Avoiding pathologies in very deep neural networks

David Duvenaud
University of Cambridge

Oren Rippel
M.I.T., Harvard University

Ryan Adams
Harvard University

Zoubin Ghahramani
University of Cambridge

Abstract

Choosing appropriate architectures and regularization strategies of deep networks is crucial to good predictive performance. To shed light on this problem, we analyze the analogous problem of constructing useful priors on compositions of functions. Specifically, we study deep Gaussian processes, a type of infinitely-wide, deep neural network. We show that in these architectures, the representational capacity of the network tends to capture fewer degrees of freedom as the number of layers increases, retaining only a single degree of freedom in the limit. We propose alternate network architectures which do not suffer from these pathologies. We also derive deep covariance functions, obtained by composing infinitely many feature transforms. Lastly, we characterize the models obtained by performing dropout on Gaussian processes.

1 Introduction

Much recent work on deep networks has focused on weight initialization (Martens, 2010), regularization (Lee *et al.*, 2007) and network architecture (Gens and Domingos, 2013). However, the interactions between these different design decisions can be complex and difficult to characterize. We propose to approach the design of deep architectures by examining the problem of assigning priors to deep functions. Well-defined priors allow us to explicitly examine the assumptions being made about functions we may wish to learn. Once we identify classes of priors that imbue our models with desirable properties, these in turn may suggest regularization, initialization, and architecture choices that delineate such properties.

Fundamentally, a multilayer neural network implements a composition of vector-valued functions, one per layer. Hence, understanding properties of such function compositions helps us gain insight into deep networks. In this paper, we examine a simple and flexible class of priors on compositions of functions, namely deep Gaussian processes (Damianou and Lawrence, 2013). Deep GPs are simply priors on compositions of vector-valued functions, where each output of each layer is drawn independently from a GP prior:

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

$$\mathbf{f}_d^{(\ell)} \overset{\text{ind}}{\sim} \mathcal{GP}\big(0, k_d^{\ell}(\mathbf{x}, \mathbf{x}')\big) \quad (2)$$

These models correspond to a certain type of infinitely-wide multi-layer perceptron (MLP), and as such make canonical candidates for generative models of functions that closely relate to neural networks.

By characterizing these models, this paper shows that representations based on repeated composition of independently-initialized functions exhibit a pathology where the representation becomes invariant to all but one direction of variation. This corresponds to an eventual debilitating decrease in the information capacity of networks as a function of their number of layers. However, we will demonstrate that a simple change in architecture — namely, connecting the input to each layer — fixes this problem.

We also present two related analyses: First, we examine the properties of a arbitrarily deep fixed feature transforms ("deep kernels"). Second, we characterise the prior obtained by performing dropout on GPs, showing equivalences to existing models.

2 Relating deep neural nets and deep Gaussian processes

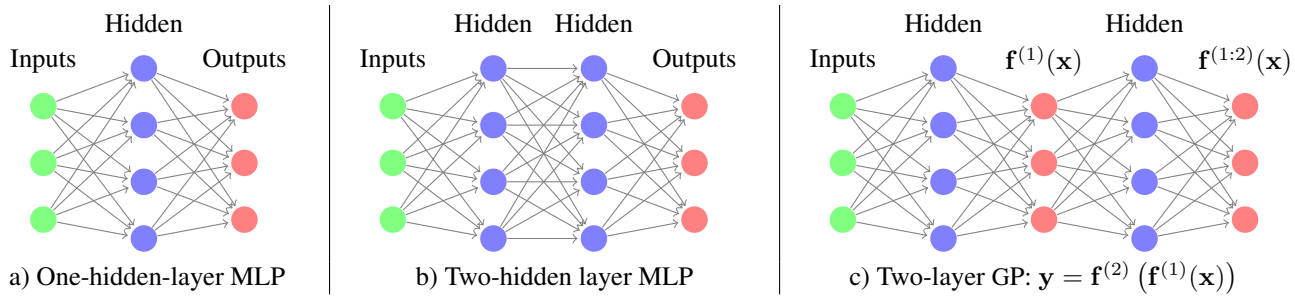
2.1 Neural nets with one hidden layer

In the typical definition of an MLP, the hidden units of the first layer are defined as:

$$\mathbf{h}(\mathbf{x}) = \mathbf{h}^{(1)}(\mathbf{x}) = \sigma\left(\mathbf{b}^{(1)} + \mathbf{W}^{(1)}\mathbf{x}\right) \quad (3)$$

where $\mathbf{h}$ are the hidden unit activations, $\mathbf{b}$ is a bias vector, $\mathbf{W}$ is a weight matrix and σ is a sigmoidal function applied

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.



a) One-hidden-layer MLP b) Two-hidden layer MLP c) Two-layer GP: $\mathbf{y} = \mathbf{f}^{(2)}\left(\mathbf{f}^{(1)}(\mathbf{x})\right)$

Figure 1: Comparing architectures. In the deep GP models, there are two possible meanings for the hidden units. We can consider every other layer to be a linear combination of an infinite number of parametric hidden units. Alternatively, we can integrate out the hidden layers, and consider the deep GP to be a neural network with a finite number of hidden units, each with a different non-parametric activation function.

element-wise. The output vector $f(\mathbf{x})$ is simply a weighted sum of these hidden unit activations:

$$f(\mathbf{x}) = \mathbf{V}^{(1)}\sigma\left(\mathbf{b}^{(1)} + \mathbf{W}^{(1)}\mathbf{x}\right) = \mathbf{V}^{(1)}\mathbf{h}^{(1)}(\mathbf{x}) \quad (4)$$

where $\mathbf{V}^{(1)}$ is another weight matrix.

There exists a correspondence between one-layer MLPs and GPs (Neal, 1995). GP priors can be viewed as a prior on neural networks with infinitely many hidden units. More precisely, for any model of the form

$$f(\mathbf{x}) = \frac{1}{K}\alpha^{\mathsf{T}}\mathbf{h}(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K}\alpha_i h_i(\mathbf{x}), \quad (5)$$

with fixed features $[h_1(\mathbf{x}), \ldots, h_K(\mathbf{x})]^{\mathsf{T}} = \mathbf{h}(\mathbf{x})$ and i.i.d. α's with zero mean and finite variance σ^2, the central limit theorem implies that as the number of features $K \to \infty$, any two function values $f(\mathbf{x}), f(\mathbf{x}')$ have a joint distribution approaching $\mathcal{N}\left(0, \frac{\sigma^2}{K}\sum_{i=1}^{K} h_i(\mathbf{x})h_i(\mathbf{x}')\right)$. A joint Gaussian distribution between any two function values is the definition of a Gaussian process. (!) DD: any two or any set?

The result is surprisingly general: It doesn't put any constraints on what the features are (other than having bounded activation), nor does it require that the feature weights α be Gaussian distributed.

We can also work backwards to derive a one-layer MLP from a GP: Mercer's theorem implies that any positive-definite kernel function corresponds to an inner product of features: $k(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^{\mathsf{T}}\mathbf{h}(\mathbf{x}')$. Thus in the one-hidden-layer case, the correspondence between MLPs and GPs is simple: The features $h(\mathbf{x})$ of the GP correspond to the hidden units of the MLP.

2.2 Multiple hidden layers

In a neural net with multiple hidden layers, the correspondence is a little more complicated. In an MLP, the n^{th} layer units are given by the recurrence:

$$\mathbf{h}^{(n)}(\mathbf{x}) = \sigma\left(\mathbf{b}^{(n)} + \mathbf{W}^{(n)}\mathbf{h}^{(n-1)}(\mathbf{x})\right) \quad (6)$$

This architecture is shown in figure 1b. In this model, each hidden layer's output feeds directly into the next layer's input, weighted by the corresponding element of $\mathbf{W}^{(n)}$.

However, in a deep GP, the D outputs $\mathbf{f}^{(n)}(\mathbf{x})$ in between each layer are weighted sums of the hidden units of the layer below, and the next layer's hidden units depend only on these D outputs. Thus deep GPs have an extra set of layers that a MLP doesn't have, shown in figure 1c.

There are two ways to directly relate deep GPs to MLPs. First, we can note that, if the hidden units in a deep GP implied by Mercer's theorem $h^{(n)}(\mathbf{x})$ depend only on a linear projection of their inputs, as in the sigmoidal activation function $h(\mathbf{x}) = \sigma\left(\mathbf{b}^{(n)} + \mathbf{W}^{(n)}f^{(n-1)}(\mathbf{x})\right)$, then we can simply substitute $\mathbf{f}^{(n-1)}(\mathbf{x}) = \mathbf{V}^{(n-1)}h^{(n-1)}(\mathbf{x})$ to recover $h(\mathbf{x}) = \sigma\left(\mathbf{b}^{(n)} + \mathbf{W}^{(n)}\mathbf{V}^{(n-1)}h^{(n-1)}(\mathbf{x})\right)$. Thus, we can ignore the intermediate outputs $\mathbf{f}^{(n)}(\mathbf{x})$, and exactly recover an MLP with activation functions given by Mercer's theorem, but with rank-D weight matrices between layers.

The second, more general way we can relate the two model classes is to integrate out all $\mathbf{V}^{(n)}$, and view deep GP models as a neural network with a finite number of nonparametric, GP-distributed basis functions, where the D outputs of $\mathbf{f}^{1:\ell}(\mathbf{x})$ represent the output of the hidden nodes at the ℓ^{th} layer. This second view lets us compare deep GP models to multilayer perceptrons more directly, examining the activations and shapes of the finite number of basis functions at each layer.

3 One-dimensional asymptotics

One way to understand properties of functions drawn from deep GPs and deep networks by looking at the distribution of the derivative of these functions. We first focus on the

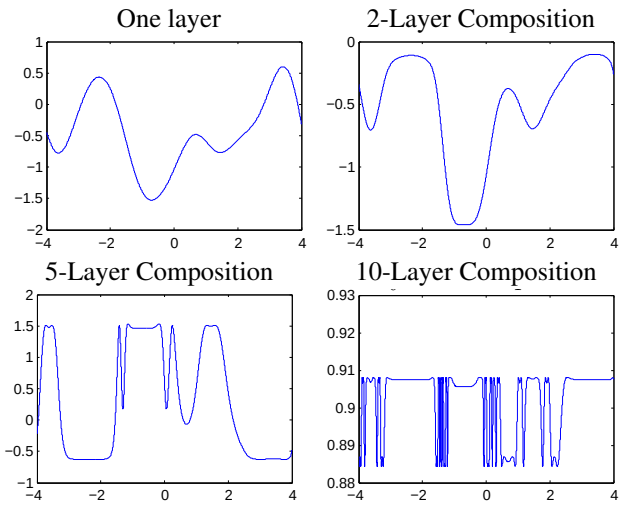


Figure 2: One-dimensional draws from a deep GP prior. After a few layers, the functions begin to be either nearly flat, or highly varying, everywhere. This is a consequence of the distribution on derivatives becoming heavy-tailed.

one-dimensional case. In this section, we derive the limiting distribution of the derivative of an arbitrarily deep, one-dimensional GP with a squared-exp kernel:

$$k_{\text{SE}}(x, x') = \sigma_f^2 \exp\left(\frac{-(x-x')^2}{2\ell^2}\right) \tag{7}$$

The hyperparameter σ_f^2 controls the variance of functions drawn from the prior, and the hyperparameter ℓ controls the smoothness. The derivative of a GP with a squared-exp kernel is distributed as $\mathcal{N}\left(0, \sigma_f^2/\ell^2\right)$. Intuitively, a GP is likely to have large derivatives if it has high variance and small lengthscales.

By the chain rule, the derivative of a one-dimensional deep GP is simply a product of its (independent) derivatives. The distribution of the absolute value of this derivative is a product of half-normals, each with mean $\sqrt{2\sigma_f^2/\pi\ell^2}$.

Thus, if we choose kernel parameters such that $\sigma_f^2/\ell_{d_1}^2 = \pi/2$, then $\mathbb{E}\left[\left|\partial f(x)/\partial x\right|\right] = 1$, and so $\mathbb{E}\left[\left|\partial f^{(1:L)}(x)/\partial x\right|\right] = 1$, that is to say, the expected magnitude of the derivative remains constants no matter the depth. If σ_f^2/ℓ^2 is less than $\pi/2$, the expected derivative magnitude goes to zero, and if greater, the expected magnitude goes to infinity as a function of L.

The log of the magnitude of the derivatives has moments:

$$m_{\log} = \mathbb{E}\left[\log\left|\frac{\partial f(x)}{\partial x}\right|\right] = 2\log\left(\frac{\sigma_f}{\ell}\right) - \log 2 - \gamma \tag{8}$$

$$v_{\log} = \mathbb{V}\left[\log\left|\frac{\partial f(x)}{\partial x}\right|\right] = \frac{\pi^2}{4} + \frac{\log^2 2}{2} - \gamma^2 - \gamma\log 4 + 2\log\left(\frac{\sigma_f}{\ell}\right)\left[\gamma + \log 2 - \log\left(\frac{\sigma_f}{\ell}\right)\right] \tag{9}$$

where $\gamma \approxeq 0.5772$ is Euler's constant. Since the second moment is finite, by the central limit theorem, the limiting distribution of the size of the gradient approaches log-normal as L grows:

$$\log\left|\frac{\partial f^{1:L}(x)}{\partial x}\right| = \sum_{i=1}^{L} \log\left|\frac{\partial f^i(x)}{\partial x}\right|$$

$$\implies \log\left|\frac{\partial f^{1:L}(x)}{\partial x}\right| \overset{L\to\infty}{\sim} \mathcal{N}\left(Lm_{\log}, L^2 v_{\log}\right) \tag{10}$$

Even if the expected magnitude of the derivative remains constant, the variance of the log-normal distribution grows without bound as the depth increases. Because the log-normal distribution is heavy-tailed, and its domain is bonded below by zero, the derivative will become very small almost everywhere, with rare but very large jumps.

Figure 2 shows this behavior in a draw from a 1D deep GP prior, at varying depths. This figure also shows that once the derivative in one region of the input space becomes very large or very small, it is likely to remain that way in subsequent layers.

4 The Jacobian of deep GP is a product of independent normal matrices

We now derive the distribution on Jacobians of multivariate functions drawn from a deep GP prior.

Lemma 4.1. *The partial derivatives of a function mapping $\mathbb{R}^D \to \mathbb{R}$ drawn from a GP prior with a product kernel are independently Gaussian distributed.*

Proof. Because differentiation is a linear operator, the derivatives of a function drawn from a GP prior are also jointly Gaussian distributed. The covariance between partial derivatives w.r.t. input dimensions d_1 and d_2 of vector $\mathbf{x}$ are given by Solak *et al.* (2003):

$$\text{cov}\left(\frac{\partial f(\mathbf{x})}{\partial x_{d_1}}, \frac{\partial f(\mathbf{x})}{\partial x_{d_2}}\right) = \frac{\partial^2 k(\mathbf{x}, \mathbf{x}')}{\partial x_{d_1} \partial x'_{d_2}}\bigg|_{\mathbf{x}=\mathbf{x}'} \tag{11}$$

If our kernel is a product over individual dimensions $k(\mathbf{x}, \mathbf{x}') = \prod_d^D k_d(x_d, x'_d)$, as in the case of the squared-exp kernel, then the off-diagonal entries are zero, implying that all elements are independent. □

In the case of the multivariate squared-exp kernel, the covariance between derivatives has the form:

$$f(\mathbf{x}) \sim \text{GP}\left(0, \sigma_f^2 \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(x_d - x'_d)^2}{\ell_d^2}\right)\right)$$

$$\implies \text{cov}\left(\frac{\partial f(\mathbf{x})}{\partial x_{d_1}}, \frac{\partial f(\mathbf{x})}{\partial x_{d_2}}\right) = \begin{cases} \frac{\sigma_f^2}{\ell_{d_1}^2} & \text{if } d_1 = d_2 \\ 0 & \text{if } d_1 \neq d_2 \end{cases} \tag{12}$$

Lemma 4.2. *The Jacobian of a set of D functions $\mathbb{R}^D \to \mathbb{R}$ drawn independetly from a GP prior with a product kernel is a $D \times D$ matrix of independent Gaussian R.V.'s*

Proof. The Jacobian of the vector-valued function $\mathbf{f}(\mathbf{x})$ is a matrix J with elements $J_{ij} = \frac{\partial \mathbf{f}_i(\mathbf{x})}{\partial x_j}$. Because we've assumed that the GPs on each output dimension $f_d(\mathbf{x}) \sim \mathcal{GP}$ are independent, it follows that each row of J is independent. Lemma 4.1 shows that the elements of each row are independent Gaussian. Thus all entries in the Jacobian of a GP-distributed transform are independent Gaussian R.V.s. □

Theorem 4.3. *The Jacobian of a deep GP with a product kernel is a product of independent Gaussian matrices, with each entry in each matrix being drawn independently.*

Proof. When composing L different functions, we'll denote the *immediate* Jacobian of the function mapping from layer $\ell-1$ to layer ℓ as $J^\ell(\mathbf{x})$, and the Jacobian of the entire composition of L functions by $J^{1:L}(\mathbf{x})$.

By the multivariate chain rule, the Jacobian of a composition of functions is simply the product of the Jacobian matrices of each function. Thus the Jacobian of the composed (deep) function $\mathbf{f}^{(L)}(\mathbf{f}^{(L-1)}(\dots\mathbf{f}^{(3)}(\mathbf{f}^{(2)}(\mathbf{f}^{(1)}(\mathbf{x})))\dots))$ is

$$J^{1:L}(x) = J^L J^{L-1} \dots J^3 J^2 J^1. \tag{13}$$

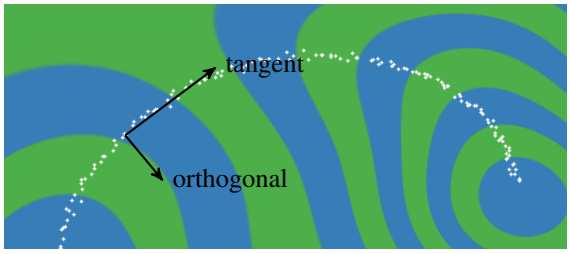
By lemma 4.2, each $J^\ell_{i,j} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \frac{\sigma_f^2}{\ell^2}\right)$, so the complete Jacobian is a product of independent Gaussian matrices, and each entry of those matrices is drawn independently. □

Theorem 4.3 allows us to analyze the representational properties of a deep Gaussian process by simply examining the properties of products of independent Gaussian matrices, a well-studied object.

5 Formalizing a pathology

Rifai *et al.* (2011b) 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. Figure 3 visualizes this idea. We follow Rifai *et al.* (2011a) in characterizing the representational properties of a function by the singular value spectrum of the Jacobian[1]. Figure 4 shows the spectrum for 5-dimensional deep GPs of different depths. As the net gets deeper, the largest singular value dominates, implying there is usually only one effective degree of freedom in representation being computed.

[1]Rifai *et al.* (2011a) examine the Jacobian at the training points, but the models we are examining are stationary, so it doesn't matter where we examine the function.



a) A noise-tolerant representation (blue & green) of a one-dimensional manifold (white)

Figure 3: Representing a 1-D manifold. Colors show the output of the computed representation as a function of the input space. The representation is invariant in directions orthogonal to the data manifold, making it robust to noise in those directions, and reducing the number of parameters needed to represent a datapoint. It also changes in directions tangent to the manifold, preserving information for later layers.

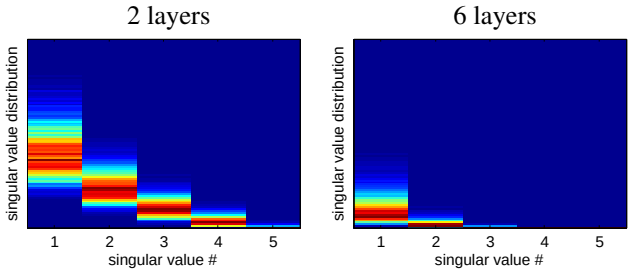


Figure 4: Normalized singular value spectrum of the Jacobian of a deep GP. As the net gets deeper, the largest singular value dominates. This implies that with high probability, there is only one effective degree of freedom in the representation being computed. As depth increases, the distribution on singular values also becomes heavy-tailed.

Figure 5 demonstrates a related pathology that arises when composing functions to produce a deep density model. The density in observed space eventually becomes locally concentrated onto one-dimensional manifolds, or *filaments*, implying that such models are insuitable to model manifolds of greater than one dimension.

To visualize this pathology in another way, figure 6 illustrates the value that at each point in the input space is mapped to after successive warpings. After 40 warpings, we can see that locally, there is usually only one direction that one can move in $\mathbf{x}$-space in order to change the value of the function.

To what extent are these pathologies present in nets being used today? In simulations, we found that for deep functions with a fixed latent dimension D, the singular value spectrum remained relatively flat for hundreds of layers as long as $D > 100$. Thus, these pathologies are unlikely to severely affect relatively shallow, wide networks.

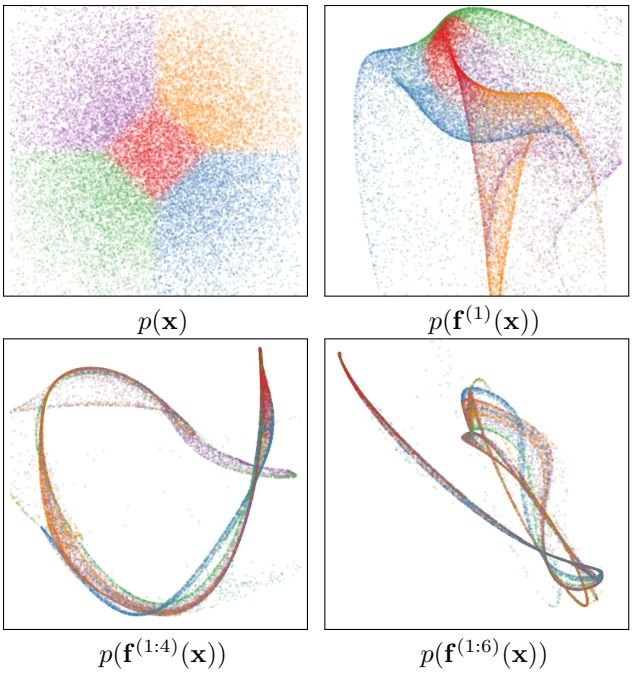


Figure 5: Draws from a deep GP. A distribution is warped by successive functions drawn from a GP prior. As the number of layers increases, the density concentrates along one-dimensional filaments.

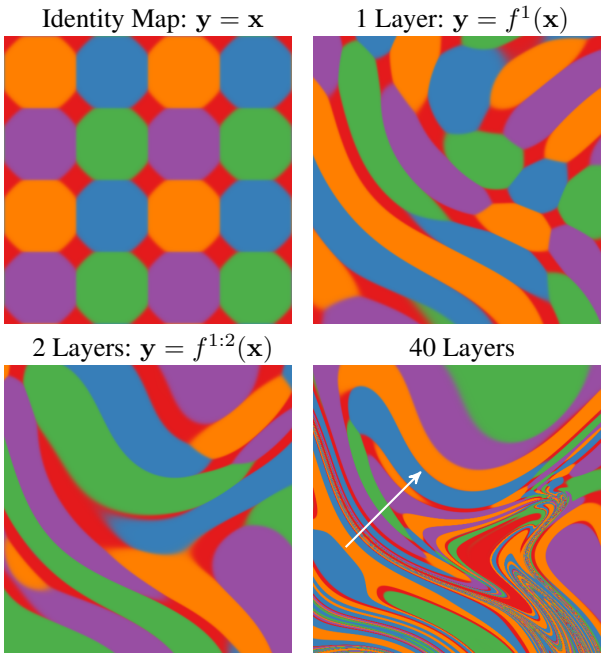


Figure 6: Feature Mapping of a deep GP. Colors correspond to the location $\mathbf{y} = \mathbf{f}(\mathbf{x})$ that each point is mapped to after being warped by a deep GP. Just as the densities in figure 5 became locally one-dimensional, there is usually only one direction that one can move $\mathbf{x}$ in locally to change $\mathbf{y}$. The number of directions in which the color changes rapidly corresponds to the number of large singular values in the Jacobian.

6 Fixing the pathology

Following a suggestion from Neal (1995), we can fix the pathologies exhibited in figures 5 and 6 by simply making each layer depend not only on the output of the previous layer, but also on the original input $\mathbf{x}$. We refer to these models as *input-connected networks*. Figure 7 shows a graphical representation of the two connectivity architectures. Similar connections between non-adjacent layers can also be found the primate visual cortex (Maunsell and van Essen, 1983). Formally,

$$\mathbf{f}^{(1:L)}(\mathbf{x}) = \mathbf{f}^{(L)}\left(\mathbf{f}^{(1:L-1)}(\mathbf{x}), \mathbf{x}\right), \quad \forall L \qquad (14)$$

Draws from the resulting prior are shown in figures 8, 9 and 11. The Jacobian of the composed, input-connected deep function is defined by the recurrence: $J^{1:L}(\mathbf{x}) = J^L \begin{bmatrix} J^{1:L-1} \\ I_D \end{bmatrix}$. Figure 10 shows that with this architecture, even 50-layer deep GPs have well-behaved singular value spectra.

7 Deep kernels

Bengio *et al.* (2006) showed that kernel machines have limited generalization ability when they use a local kernel such as the squared-exp. However, many interesting non-local kernels can be constructed which allow non-trivial extrapolation, for example, periodic kernels. Periodic kernels can be viewed as a 2-layer-deep kernel, in which the first layer maps $x \rightarrow [\sin(x), \cos(x)]$, and the second layer maps through a set of radial-basis functions.

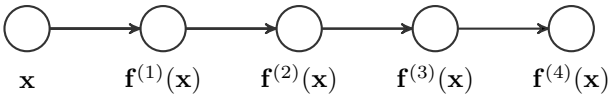
Can we construct useful kernels by composing fixed feature maps several times, creating deep kernels? Cho (2012) constructed kernels of this form, repeatedly applying multiple layers of feature mappings. Given a kernel $k_1(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x})^\mathsf{T}\mathbf{h}(\mathbf{x}')$, we can sometimes construct a kernel by composing the feature mapping:

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

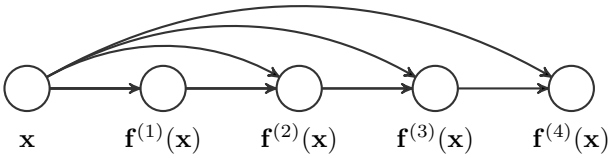
For the squared-exp kernel, this composition operation has a closed form:

$$\begin{aligned} k_{L+1}\left(\mathbf{x}, \mathbf{x}'\right) &= k_{SE}\left(\mathbf{h}(\mathbf{x}), \mathbf{h}(\mathbf{x}')\right) = \\ &= \exp\left(-\frac{1}{2}||\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x}')||_2^2\right) \\ &= \exp\left(-\frac{1}{2}\left[\mathbf{h}(\mathbf{x})^\mathsf{T}\mathbf{h}(\mathbf{x}) - 2\mathbf{h}(\mathbf{x})^\mathsf{T}\mathbf{h}(\mathbf{x}') + \mathbf{h}(\mathbf{x}')^\mathsf{T}\mathbf{h}(\mathbf{x}')\right]\right) \\ &= \exp\left(-\frac{1}{2}\left[k_L(\mathbf{x}, \mathbf{x}) - 2k_L(\mathbf{x}, \mathbf{x}') + k_L(\mathbf{x}', \mathbf{x}')\right]\right) \end{aligned}$$

Thus, we can express k_{L+1} exactly in terms of k_L.



a) The standard MLP connectivity architecture.



b) Input-connected architecture.

Figure 7: Two different architectures for deep neural networks. The standard architecture connects each layer's outputs to the next layer's inputs. The input-connected architecture connects also connects the original input $\mathbf{x}$ to each layer.

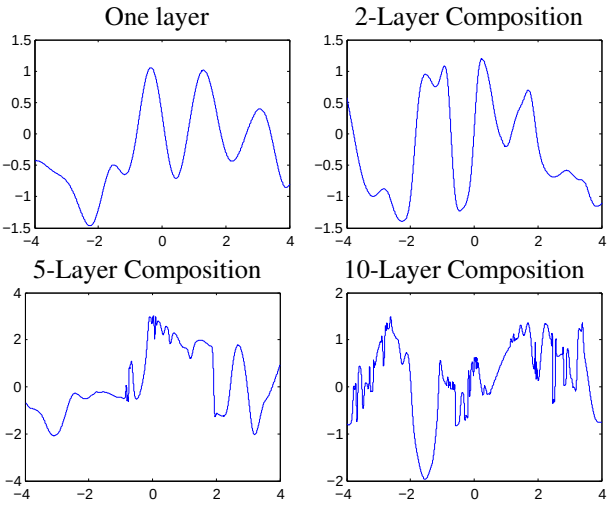


Figure 8: Draws from a 1D deep GP prior with each layer connected to the input. Even after many layers, the functions remain smooth in some regions, while varying rapidly in other regions.

Infinitely deep kernels What happens when repeat this feature mapping many times, starting with the squared-exp kernel? In the infinite limit, this recursion converges to $k(\mathbf{x}, \mathbf{x}') = 1$ for all pairs of inputs, which is simply a prior on constant functions $f(\mathbf{x}) = c$.

A non-degenerate construction As before, we can overcome this degeneracy by connecting the inputs $\mathbf{x}$ to each layer. To do so, we simply augment the feature vector

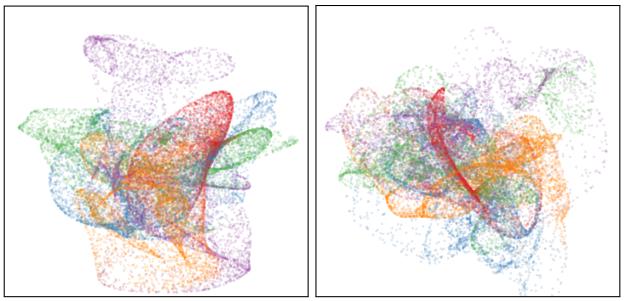


Figure 9: Left: Densities defined by a draw from a deep GP, with each layer connected to the input $\mathbf{x}$. As depth increases, the density becomes more complex without concentrating along filaments.

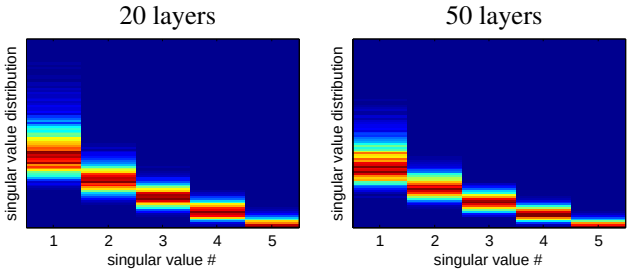


Figure 10: The distribution of singular values drawn from a 5-dimensional input-connected deep GP prior 25 and 50 layers deep. The singular values remain roughly the same scale as one another.

$\mathbf{h}_n(\mathbf{x})$ with $\mathbf{x}$ at each layer:

$$k_{L+1}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\left|\left|\begin{bmatrix}\mathbf{h}_L(\mathbf{x}) \\ \mathbf{x}\end{bmatrix} - \begin{bmatrix}\mathbf{h}_L(\mathbf{x}') \\ \mathbf{x}'\end{bmatrix}\right|\right|_2^2\right)$$
$$= \exp\Big(-\frac{1}{2}\big[k_L(\mathbf{x},\mathbf{x}) - 2k_L(\mathbf{x},\mathbf{x}') + k_L(\mathbf{x}',\mathbf{x}') - ||\mathbf{x}-\mathbf{x}'||_2^2\big]\Big) \quad (15)$$

For the SE kernel, this repeated mapping satisfies

$$k_\infty(\mathbf{x}, \mathbf{x}') - \log\left(k_\infty(\mathbf{x}, \mathbf{x}')\right) = 1 + \frac{1}{2}||\mathbf{x} - \mathbf{x}'||_2^2 \quad (16)$$

The solution to this recurrence has no closed form, but it has a similar shape to the Ornstein-Uhlenbeck covariance $k_{\text{OU}}(x, x') = \exp(-|x - x'|)$, with lighter tails. Samples from a GP prior with this kernel are not differentiable, and are locally fractal.

7.1 When are deep kernels useful models?

Kernels corresponding to fixed feature maps can compute rich structure, and can enable many types of generalization, such as translation and rotation invariance in images (Kondor, 2008). Salakhutdinov and Hinton (2008) used a

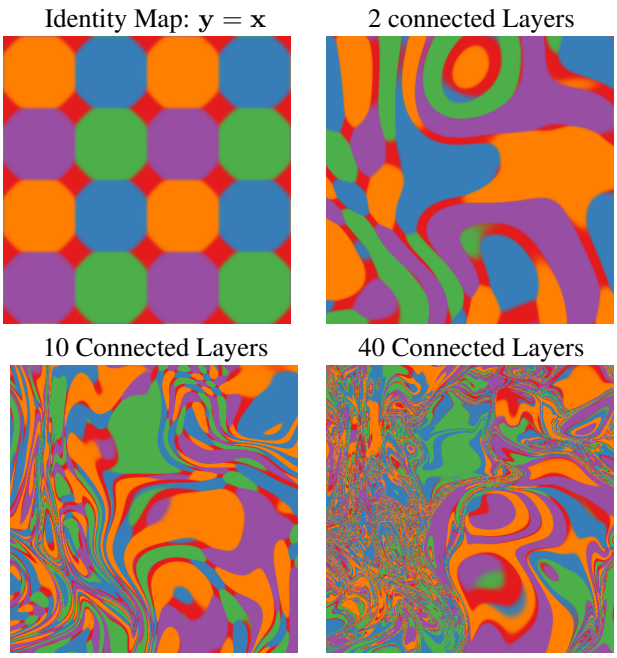
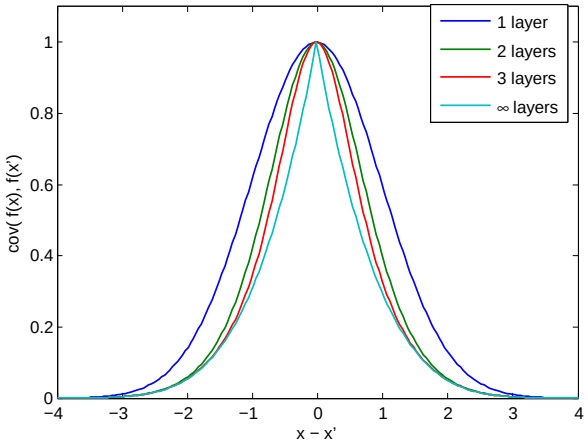


Figure 11: Feature Mapping of a deep GP with each layer connected to the input $\mathbf{x}$. Just as the densities in figure 9 remained locally two-dimensional even after many transformations, in this mapping there are usually two directions that one can move locally in $\mathbf{x}$ to change $\mathbf{y}$.



Connected transform kernel

Figure 12: A non-degenerate version of the infinitely deep feature transform kernel. By connecting the inputs $\mathbf{x}$ to each layer, the function can still depend on its input even after arbitrarily many layers of computation.

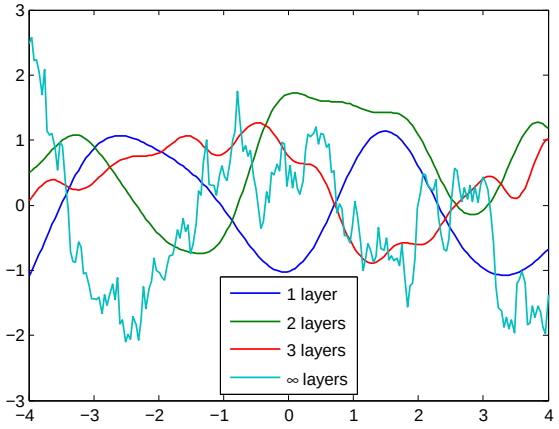


Figure 13: Draws from the deep input-connected kernel.

deep neural network to learn feature transforms for kernels, which learn invariances in an unsupervised manner. The relatively uninteresting properties of the kernels derived in this section simply reflect the fact that an arbitrary deep computation is not usually a useful representation, unless combined with learning.

8 Dropout in Gaussian processes

Dropout is a method for regularizing neural networks (Hinton *et al.*, 2012; Srivastava, 2013). Training with dropout entails randomly "dropping" (setting to zero) some proportion p of features or inputs, in order to improve the robustness of the resulting network by reducing co-dependence between neurons. To maintain similar overall activation levels, weights are multiplied by $1/p$ at test time. Alternatively, feature activations are multiplied by $1/p$ during training. At test time, the set of models defined by all possible ways of dropping-out neurons is averaged over, usually in an approximate way.

Baldi and Sadowski (2013) and Wang and Manning (2013) analyzed dropout in terms of the effective prior induced by this procedure in several models, such as linear and logistic regression. In this section, we examine the priors on functions that result from performing dropout in the one-layer neural network implicitly defined by a GP (equation (5)).

8.1 Dropout on feature activations

First, we examine the prior that results from randomly dropping features from $\mathbf{h}(x)$ with probability p. If these features have a weight distribution with finite moments $\mathbb{E}[\alpha_i] = \mu, \mathbb{V}[\alpha_i] = \sigma^2$, then the distribution of weights after each one has been dropped out with probability p is:

$$r_i \overset{\text{iid}}{\sim} \text{Ber}(p) \qquad \mathbb{E}[r_i\alpha_i] = p\mu, \mathbb{V}[r_i\alpha_i] = p^2\sigma^2 \quad (17)$$

However, after we increase the remaining activations to maintain the same expected activation, (by multiplying them by $1/p$), the resulting moments are again:

$$\mathbb{E}[{}^{p}/_{p} r_i\alpha_i] = \mu, \mathbb{V}[{}^{p}/_{p} r_i\alpha_i] = \sigma^2. \quad (18)$$

Thus, dropping out features of an infinitely-wide MLP does not change the model at all, since no individual feature can have more than infintesimal contribution to the activations.

8.2 Dropping out inputs

In a GP with a kernel $k(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^{D} k_d(\mathbf{x}_d, \mathbf{x}'_d)$, exact averaging over all possible ways of dropping out inputs with probability $1/2$ results in a mixture of GPs, each depending on only a subset of the inputs:

$$p\left(f(\mathbf{x})\right) = \frac{1}{2^D} \sum_{\mathbf{r}\in\{0,1\}^D} \mathrm{GP}\left(0, \prod_{d=1}^{D} k_d(\mathbf{x}_d, \mathbf{x}'_d)^{r_d}\right) \quad (19)$$

For an SE kernel, this model corresponds to a spike-and-slab prior on the inverse lengthscales (relevances).

This dropout mixture has the same covariance as

$$f(\mathbf{x}) \sim \mathrm{GP}\left(0, \frac{1}{2^D} \sum_{\mathbf{r}\in\{0,1\}^D} \prod_{d=1}^{D} k_d(\mathbf{x}_d, \mathbf{x}'_d)^{r_d}\right). \quad (20)$$

In GP models, a sum of covariance functions corresponds to a sum of functions. Therefore, (20) describes a sum of 2^D functions, each depending on a different subset of the inputs. This model class was studied by Duvenaud *et al.* (2011), who showed that exact inference in these models can be performed in $\mathcal{O}(N^2D^2) + \mathcal{O}(N^3)$.

9 Related work

Deep GPs were first proposed by Lawrence and Moore (2007). Variational inference in deep GPs was developed by Damianou and Lawrence (2013), who also analyzed the effect of automatic relevance determination in that model.

Adams *et al.* (2010) proposed a prior on deep Bayesian networks. Their architecture has no connections except between adjacent layers, and may also be expected to have similar pathologies as deep GPs as the number of layers increases. Deep Density Networks (Rippel and Adams, 2013) were constructed with invertibility in mind, with penalty terms encouraging the preservation of information about lower layers. Such priors are a promising approach to alleviating the pathology discussed in this paper.

Recurrent networks Bengio *et al.* (1994) and Pascanu *et al.* (2012) analyze a related problem with gradient-based learning in recurrent nets, the "exploding-gradients" problem. Hermans and Schrauwen (2012) analyze deep kernels corresponding to recurrent neural networks.

Analysis of deep learning Montavon *et al.* (2010) perform a layer-wise analysis of deep networks, and note that the performance of MLPs degrades as the number of layers with random weights increases. The importance of network architectures relative to learning has been examined by Saxe *et al.* (2011). Saxe *et al.* (2013) also looked at learning dynamics in deep linear models.

10 Conclusions

In this work, we established a number of propositions which help us gain insight into the properties of very deep models, and allow making informed choices regarding their architecture.

First, we identified a strong equivalence between deep GPs and MLPs — namely, that deep GPs can be written as MLPs with a finite number of nonparametric hidden units.

Second, we showed that deep GPs can be characterized using random matrix theory, which we applied to establish results regarding the distribution over the Jacobian of the composition transformation.

Next, we demonstrated that representations based on repeated composition of independently-initialized functions exhibit a pathology where the representations becomes invariant to all directions of variation, but one. This leads to extremely restricted expressiveness of such deep models in the limit of increasing number of layers.

Finally, we proposed a way to alleviate this problem: connecting the input to each layer of a deep representation allows us to construct priors on deep functions that do not exhibit the information-capacity pathology.

Acknowledgements

We thank Carl Rasmussen, Andrew McHutchon, Neil Lawrence, Andreas Damianou, James Lloyd, Creighton Heaukulani, Dan Roy and Mark van der Wilk for helpful discussions.

References

Adams, R. P., Wallach, H. M., and Ghahramani, Z. (2010). Learning the structure of deep sparse graphical models. In *International Conference on Artificial Intelligence and Statistics*, pages 1–8.

Baldi, P. and Sadowski, P. J. (2013). Understanding dropout. In *Advances in Neural Information Processing Systems*, pages 2814–2822.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, **5**(2), 157–166.

Bengio, Y., Delalleau, O., and Le Roux, N. (2006). The curse of highly variable functions for local kernel machines. pages 107–114.

Cho, Y. (2012). *Kernel methods for deep learning*. Ph.D. thesis, University of California, San Diego.

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.

Duvenaud, D., Nickisch, H., and Rasmussen, C. E. (2011). Additive Gaussian processes. In *Advances in Neural Information Processing Systems 24*, pages 226–234, Granada, Spain.

Gens, R. and Domingos, P. (2013). Learning the structure of sum-product networks. In *International Conference on Machine learning*.

Hermans, M. and Schrauwen, B. (2012). Recurrent kernel machines: Computing with infinite echo state networks. *Neural Computation*, **24**(1), 104–133.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Kondor, I. R. (2008). *Group theoretical methods in machine learning*. Ph.D. thesis, Columbia University.

Lawrence, N. D. and Moore, A. J. (2007). Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, pages 481–488. ACM.

Lee, H., Ekanadham, C., and Ng, A. (2007). Sparse deep belief net model for visual area v2. In *Advances in neural information processing systems*, pages 873–880.

Martens, J. (2010). Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, pages 735–742.

Maunsell, J. and van Essen, D. C. (1983). The connections of the middle temporal visual area (mt) and their relationship to a cortical hierarchy in the macaque monkey. *The Journal of neuroscience*, **3**(12), 2563–2586.

Montavon, G., Braun, D., and Müller, K.-R. (2010). Layer-wise analysis of deep networks with Gaussian kernels. *Advances in Neural Information Processing Systems*, **23**, 1678–1686.

Neal, R. M. (1995). *Bayesian learning for neural networks*. Ph.D. thesis, University of Toronto.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012). Understanding the exploding gradient problem. *arXiv preprint arXiv:1211.5063*.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011a). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, pages 833–840.

Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. (2011b). Higher order contractive auto-encoder. In *Machine Learning and Knowledge Discovery in Databases*, pages 645–660. Springer.

Rippel, O. and Adams, R. P. (2013). High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*.

Salakhutdinov, R. and Hinton, G. (2008). Using deep belief nets to learn covariance kernels for Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 20.

Saxe, A., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., and Ng, A. Y. (2011). On random weights and unsupervised feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1089–1096.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Dynamics of learning in deep linear neural networks.

Solak, E., Murray-Smith, R., Solak, E., Leithead, W., Rasmussen, C., and Leith, D. (2003). Derivative observations in Gaussian process models of dynamic systems.

Srivastava, N. (2013). *Improving neural networks with dropout*. Ph.D. thesis, University of Toronto.

Wang, S. and Manning, C. (2013). Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 118–126.