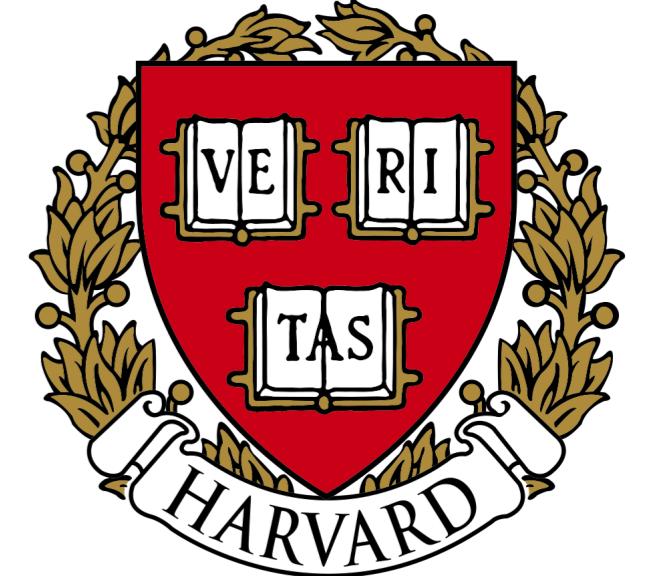
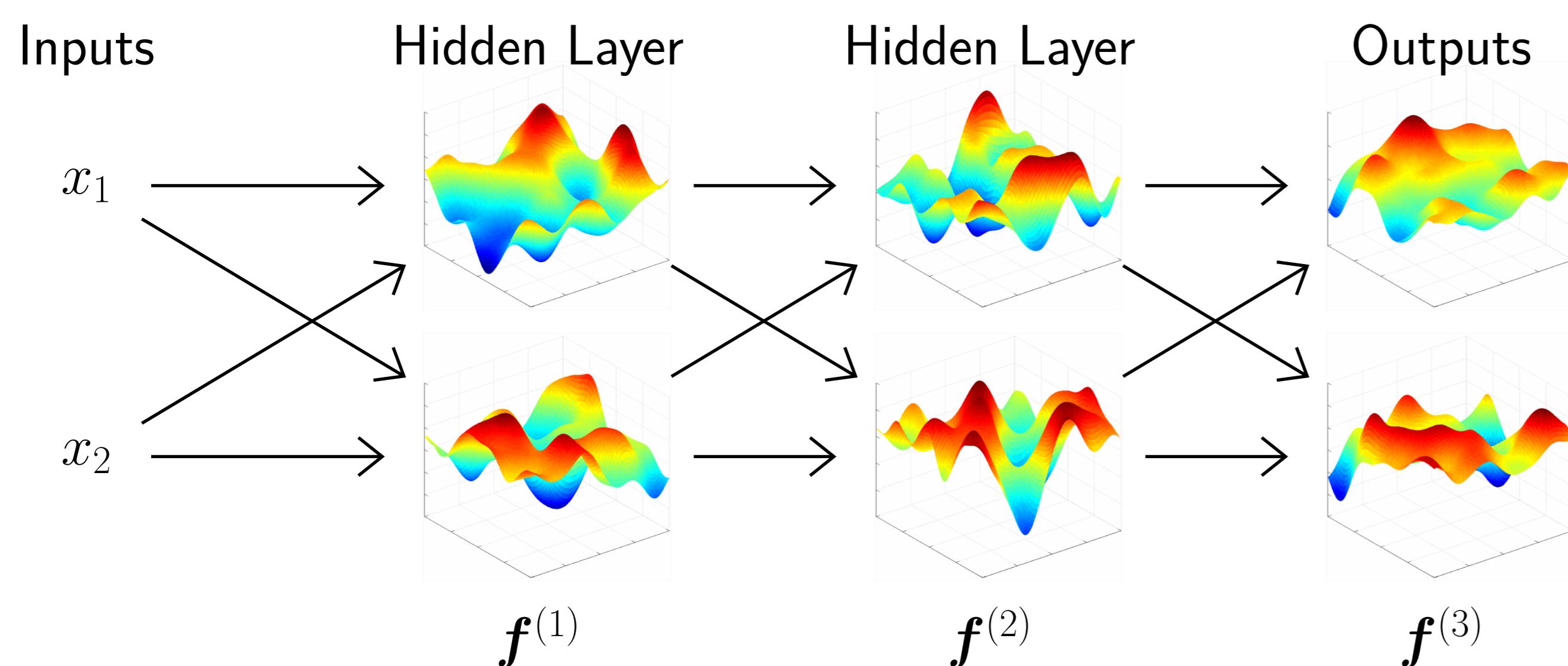


Avoiding Pathologies in Very Deep Networks

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



Nonparametric Priors on Deep Neural Networks

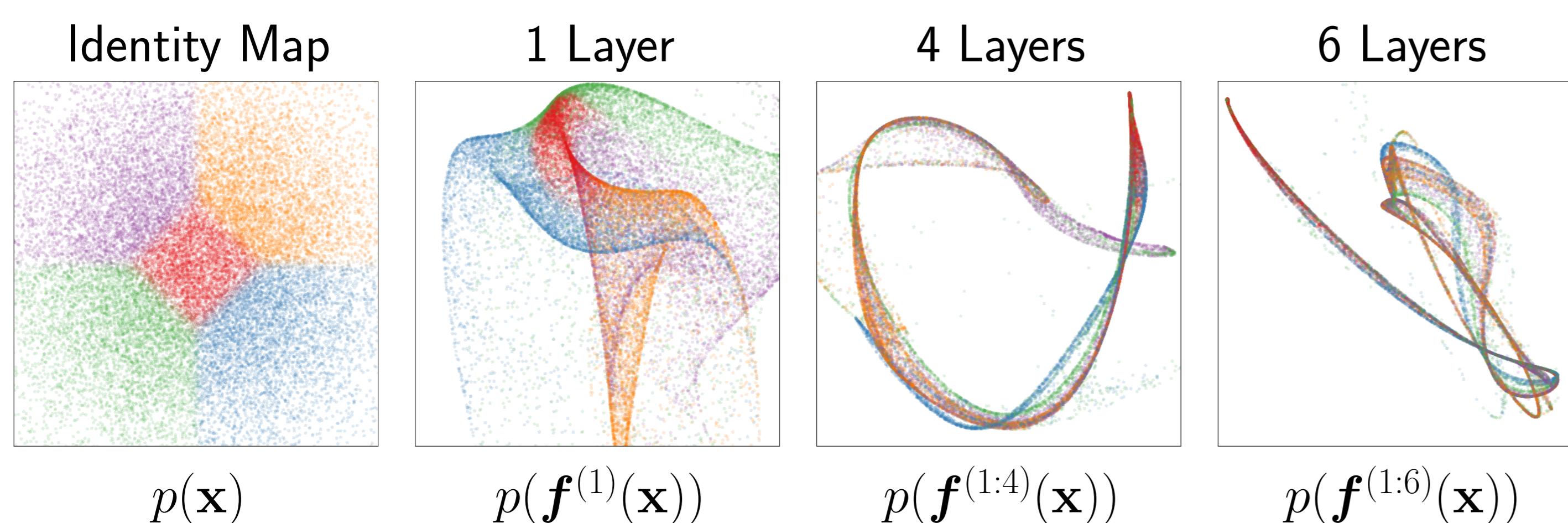


Deep GPs are compositions of functions, each $f^{(\ell)} \stackrel{\text{ind}}{\sim} \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

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

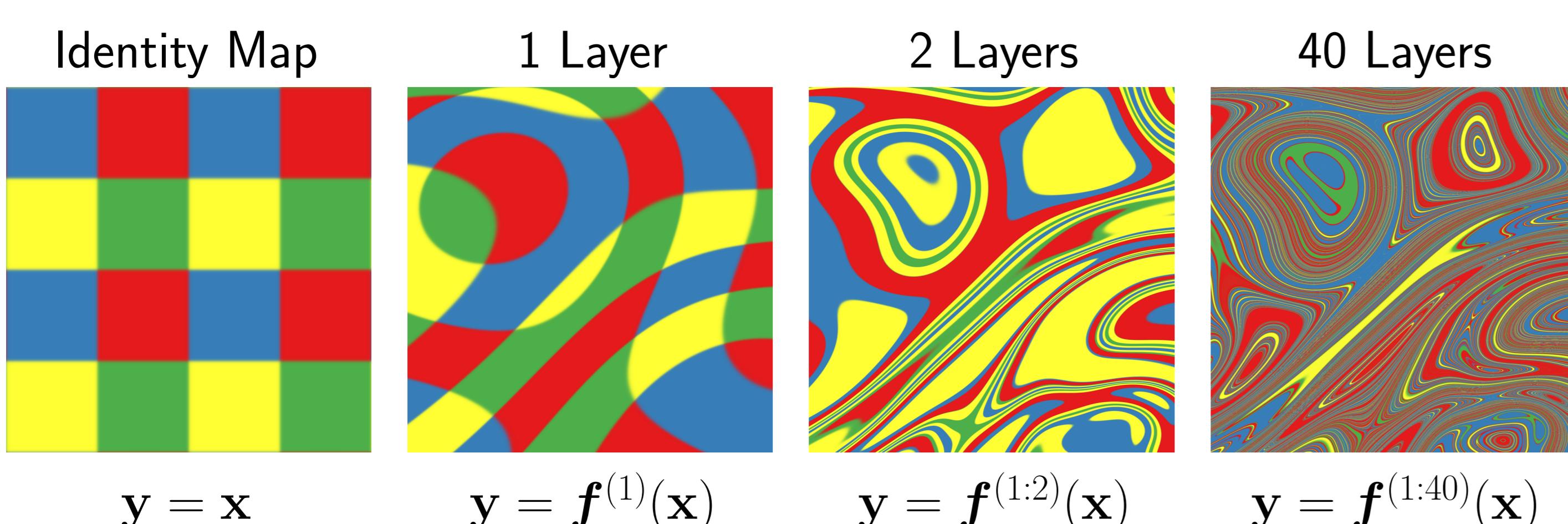
Random Deep Nets Capture Few Degrees of Freedom

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



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

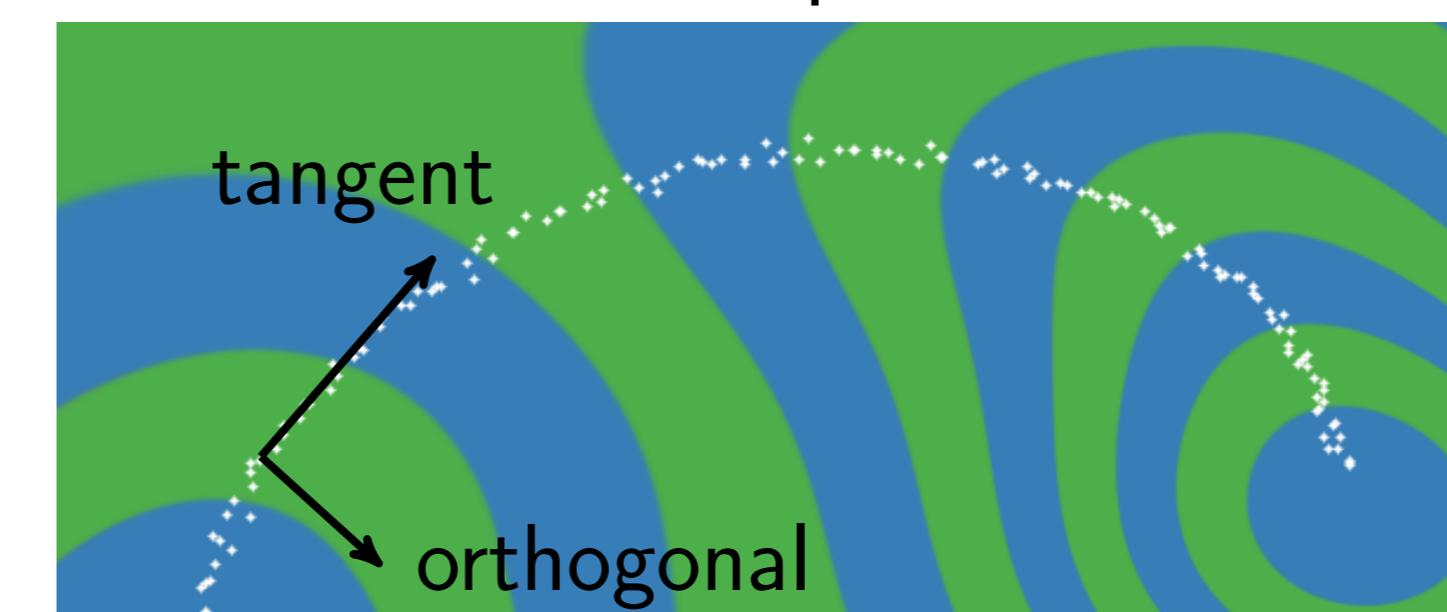
Sampled mappings illustrate properties of this prior on functions:



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

Good Representations Change Along All Tangents

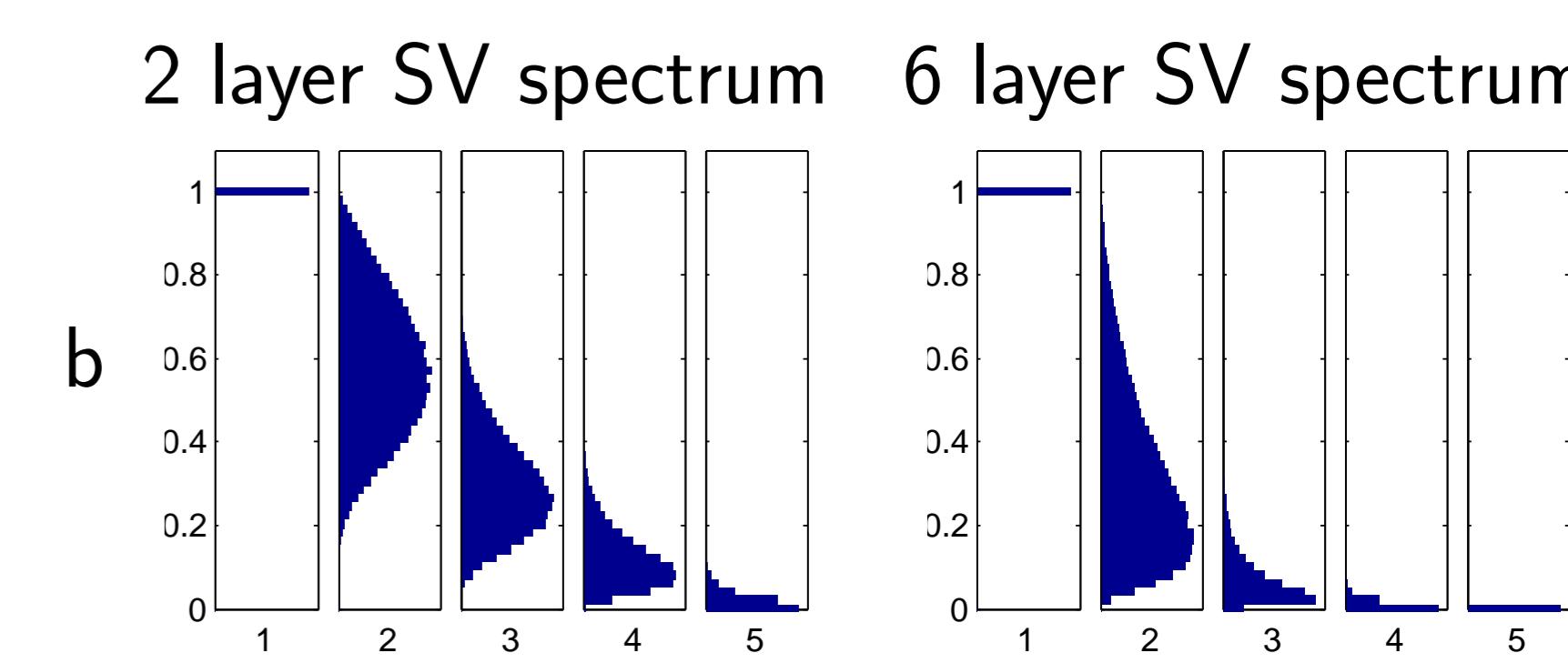
Contours of representation



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

Explaining the Pathology

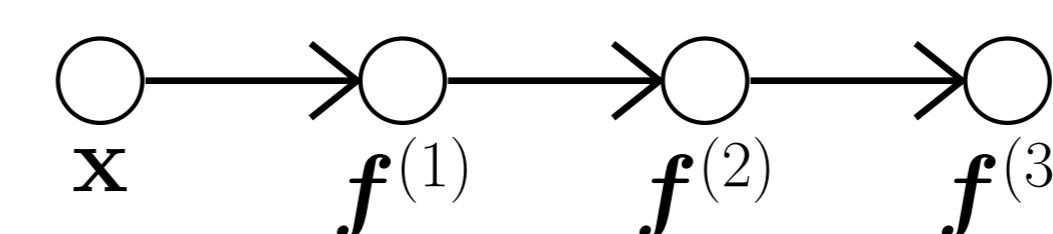
- Jacobian of a deep GP is a product of independent Gaussian matrices.
- Singular value spectrum shows relative size of derivatives.
- As net deepens, one direction has much larger derivative than others.



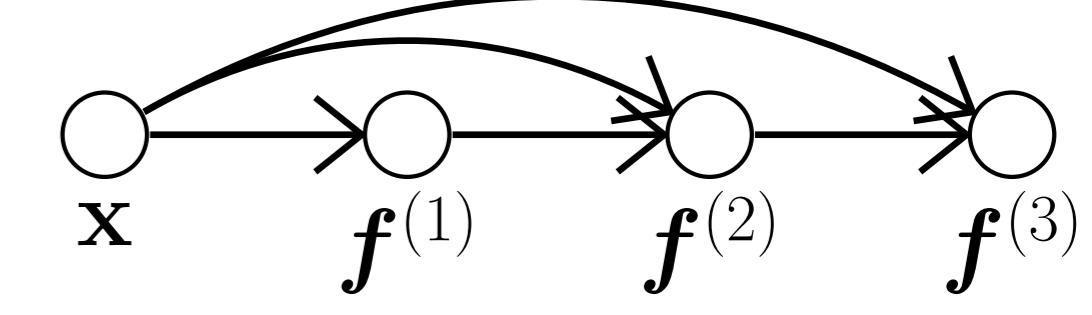
Fixing the pathology

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

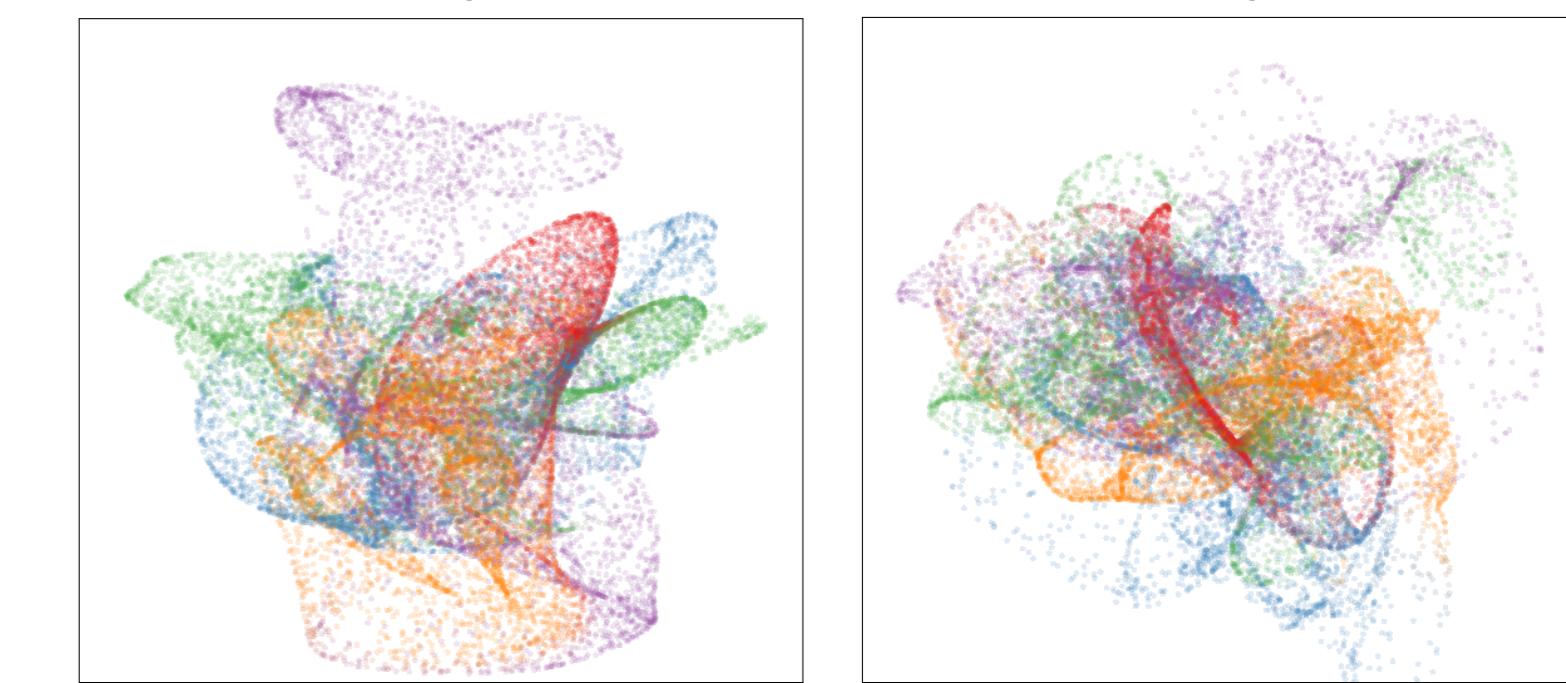
Standard deep net architecture



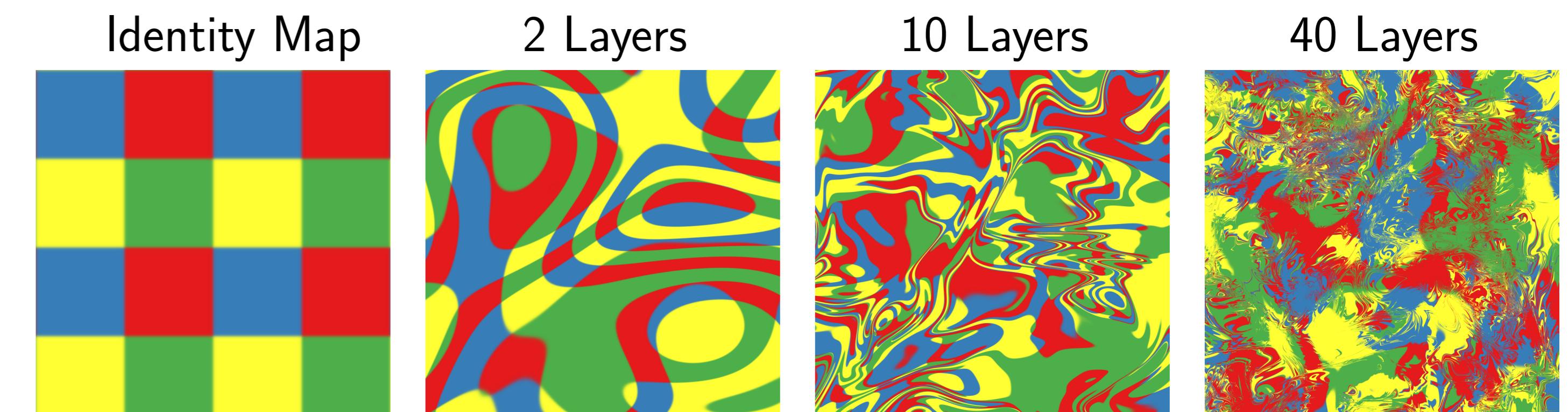
Input-connected architecture



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

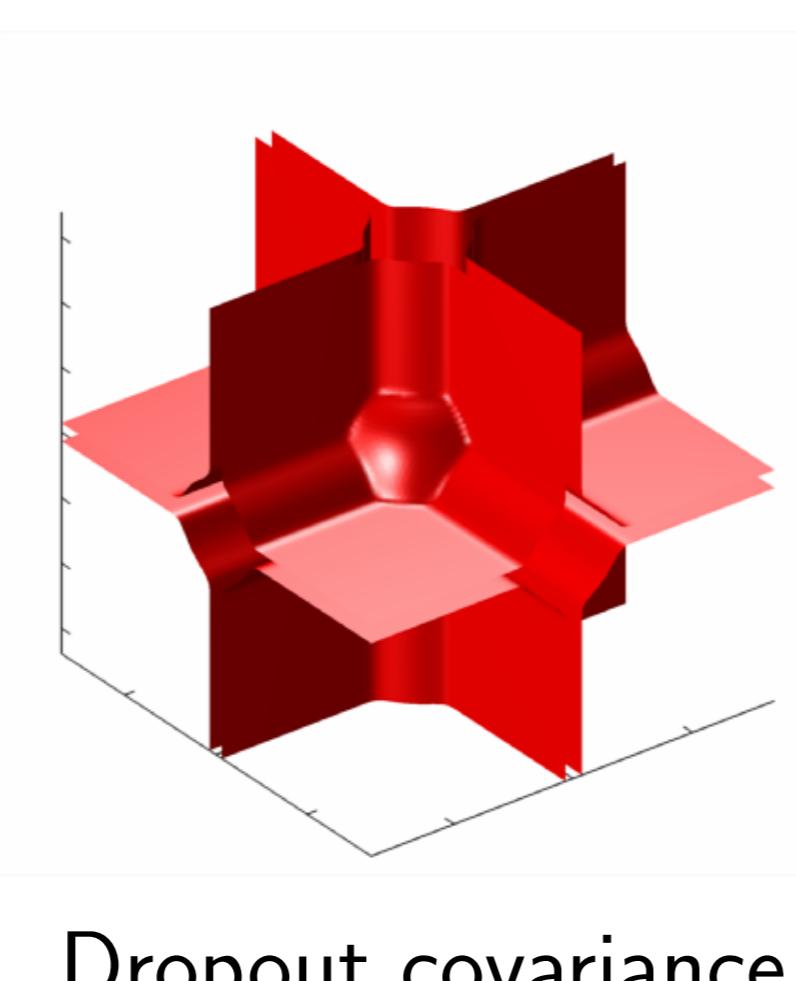


- Locally up to D degrees of freedom, at any depth:



Other Analyses

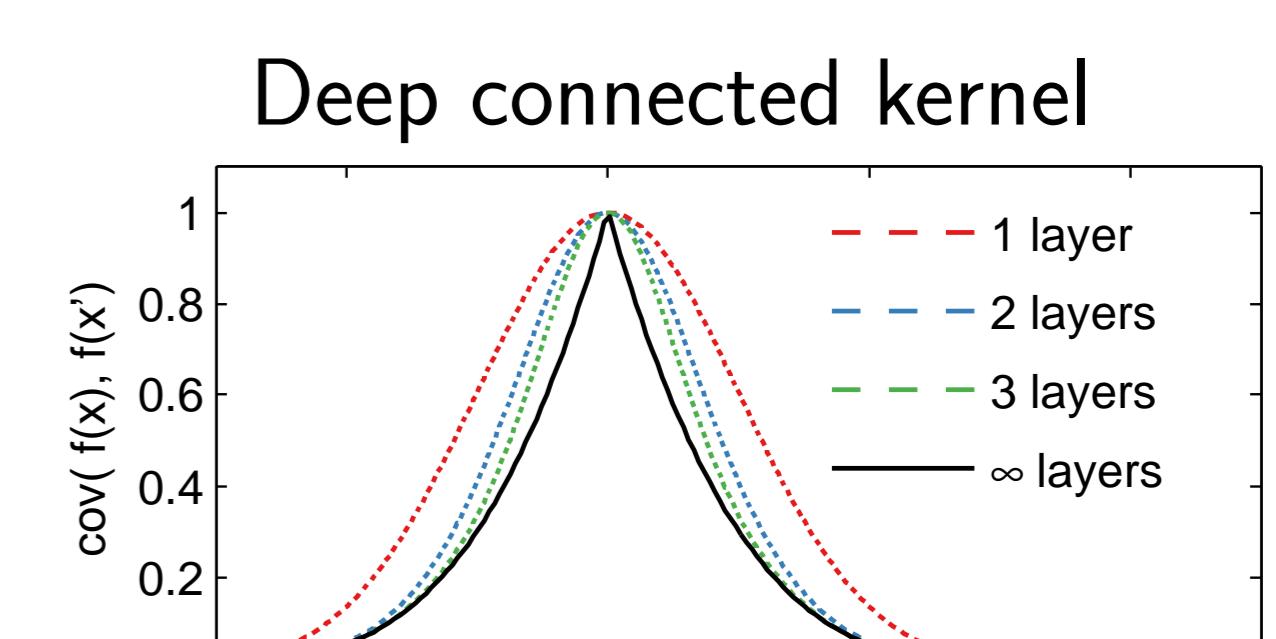
Dropout in Gaussian Processes



- 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

Infinitely Deep Kernels

- If $k_1(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\top \mathbf{h}(\mathbf{x}')$,
 $k_2(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{h}(\mathbf{x}))^\top \mathbf{h}(\mathbf{h}(\mathbf{x}'))$.
- Recurrent limit for squared-exp kernel:
- $k_\infty - \log(k_\infty) = 1 + \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_2^2$



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