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.
- We characterize a pathology in standard architectures.
- We show a simple alternative architecture that fixes the problem.

Nonparametric priors on deep neural networks

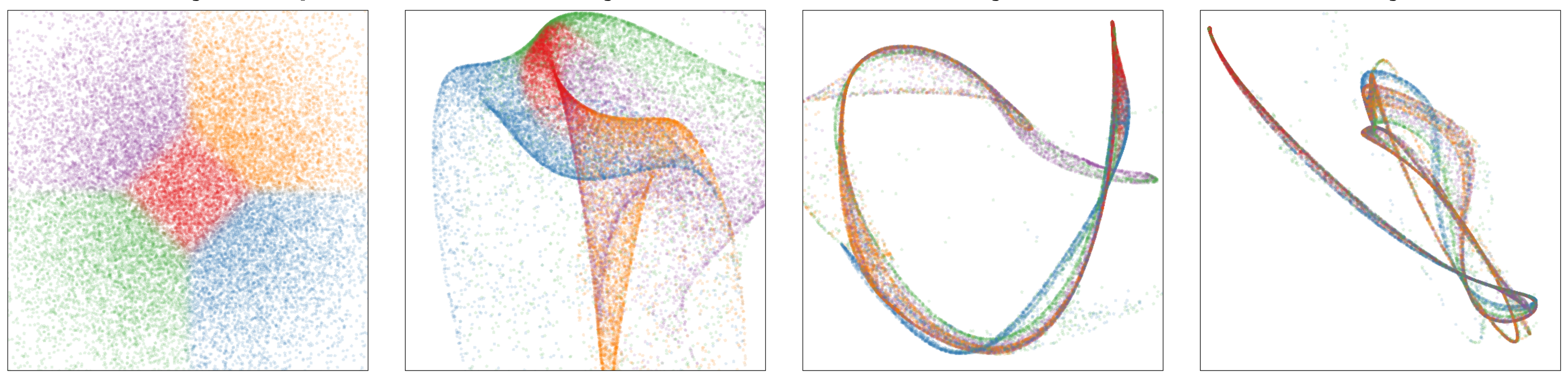
Deep GPs are compositions of functions, each $f^{(\ell)} \overset{\text{ind}}{\sim} \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

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

Random deep nets capture few degrees of freedom

A density warped by a deep-GP distributed function:

Identity Map | 1 Layer | 4 Layers | 6 Layers

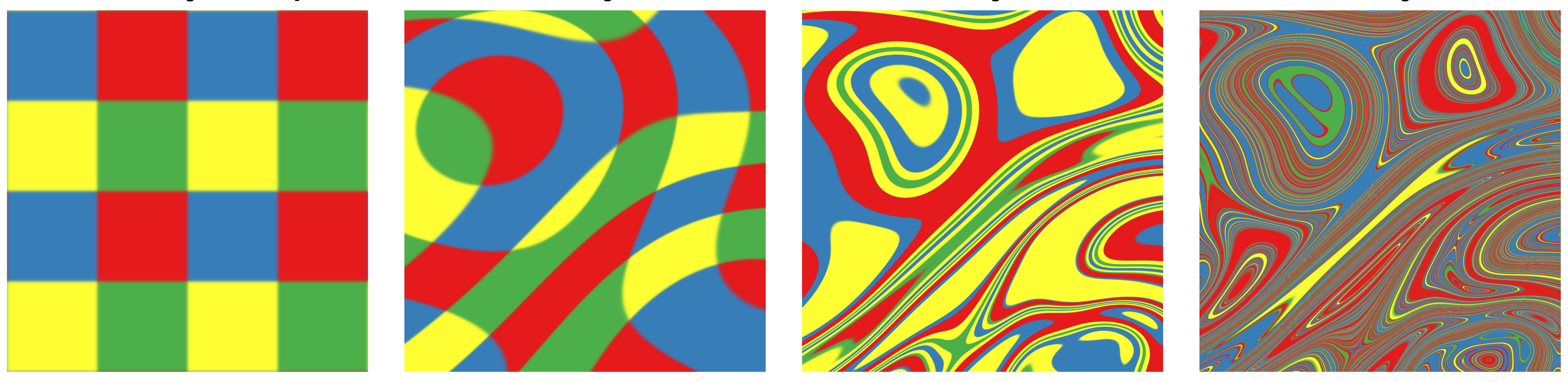


$p(\mathbf{x})$ | $p(\boldsymbol{f}^{(1)}(\mathbf{x}))$ | $p(\boldsymbol{f}^{(1:4)}(\mathbf{x}))$ | $p(\boldsymbol{f}^{(1:6)}(\mathbf{x}))$

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

Sampled mappings illustrate properties of this prior on functions:

Identity Map | 1 Layer | 2 Layers | 40 Layers

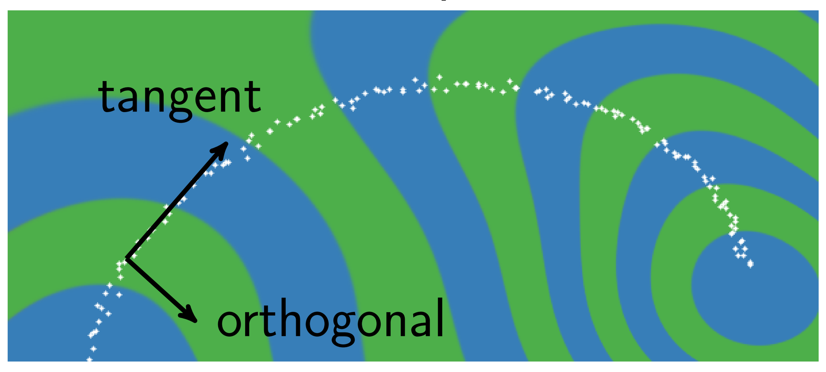


$\mathbf{y} = \mathbf{x}$ | $\mathbf{y} = \boldsymbol{f}^{(1)}(\mathbf{x})$ | $\mathbf{y} = \boldsymbol{f}^{(1:2)}(\mathbf{x})$ | $\mathbf{y} = \boldsymbol{f}^{(1:40)}(\mathbf{x})$

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

Good representations change along all tangents

Contours of representation

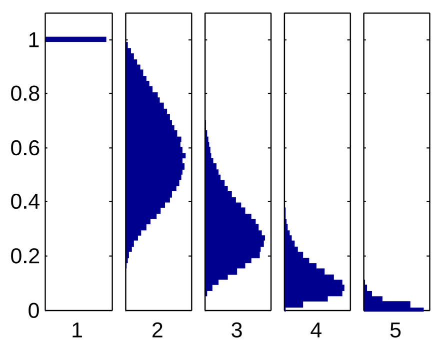


Representation $\mathbf{y} = \boldsymbol{f}(\mathbf{x})$ must change in directions tangent to the data manifold, to preserve information. (Rifai et. al., 2011)

Explaining the pathology

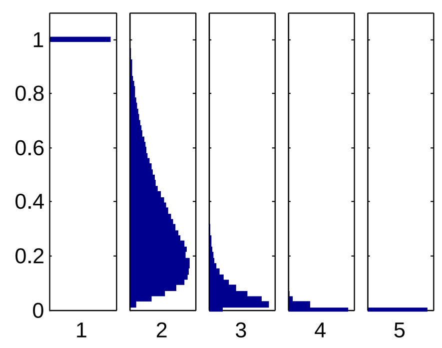
- The Jacobian of a deep GP is a product of independent Gaussian matrices.
- Singular value spectrum shows relative size of derivatives.
- As the net deepens, the derivative in one direction becomes much larger than all the others.

2 layer spectrum



Singular values

6 layer spectrum

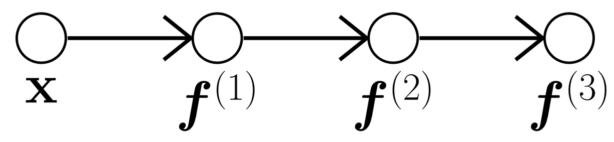


Singular values

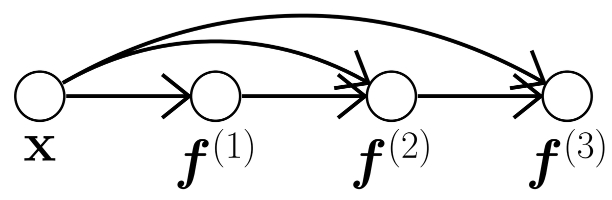
Fixing the pathology

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

Standard deep net architecture

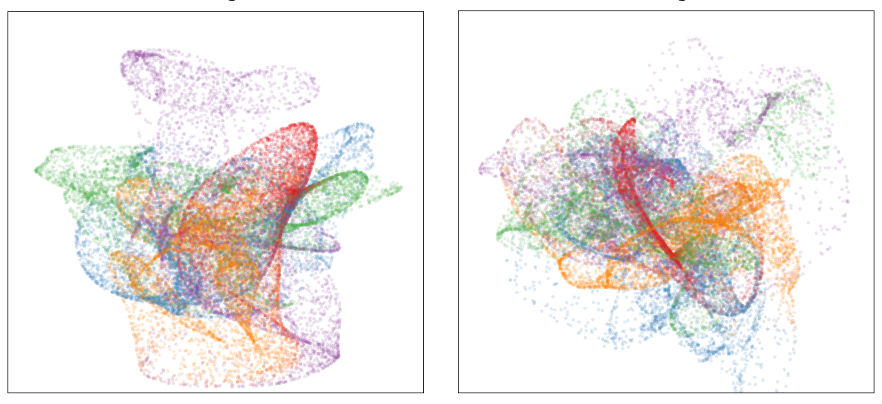


Input-connected architecture

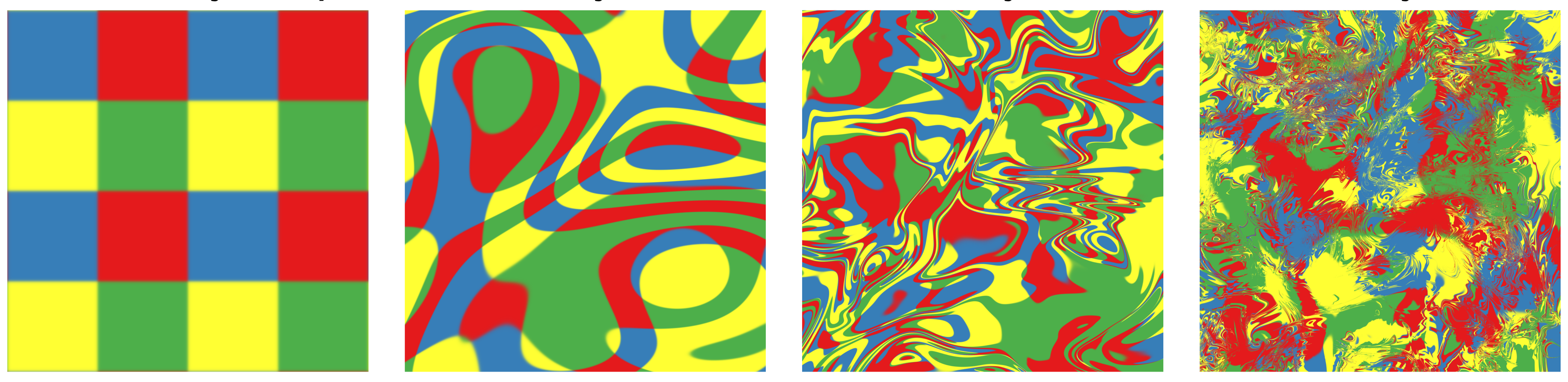


Pathology is now resolved in deep density models: Density does not concentrate along filaments when the input connects to all layers.

4 Layers | 5 Layers



Identity Map | 2 Layers | 10 Layers | 40 Layers



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

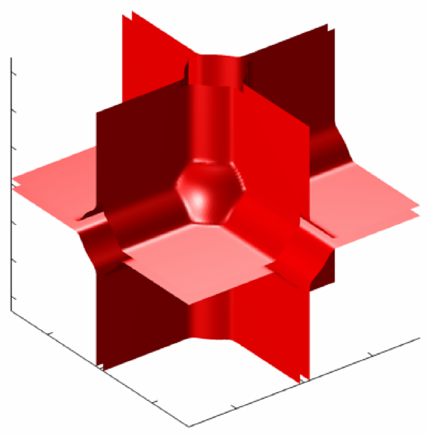
Other Results

Dropout in Gaussian processes

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

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

Isocontour of dropout covariance



$\mathbf{x} - \mathbf{x}'$

Infinitely deep kernels

- Kernels correspond to feature mappings:

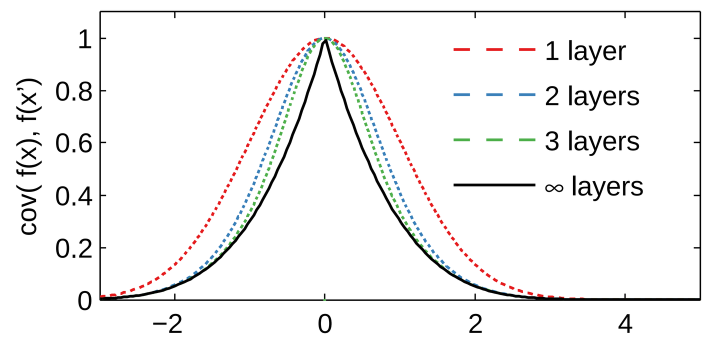
$$k_1(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^{\mathsf{T}} \mathbf{h}(\mathbf{x}')$$

- Can compose feature maps to get deep kernels: (Cho, 2012)

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

Deep connected kernel

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



$x - x'$

Code at `github.com/duvenaud/deep-limits`
Paper at `arxiv.org/abs/1402.5836`