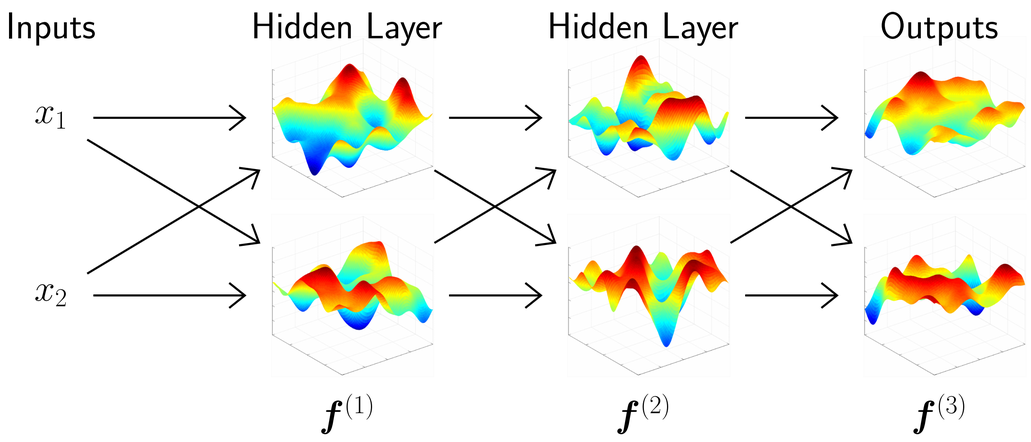
Avoiding Pathologies in Very Deep Networks

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

UNIVERSITY OF CAMBRIDGE

HARVARD

Let's Examine Priors on Deep Neural Networks

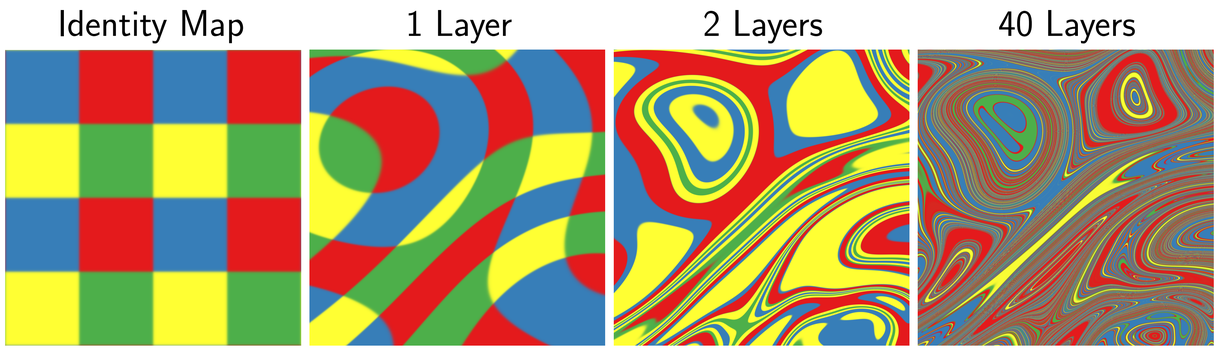


Deep GPs are compositions of functions, each $f^{(\ell)} \overset{\text{ind}}{\sim} \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

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

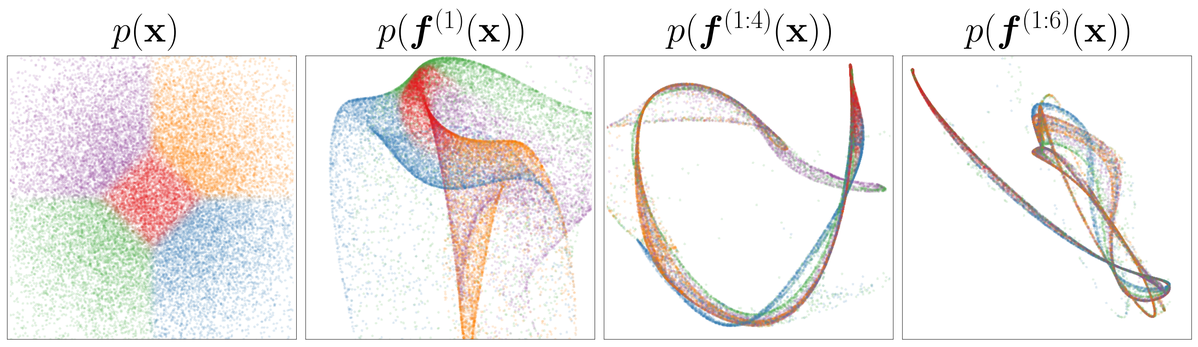
Random Deep Nets Capture Few Degrees of Freedom

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 $\mathbf{y}$.

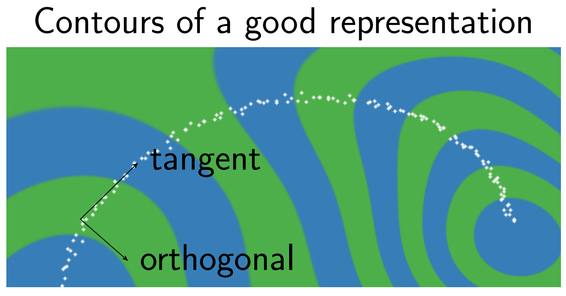
A distribution warped by successive functions drawn from a GP prior:



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

Good Representations Change Along All Tangents

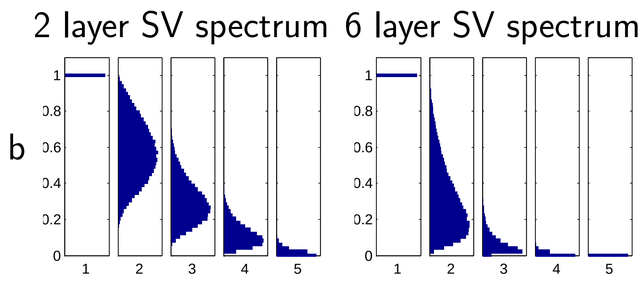
Contours of a good representation



- 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.

Explaining the Pathology

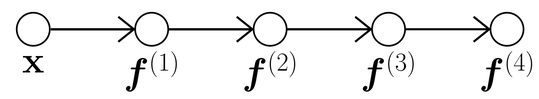
- Jacobian of a deep GP is a product of independent Gaussian matrices.
- Singular value spectrum shows relative size of derivatives.
- As net deepens, one direction has much larger derivative than others.



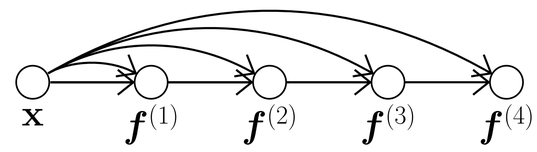
Fixing the pathology

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

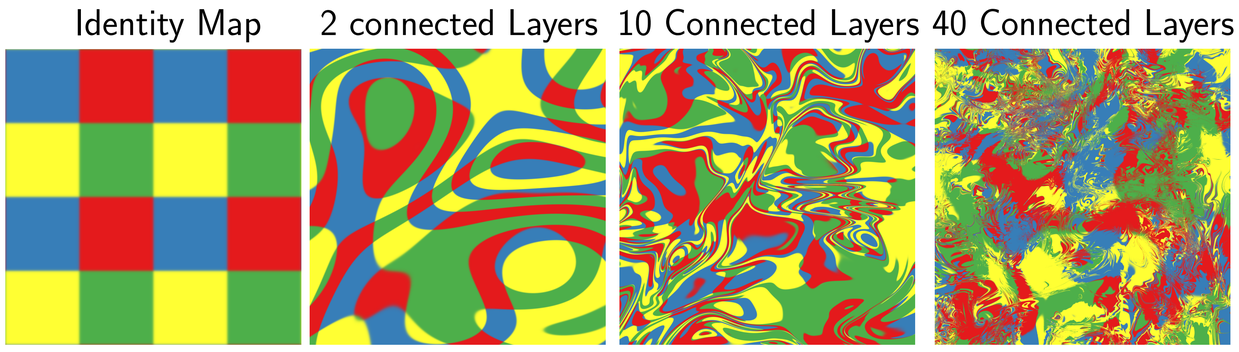
Standard deep net architecture



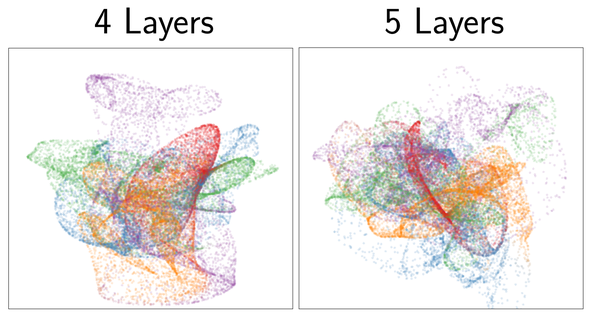
Input-connected architecture



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



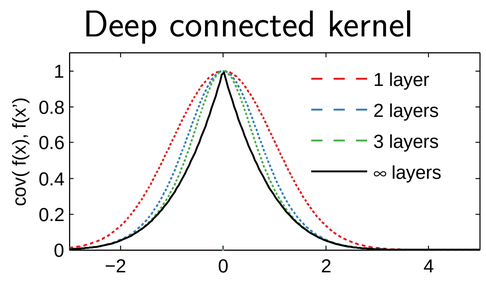
Pathology is also resolved in deep density models: Density does not concentrate along filaments when input connects to all layers.



Dropout in Gaussian Processes

Infinitely Deep Kernels

- If $k_1(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\mathsf{T}\mathbf{h}(\mathbf{x}')$,
  $k_2(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{h}(\mathbf{x}))^\mathsf{T}\mathbf{h}(\mathbf{h}(\mathbf{x}'))$.
- Recurrent limit for squared-exp kernel:
- $k_\infty - \log(k_\infty) = 1 + \frac{1}{2}||\mathbf{x} - \mathbf{x}'||_2^2$



Conclusions

- Random networks capture fewer degrees of freedom as they get deeper
- Connecting the input to each layer resolves this pathology
- Can build arbitrarily deep kernels

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

Paper at `mlg.eng.cam.ac.uk/duvenaud/papers/verydeep.pdf`