Avoiding Pathologies in Very Deep Networks

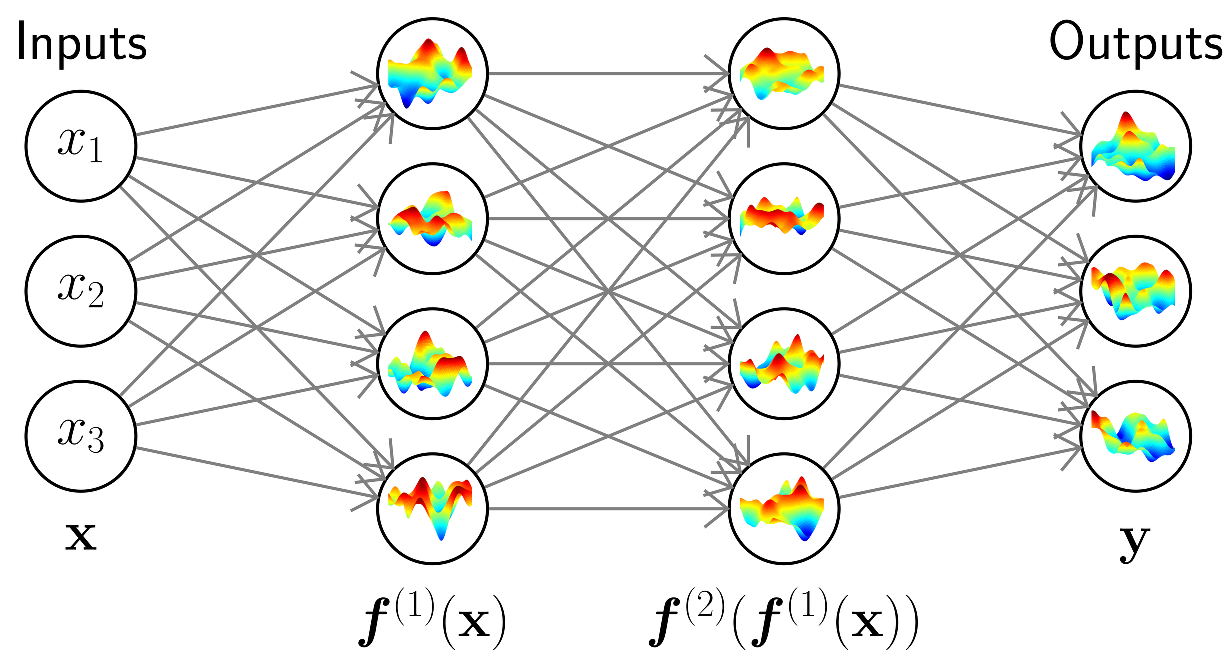
David Duvenaud, Oren Rippel, Ryan P. Adams, Zoubin Ghahramani

Main Idea

- We compare network architectures by analyzing priors on deep nets.
- We characterize a pathology in standard architectures.
- A simple alternative architecture fixes the problem.

A nonparametric prior on deep neural networks

We examine deep Gaussian processes (Damianou and Lawrence, 2012)

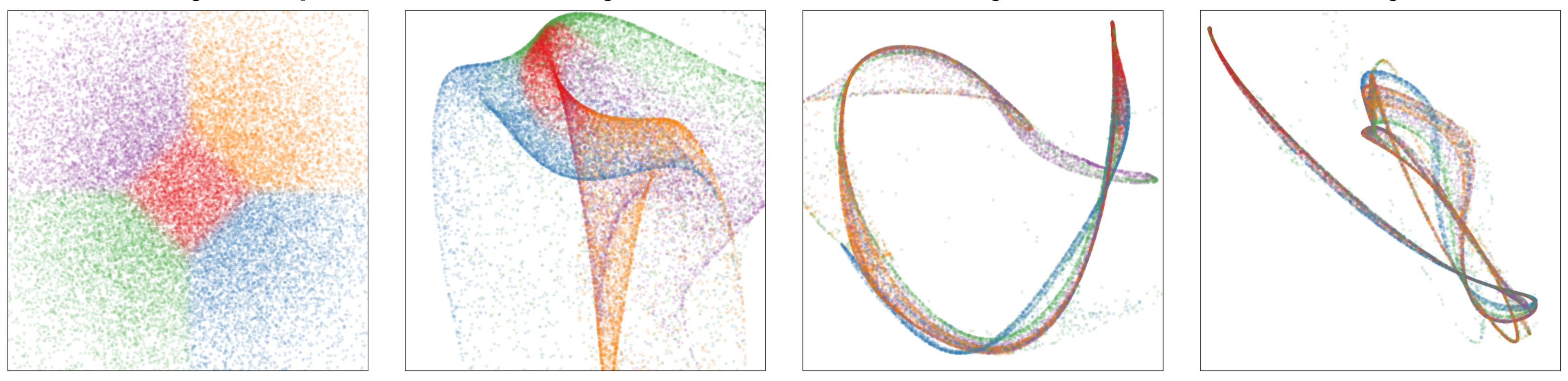


Deep GPs are compositions of functions, each $f^{(\ell)} \overset{\text{ind}}{\sim} \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

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

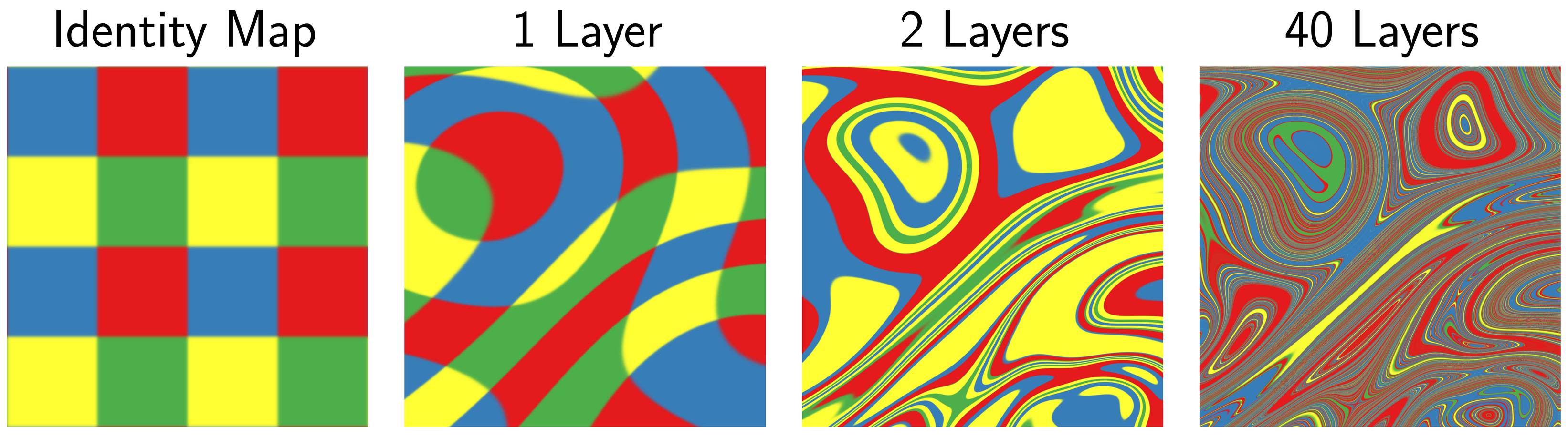
Random deep nets vary in few directions

A density warped by a deep-GP distributed function:



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

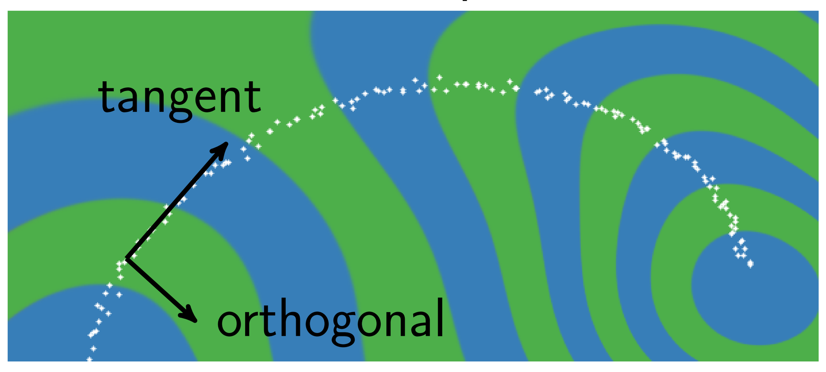
Sampled mappings illustrate properties of this prior on functions:



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

Good representations vary along the data manifold

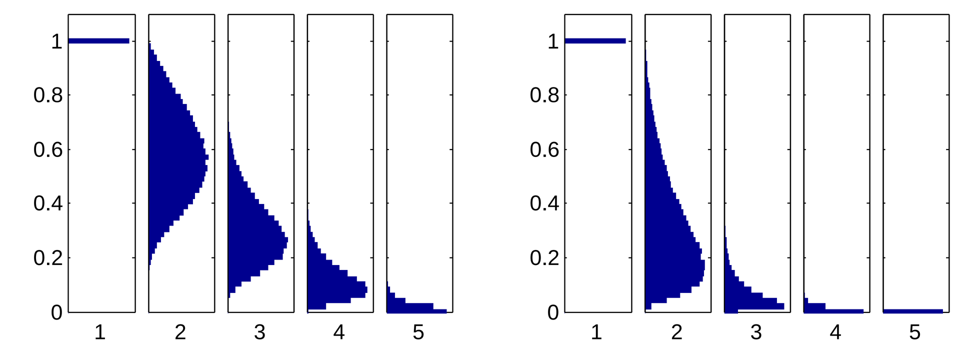
Contours of a representation



Representation $\mathbf{y} = \boldsymbol{f}(\mathbf{x})$ must vary in at least as many directions as the number of dimensions of the data manifold. (Rifai et. al., 2011)

Explaining the pathology

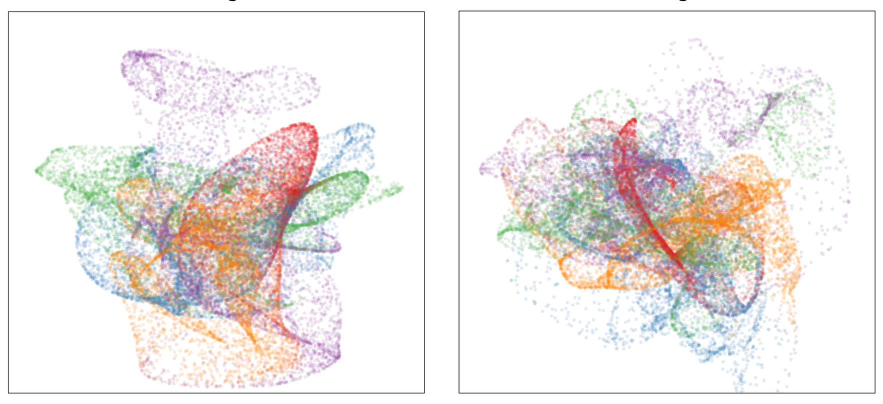
- The Jacobian of a deep GP is a product of independent Gaussian matrices.
- Singular value spectrum shows relative size of derivatives.
- As the net deepens, the derivative in one direction becomes much larger than all the others.

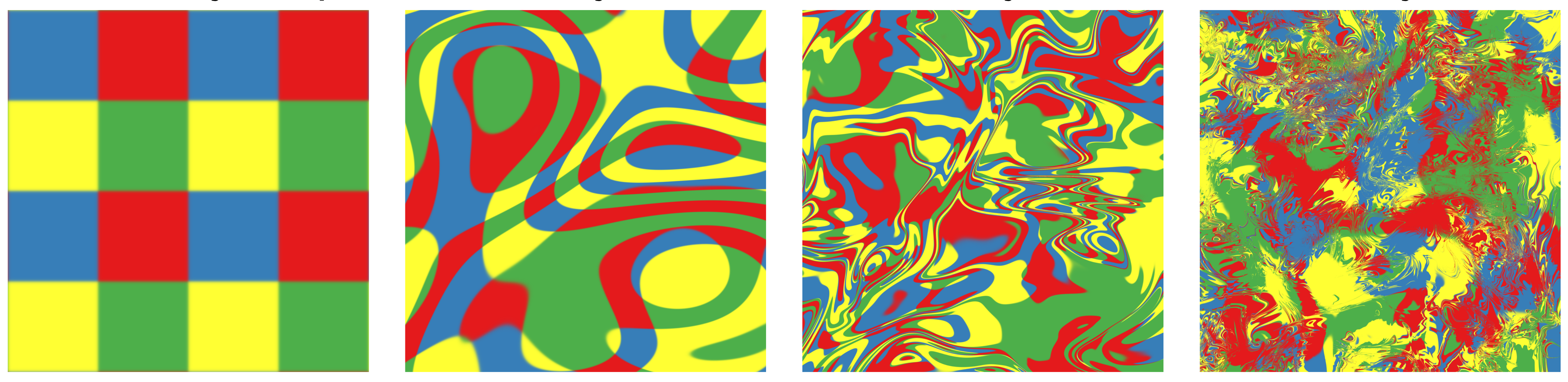


Fixing the pathology

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

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





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

Other Results

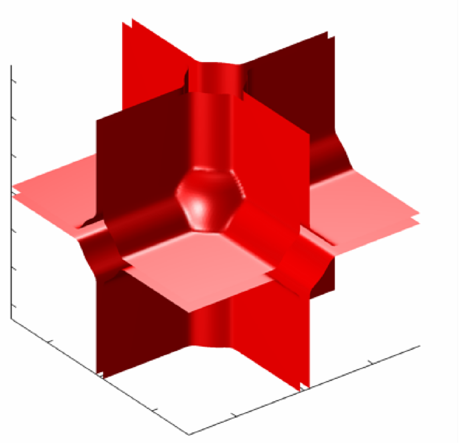
Dropout in Gaussian processes

- GPs are infinitely-wide shallow nets.
- Dropping out hidden units has no effect!
- Dropping out inputs gives a mixture of GPs, with tractable covariance:

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

- Same as additive GPs (Duvenaud et. al. 2011)

Isocontour of dropout kernel



Infinitely deep kernels

- Kernels correspond to feature mappings:

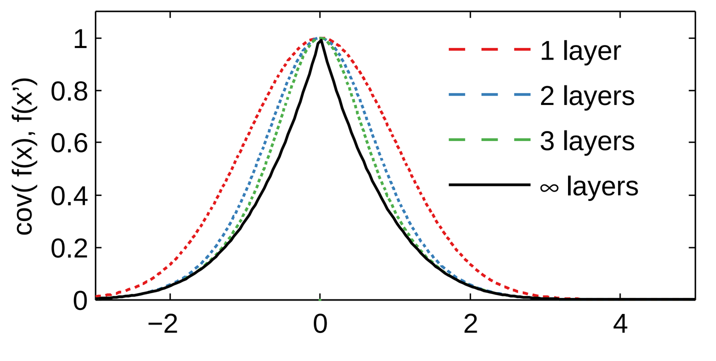
$$k_1(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^{\mathsf{T}}\mathbf{h}(\mathbf{x}')$$

- 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

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



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