

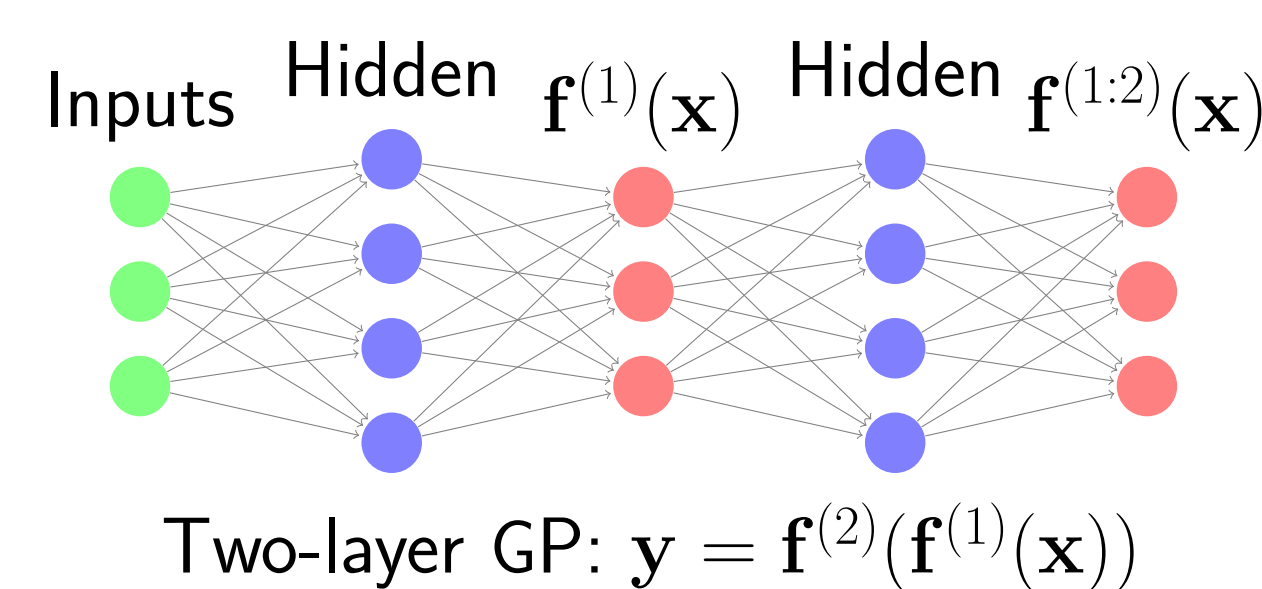
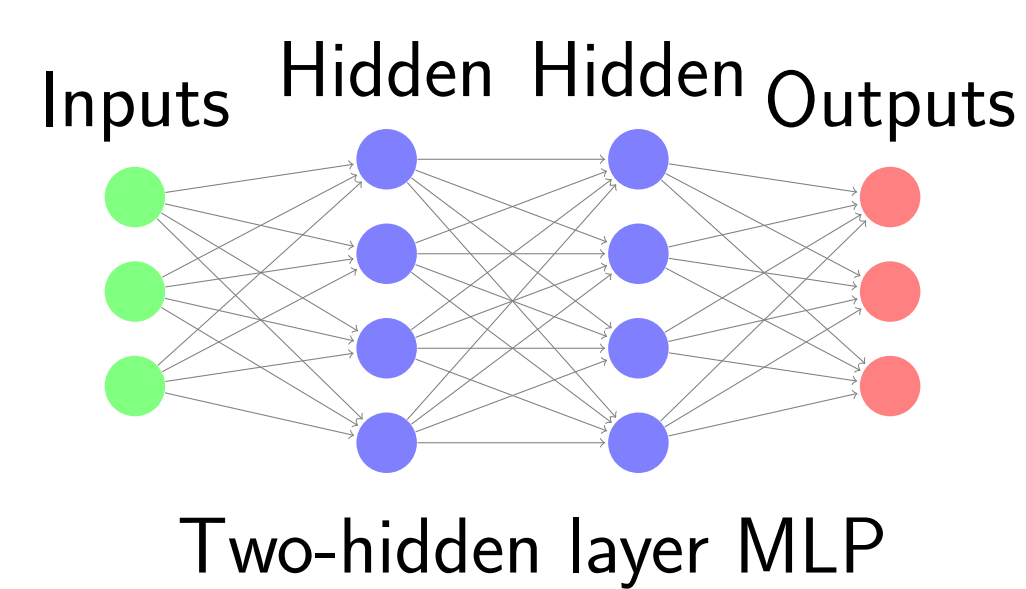
Avoiding Pathologies in Very Deep Networks

David Duvenaud, Oren Rippel, Ryan Adams, Zoubin Ghahramani

Abstract

- To study very deep networks, we analyze deep Gaussian processes as a simple proxy.
- We study distributions of deep GPs and find a pathology, then show a simple fix.
- We also derive kernels corresponding to infinitely deep nets.

Deep Nets and Deep Gaussian processes



Deep GPs are priors on compositions of functions:

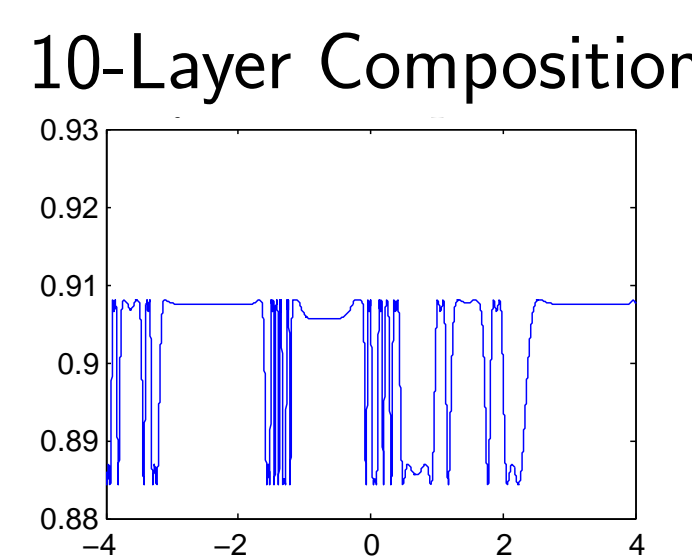
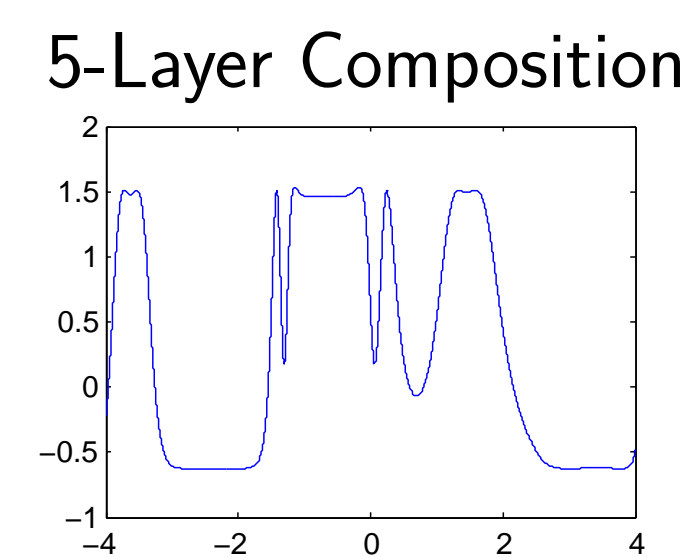
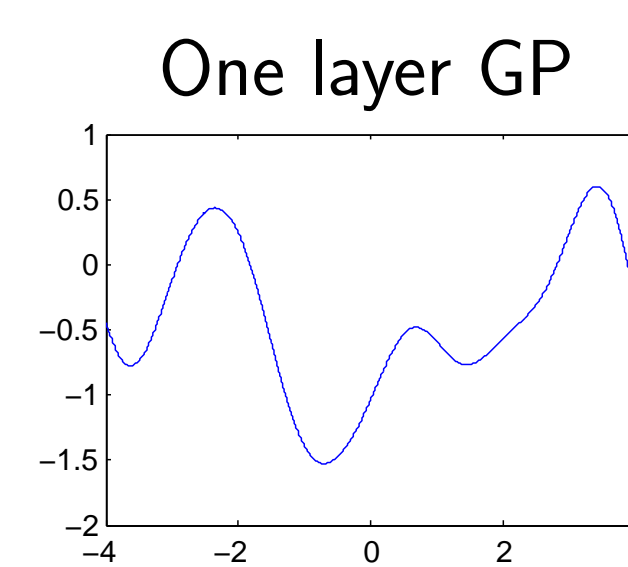
$$f^{(1:L)}(x) = f^{(L)}(f^{(L-1)}(\dots f^{(2)}(f^{(1)}(x)) \dots))$$

where each $f_d^{(\ell)} \sim \mathcal{GP}(0, k_d^\ell(x, x'))$. Can be viewed as either

1. MLPs with nonparametric activation functions
2. MLPs with infinitely-many parametric hidden nodes

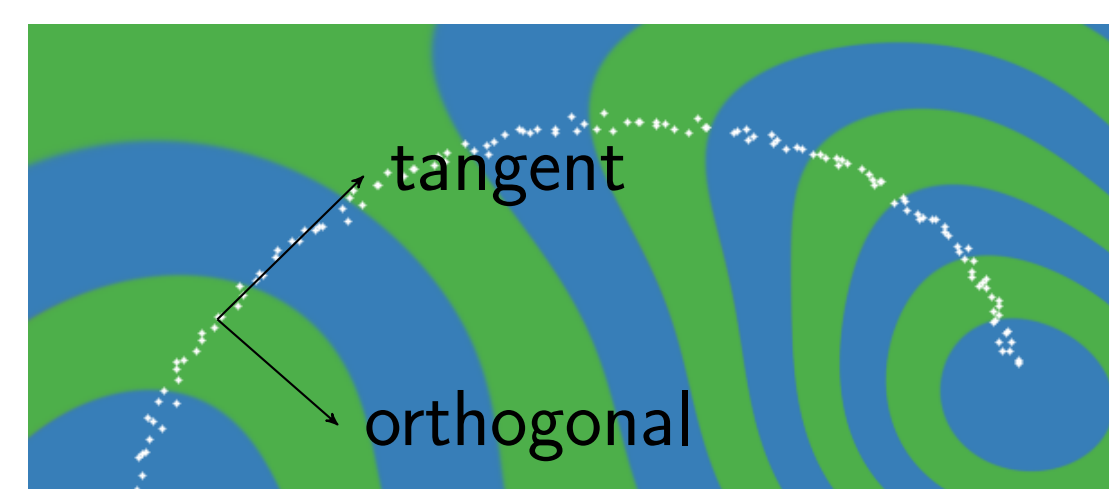
One-dimensional asymptotics

- Can examine deep GPs through distribution of derivatives.
- We use squared-exp kernel $k(x, x') = \sigma_f^2 \exp(-(x-x')^2/2\ell^2)$
- By the chain rule, the derivative of a deep GP is a product of independent derivatives, each normally distributed.
- The absolute value of derivative is a product of half-normals, each with mean $\sqrt{2\sigma_f^2/\pi\ell^2}$.
- By Central Limit Theorem, size of derivative has log-normal limiting distribution.
- Derivative becomes almost zero everywhere, with large jumps.



Degrees of Freedom of a Neural Network

- Neural net computes a representation $y = f(x)$ of data x .
- y needs to capture relevant degrees of freedom of x .

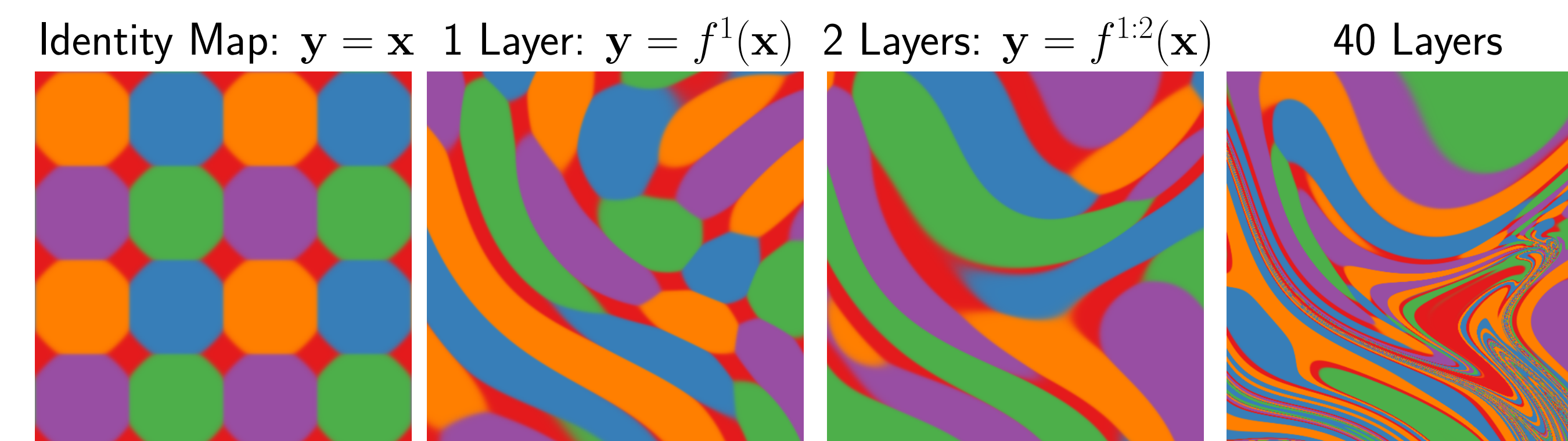


Contours show $f(x)$,
a noise-tolerant representation
of a data manifold (white)

- (Rifai et. al., 2011) argue that a good latent representation is invariant in directions orthogonal to the manifold on which the data lie.
- Conversely, a good latent representation must also change in directions tangent to the data manifold, in order to preserve relevant information.

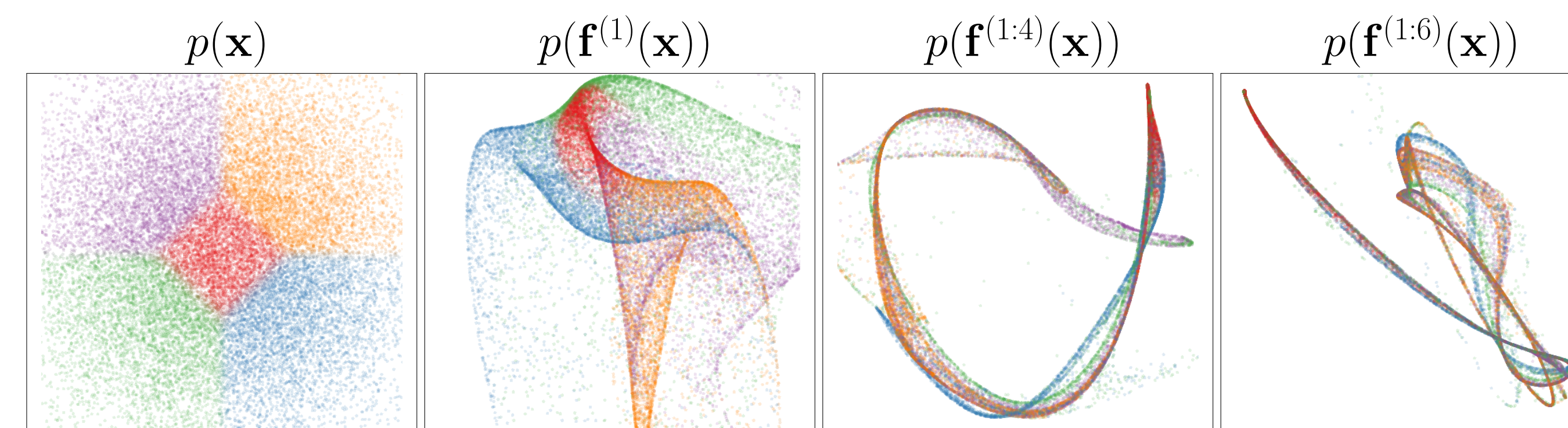
Random deep nets have few degrees of freedom

We visualize random mappings to show properties of prior on functions:



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

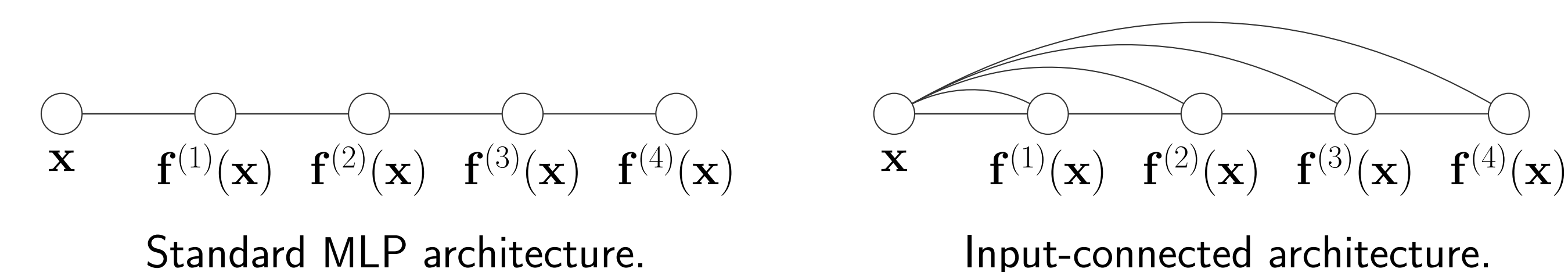
We also visualize a distribution warped by successive functions drawn from a GP prior:



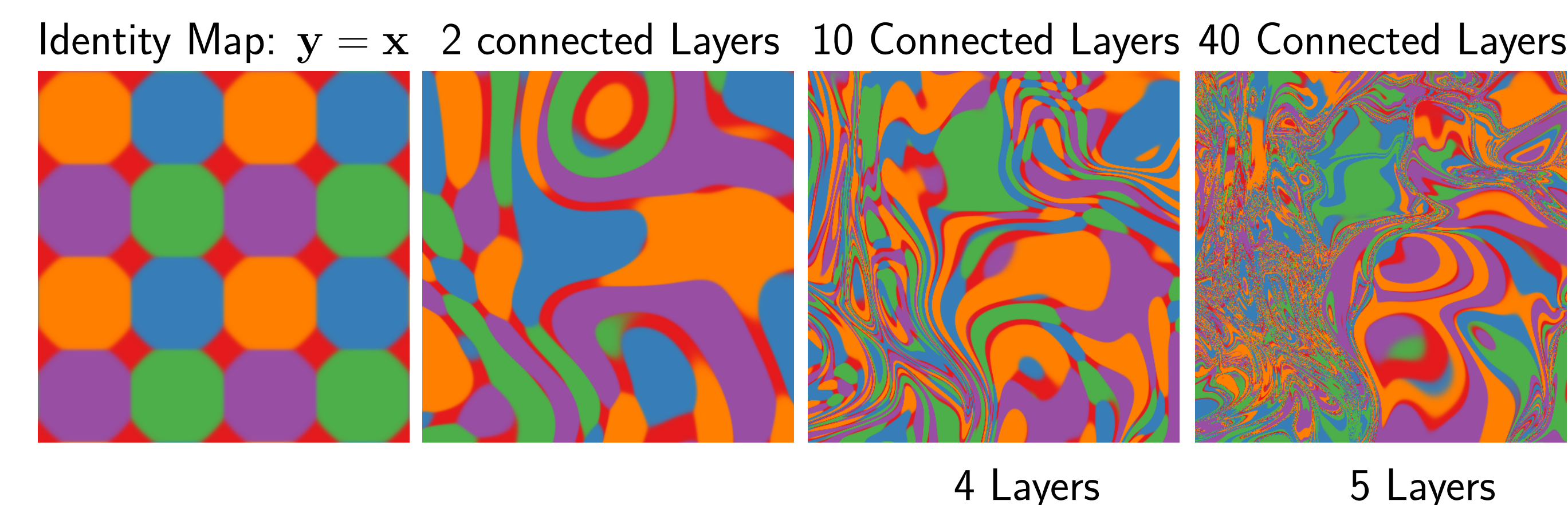
As depth increases, the density concentrates along one-dimensional filaments.

Fixing the pathology

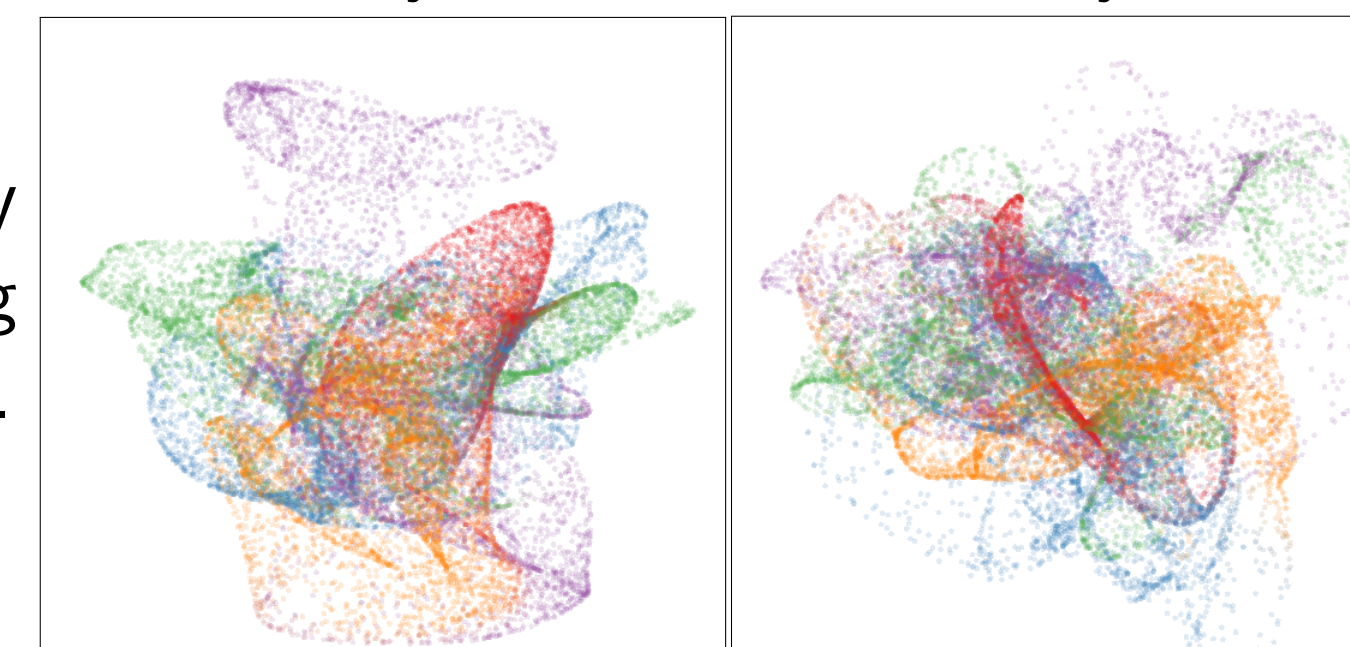
- Following a suggestion from (Neal, 1995), connect input to every layer:



- This fixes the problem: Locally there are usually D degrees of freedom, at any depth:

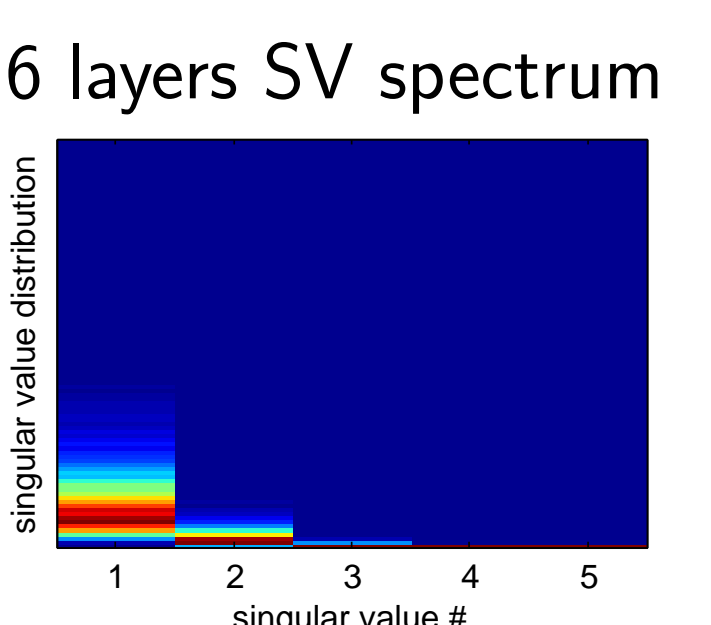
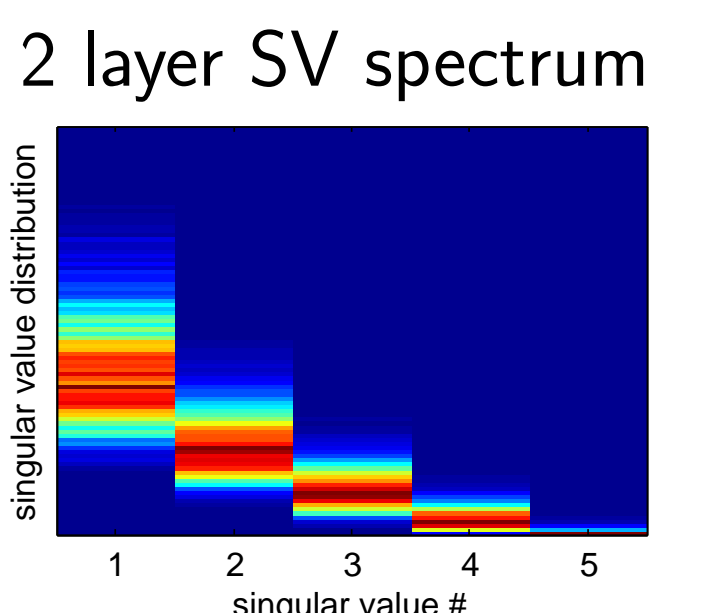


Pathology is also avoided in deep density models: Density does not concentrate along filaments when input connects to all layers.



Explaining the Pathology

- Jacobian of a deep GP is a product of independent Gaussian matrices.
- Singular values of Jacobian quantify relative magnitude of derivatives.
- As the net gets deeper, distribution becomes heavy-tailed, and the largest singular value dominates.
- Eventually, there is only one direction we can move x , in order to change y .

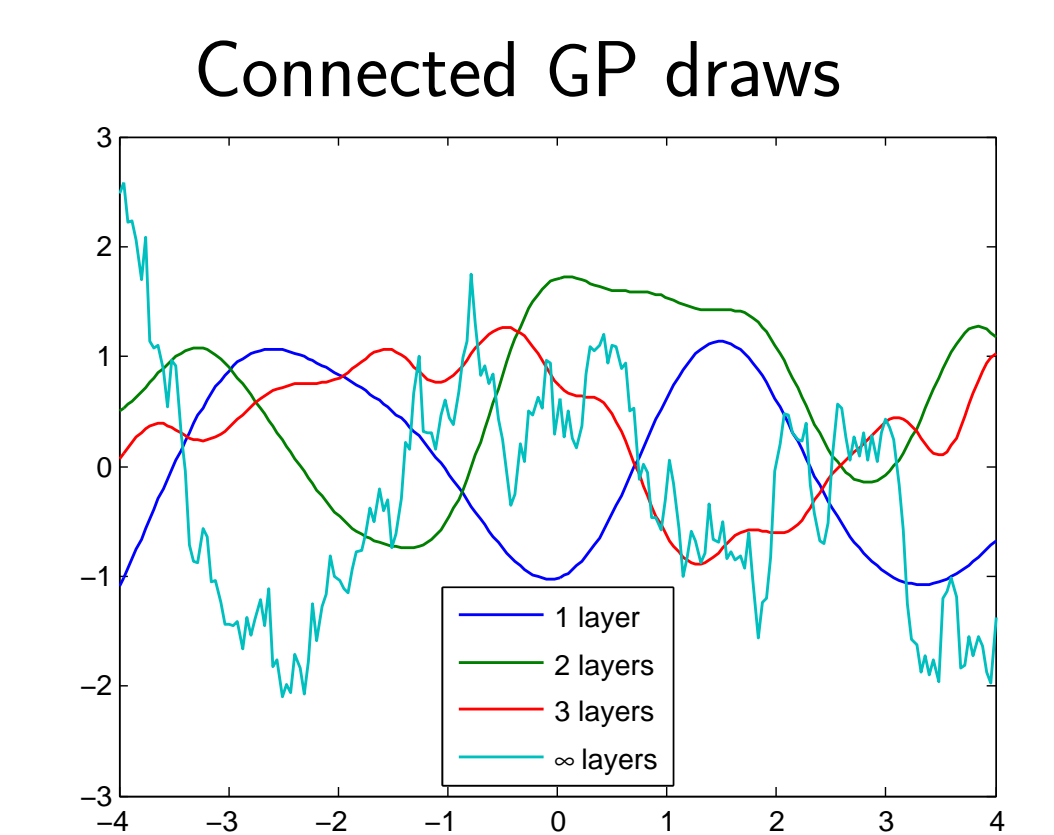
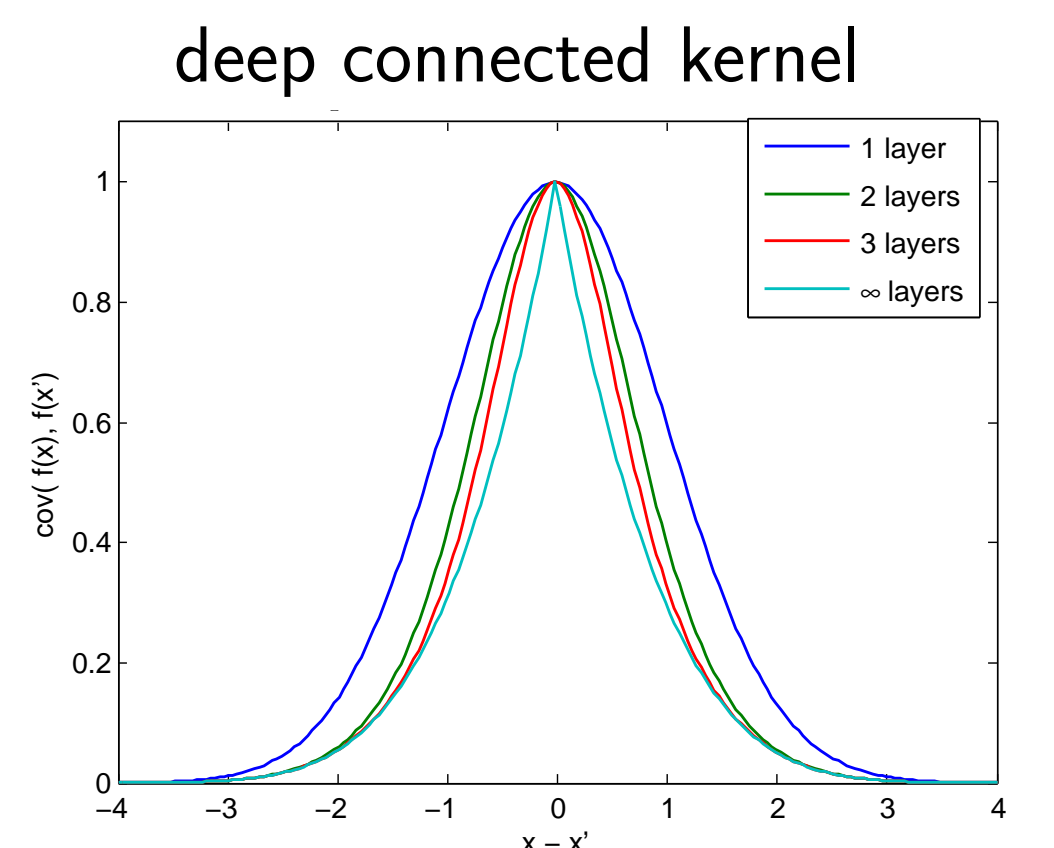


Infinitely Deep Kernels

- Can also analyze fixed deep feature mappings:
- (Cho, 2012) built kernels from multiple layers of feature mappings:
- If $k_1(x, x') = \Phi(x)^T \Phi(x')$, we can also build kernel $k_2(x, x') = k_2(\Phi(x), \Phi(x')) = \Phi(\Phi(x))^T \Phi(\Phi(x'))$.
- For the squared-exp kernel, this composition operation has a closed form:

$$\begin{aligned} k_{n+1}(x, x') &= \\ &= \exp\left(-\frac{1}{2} \left\| \begin{bmatrix} \Phi_n(x) \\ x \end{bmatrix} - \begin{bmatrix} \Phi_n(x') \\ x' \end{bmatrix} \right\|_2^2\right) \\ &= \exp\left(k_n(x, x') - 1 - \frac{1}{2} \|x - x'\|_2^2\right) \quad (1) \end{aligned}$$

- This kernel satisfies $k_\infty - \log(k_\infty) = 1 + \frac{1}{2} \|x - x'\|_2^2$
- No closed form, but it is continuous and differentiable everywhere except at $x = x'$.



Conclusions

- Random networks capture fewer degrees of freedom as they get deeper
- Connecting the input to each layer resolves this pathology
- Deep Gaussian processes are a data-independent way to characterize neural networks
- Deep One-Dimensional GPs have a log-normal distribution on the magnitude of their derivatives.
- Can build “deep net” kernels

Code at github.com/duvenaud/deep-limits

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

