom, at any depth.

Other Analyses

Dropout in Gaussian Processes

- One-layer GPs are infinitely-wide neural nets
- Dropping out features has no effect
- Dropping out inputs gives mixture of GPs
- This mixture has closed-form covariance

$$\text{Cov}\left[f(\mathbf{x}'), f(\mathbf{x})\right] = \frac{1}{2^D} \sum_{\mathbf{R} \in \{0,1\}^D} \prod_{d=1}^{D} k_d(\mathbf{x}_d, \mathbf{x}'_d)^{r_d}$$



Dropout covariance isocontour

Infinitely Deep Kernels

- Kernels correspond to feature mappings:

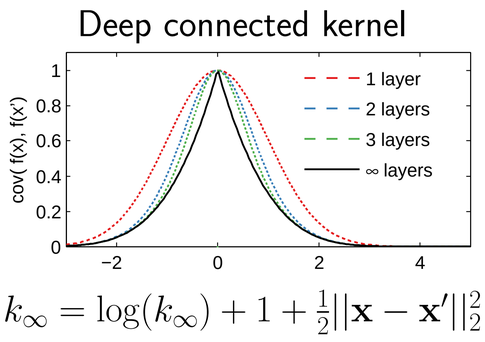
$$k_1(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\mathsf{T} \mathbf{h}(\mathbf{x}')$$

- Compose feature maps for deep kernels:

$$k_2(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{h}(\mathbf{x}))^\mathsf{T} \mathbf{h}(\mathbf{h}(\mathbf{x}'))$$

Code at `github.com/duvenaud/deep-limits`

Paper at `arxiv.org/abs/1402.5836`



$$k_\infty = \log(k_\infty) + 1 + \tfrac{1}{2}||\mathbf{x} - \mathbf{x}'||_2^2$$