Гауссовские процессы как аппроксимация нейросетей для исследования устойчивости моделей

Ольга Гребенькова

Байесовское мультимоделирование

2 марта 2022 г.

О чем пойдет речь?

Гауссовские процессы (с прошлой пары)

Назовем гауссовским процессом $GP(\mathbf{m}(\mathbf{x}), \mathbf{k}(\mathbf{x}, \mathbf{x}')$ распределение на множестве функций, такое что для каждых $\mathbf{x}, \mathbf{x}'$: $GP(\mathbf{m}(\mathbf{x}), \mathbf{k}(\mathbf{x}, \mathbf{x}')$ — гауссовское распределение.

Пример глубокого гауссовского процесса

Композиция векторных функций, где каждая функция определна независимо из $f_d^{(l)} \sim GP(0, \mathbf{k}_d^{(l)}(x, x'))$:

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

Идея

Пример похож на определение глубокой нейронной сети. Давайте попробуем поисследовать гауссовские процессы, составляя из них архитектуры, похожие на известные нам нейронные сети.

- MLP с одним скрытым слоем $f(x) = W\sigma(b + Vx) = Wh(x)$, $h(x) = \sigma(b + Vx)$.
- Здесь h — скрытая активация, b — вектор смещения, V, W — матрицы весов и σ одномерная нелинейность.
- Для всякой модели в форме

$$\mathbf{f}(x) = \frac{1}{K}\mathbf{w}^\top \mathbf{h}(x) = \frac{1}{K}\sum_{i=1}^{K} w_i h_i(x)$$

с фиксированными функциями $[h_1(x), \ldots, h_K(x)]^\top = h(x)$ и i.i.d. w с нулевым средним и конечной дисперсией по центральной предельной теореме при большом K получаем, что для любых значений функции $f(x)$ и $f(x')$ имеют совместное распределение стремящееся к нормальному:

$$\lim_{K\to\infty} p\left(\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}') \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{\sigma^2}{K}\begin{bmatrix} \sum_{i=1}^{K} h_i(\mathbf{x})h_i(\mathbf{x}) & \sum_{i=1}^{K} h_i(\mathbf{x})h_i(\mathbf{x}') \\ \sum_{i=1}^{K} h_i(\mathbf{x}')h_i(\mathbf{x}) & \sum_{i=1}^{K} h_i(\mathbf{x}')h_i(\mathbf{x}') \end{bmatrix}\right)$$

- соответствие между MLP и GP прямое: неявные признаки h(x) ядра соответствуют скрытым слоям MLP.

Multiple hidden layer

Бесконечно широкий MLPs с несколькими скрытыми слоями:

- $h^{(l)}(x) = \sigma(b^{(l)} + V^{(l)} h^{(l-1)}(x))$.
- $f(x) = \frac{1}{K} w^\top h^{(2)} h^{(1)}(x)$.
- Если функции $h^{(2)}(x) h^{(1)}(x)$ фиксированы и нам неизвестны только веса последнего слоя w, эта модель соответствует неглубокому ГП с глубоким ядром, задаваемой формулой:

$$k(x, x') = [h^{(2)}(h^{(1)}(x))]^\top [h^{(2)}(h^{(1)}(x'))]$$

Пример

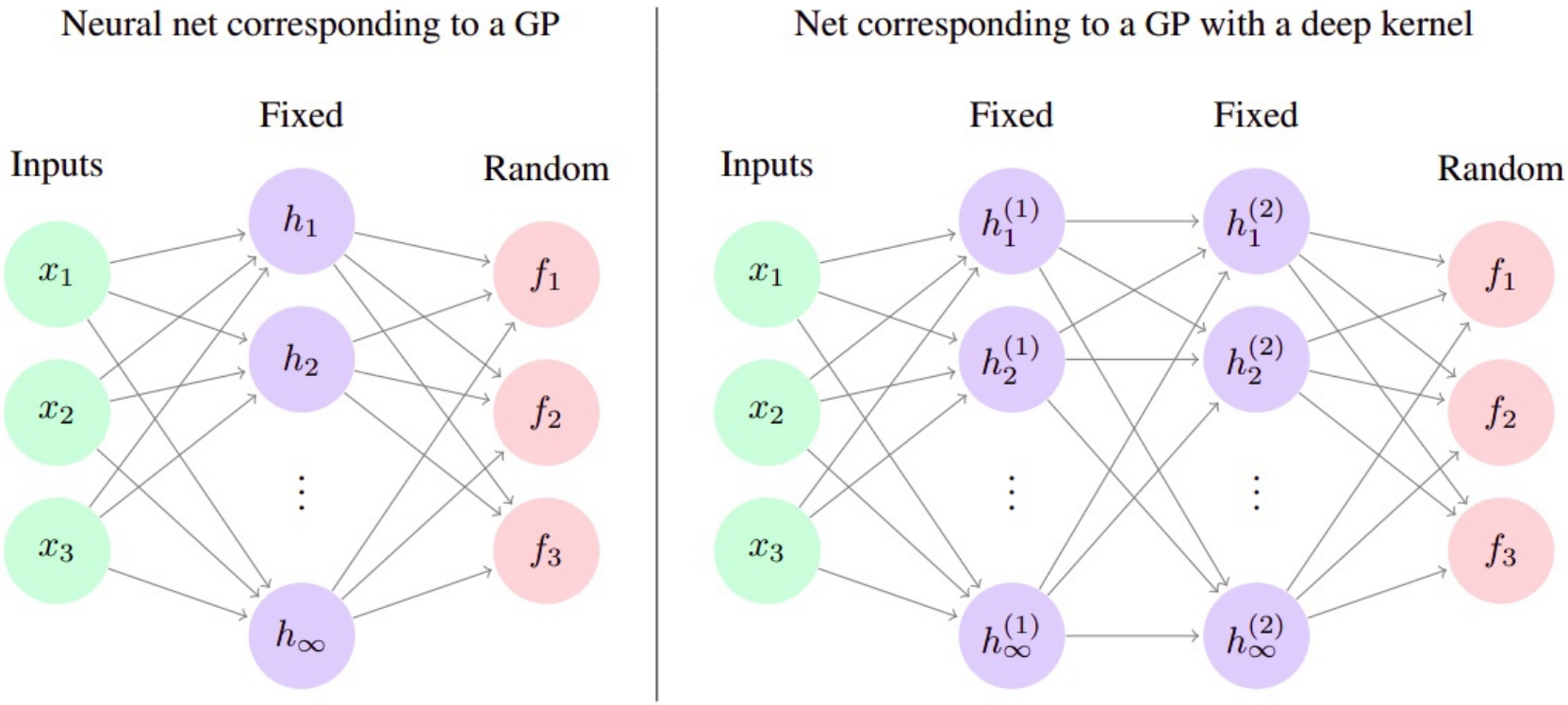
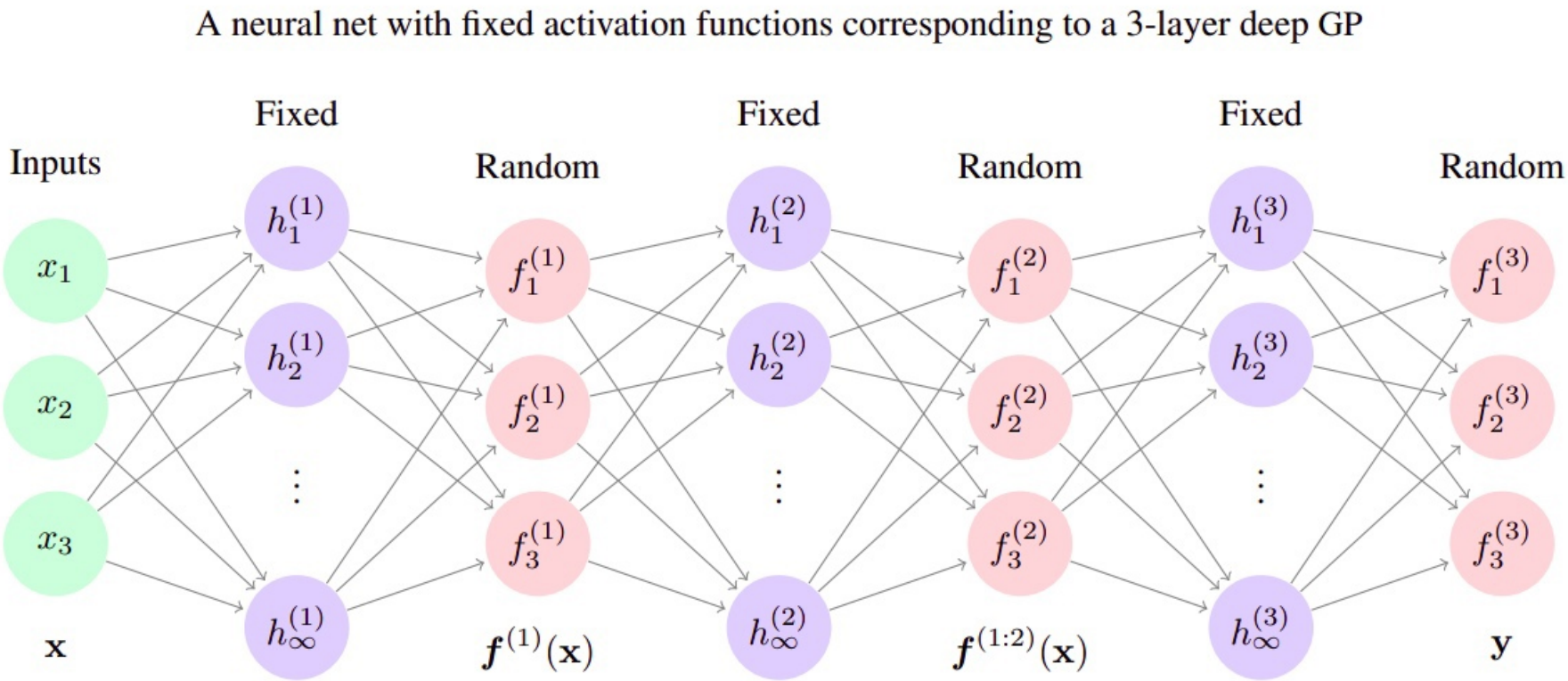


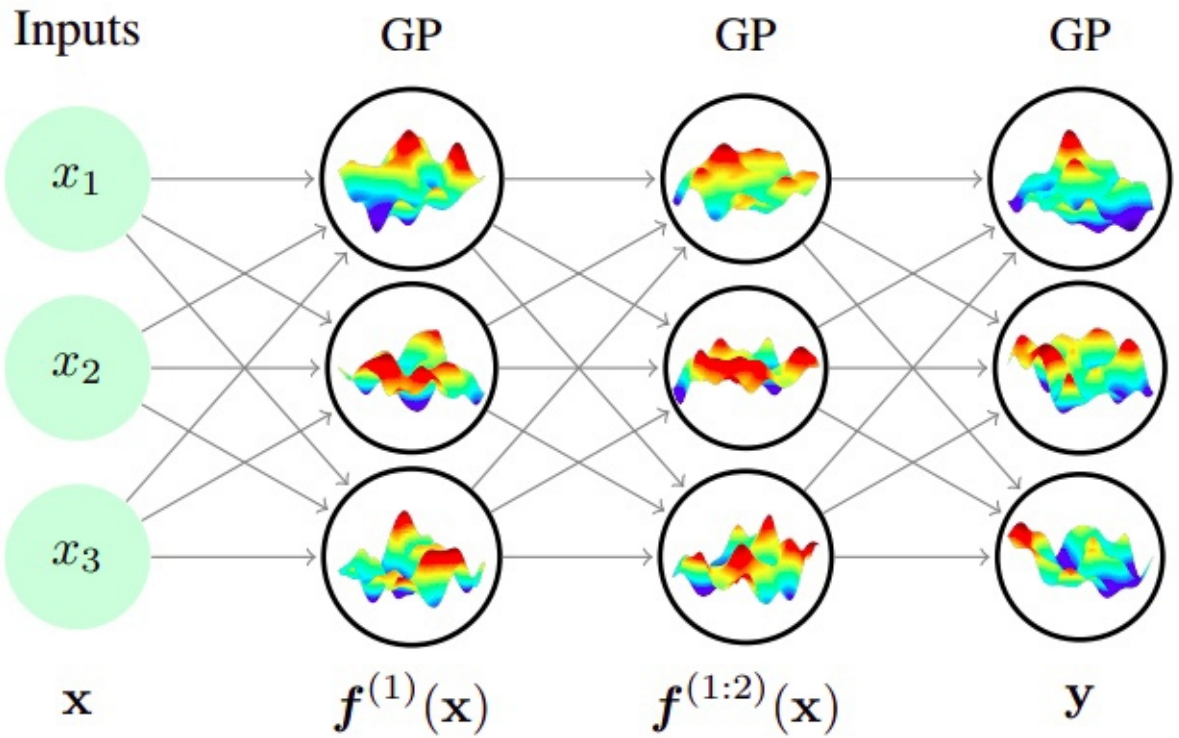
Figure 1: *Left:* GPs can be derived as a one-hidden-layer MLP with infinitely many fixed hidden units having unknown weights. *Right:* Multiple layers of fixed hidden units gives rise to a GP with a deep kernel, but not a deep GP.

Пример



Нейронная сеть, где каждый слой представляет собой взвешенную сумму бесконечного числа фиксированных скрытых нейронов, веса которых изначально неизвестны.

A net with nonparametric activation functions corresponding to a 3-layer deep GP



Нейронная сеть с конечным числом скрытых нейронов, каждый со своим неизвестным непараметрической функцией активации. Функции активации визуализируются с помощью рисунков из 2-мерного гауссовского процесса, хотя их входная размерность фактически будет такая же, как выходная предыдущего слоя.

Хороший манифолд

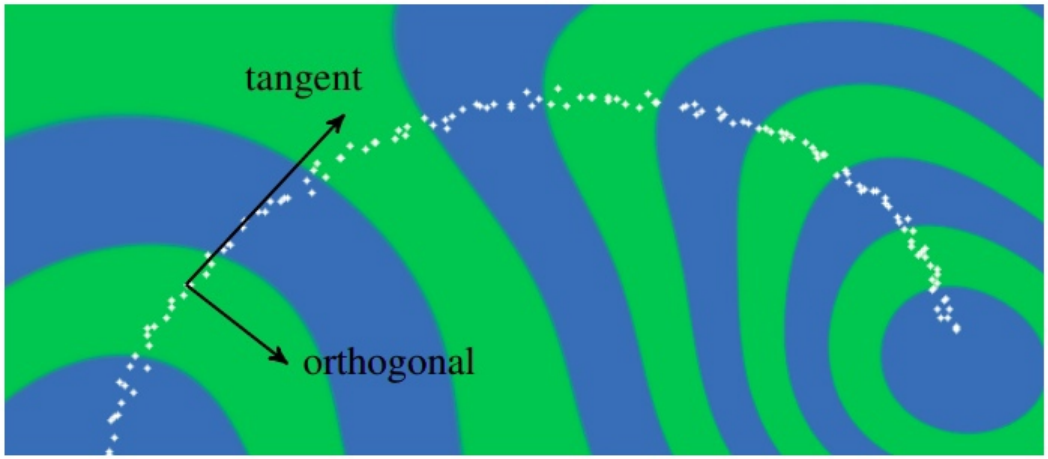


Figure 4: Representing a 1-D data manifold. Colors are a function of the computed representation of the input space. The representation (blue & green) changes little in directions orthogonal to the manifold (white), making it robust to noise in those directions. The representation also varies in directions tangent to the data manifold, preserving information for later layers.

Как оценить то, что мы построили?

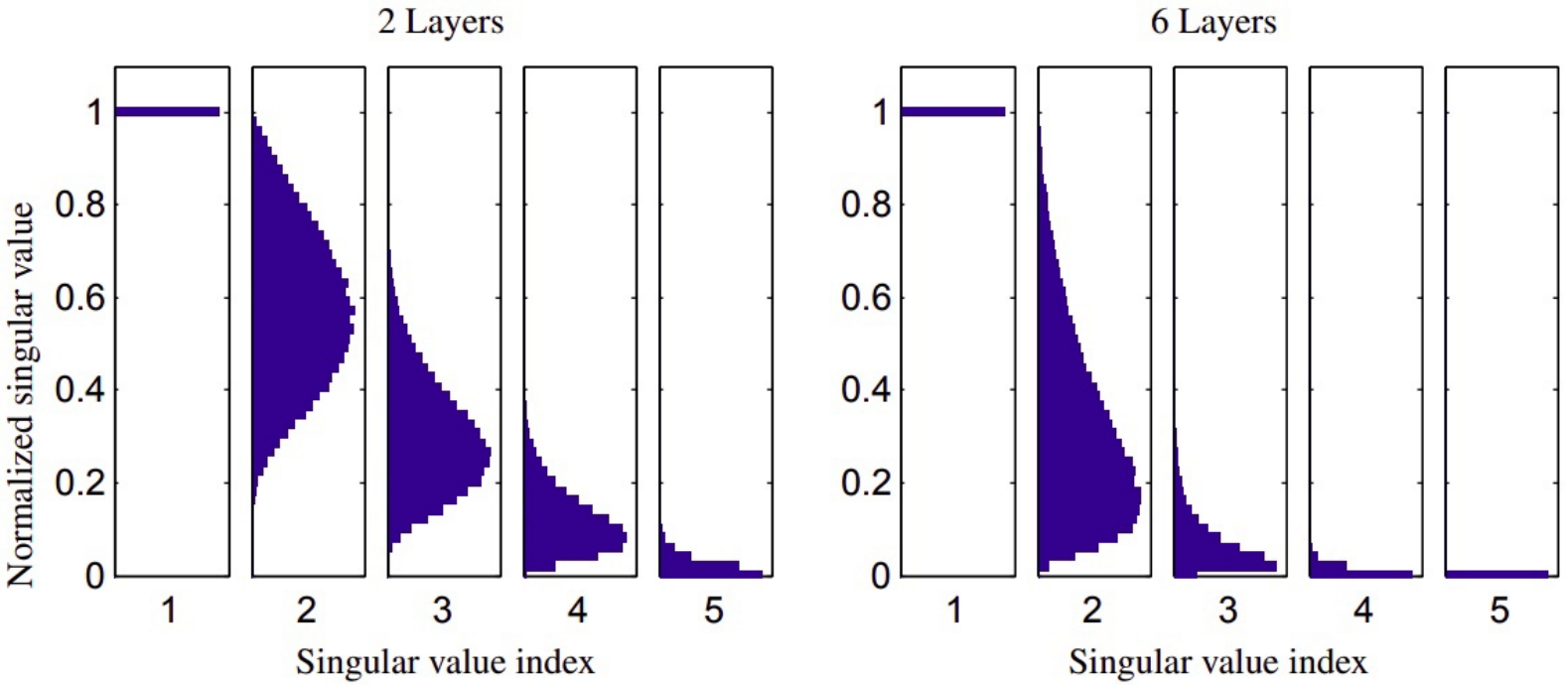
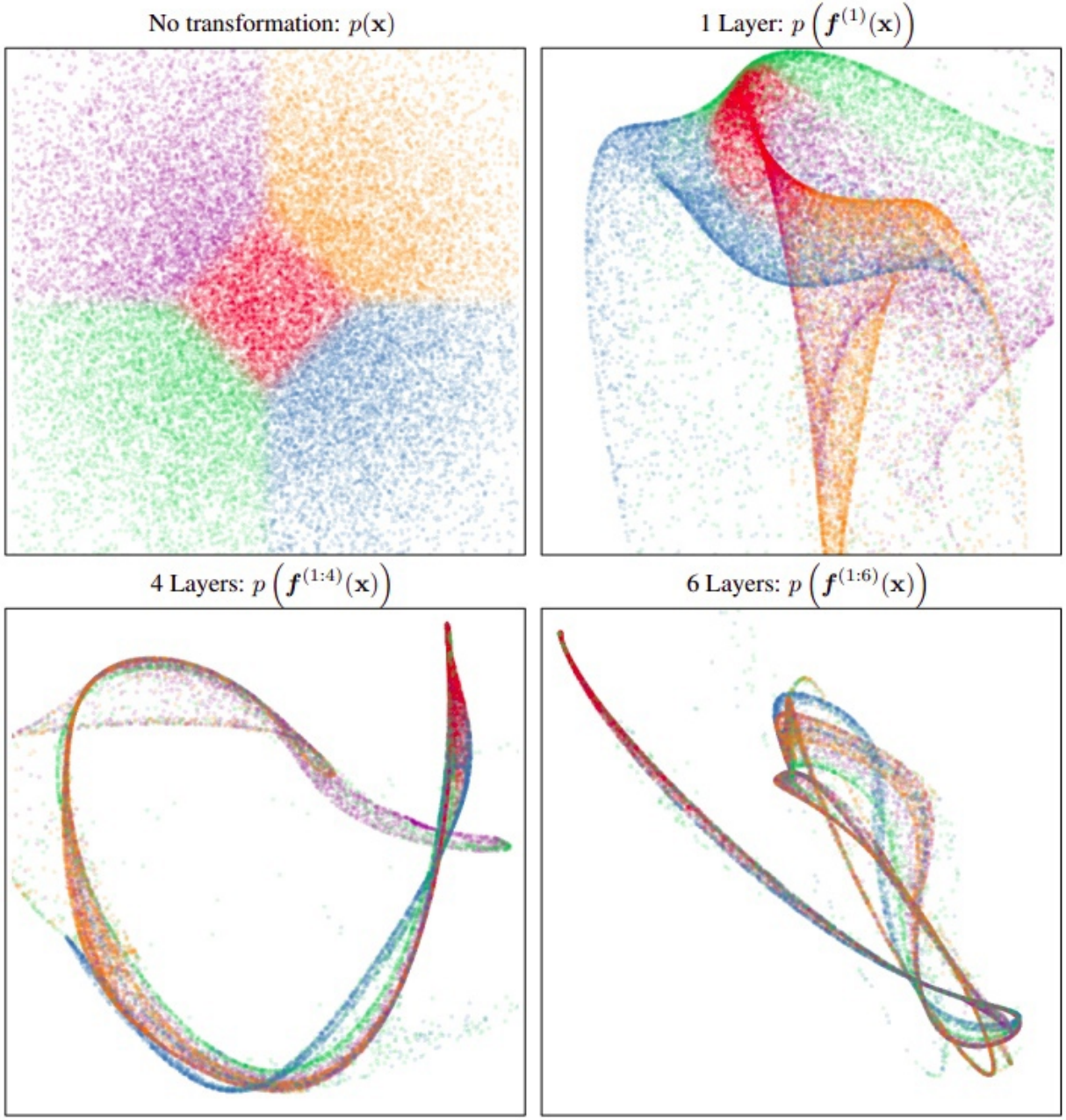


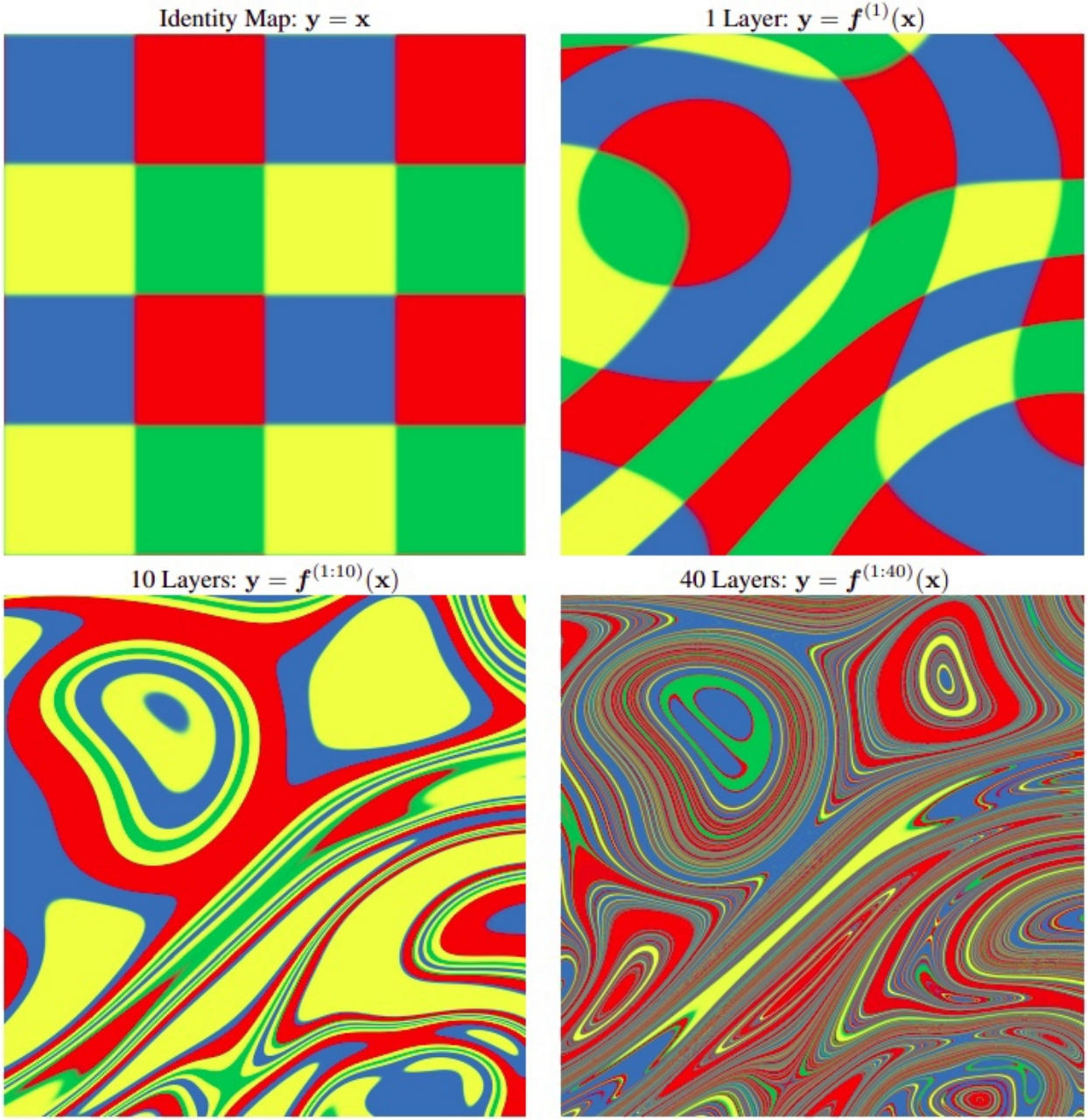
Figure 5: The distribution of normalized singular values of the Jacobian of a function drawn from a 5-dimensional deep GP prior 2 layers deep (*Left*) and 6 layers deep (*Right*). As nets get deeper, the largest singular value tends to become much larger than the others. This implies that with high probability, these functions vary little in all directions but one, making them unsuitable for computing representations of manifolds of more than one dimension.

Будем характеризовать репрезентативные свойства функции сингулярными значениями якобиана. Количество относительно больших сингулярных значений якобиана указывает на число направлений в пространстве данных, в которых представление значительно различается.

Пример распределения

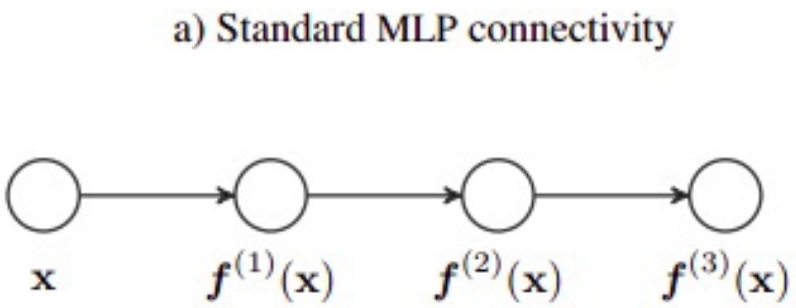


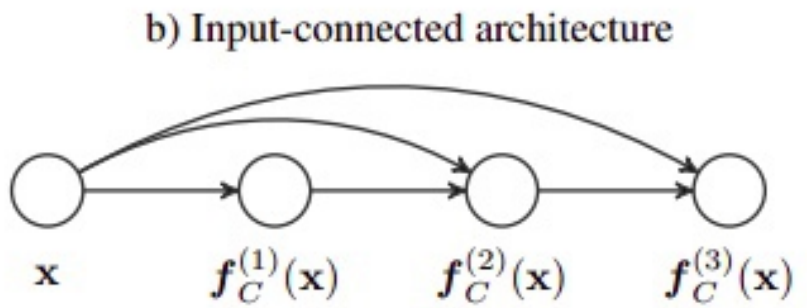
Пример распределения



Что делать?

- мы можем исправить патологии, показанные на слайдах просто делая каждый слой зависимым не только от выхода предыдущего слоя, но и от изначального входа. Назовем эти модели сетями со входными связями и обозначим глубокие функции, имеющие такую архитектуру индексом C, как в $f_C(x)$.
- $f_C^{(1:L)}(x) = f^{(L)}(f_C^{(1:L-1)}(x), x) \quad \forall L$





Пример распределения

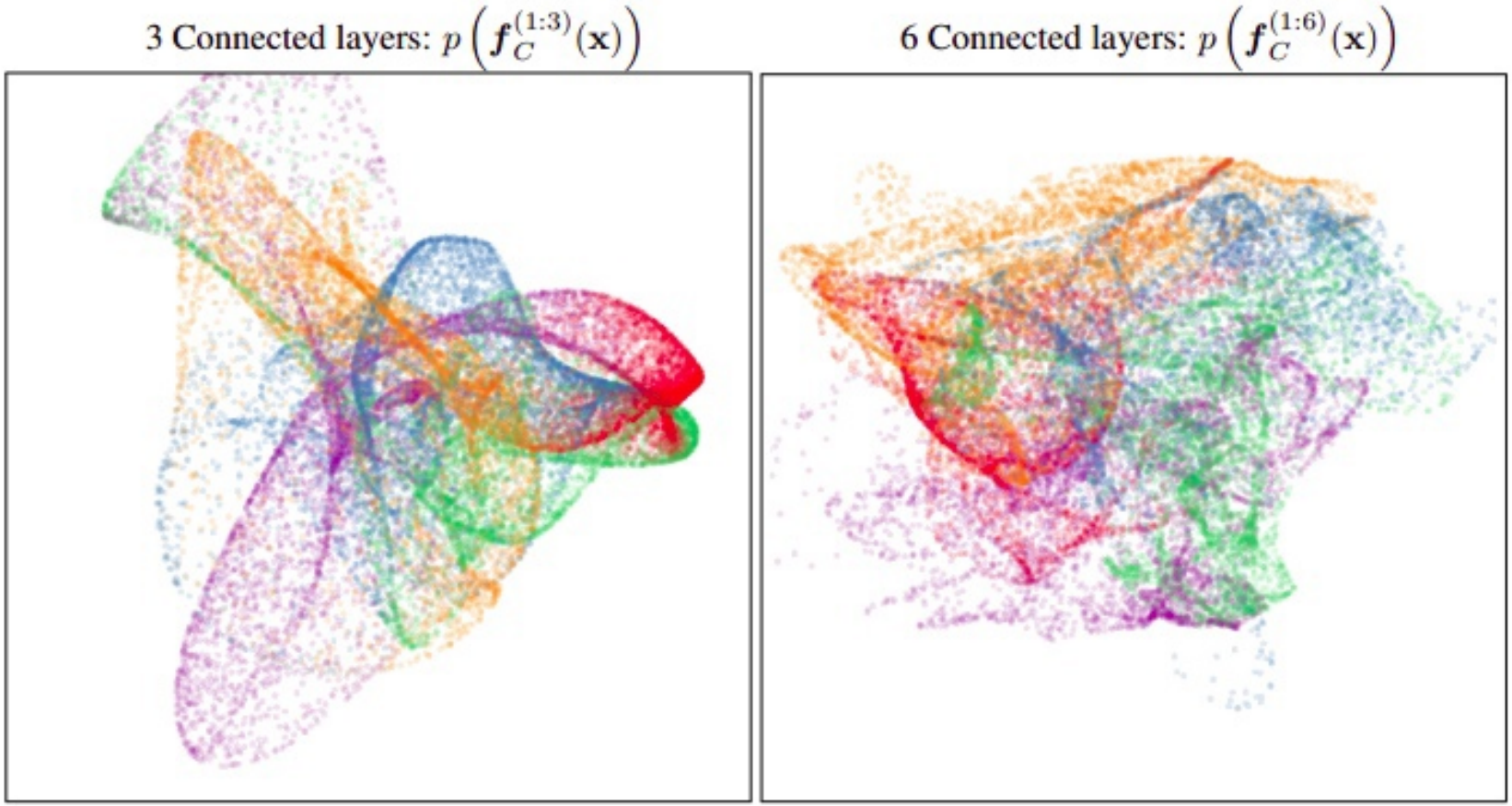


Figure 10: Points warped by a draw from a deep GP with each layer connected to the input $\mathbf{x}$. As depth increases, the density becomes more complex without concentrating only along one-dimensional filaments.

Пример распределения

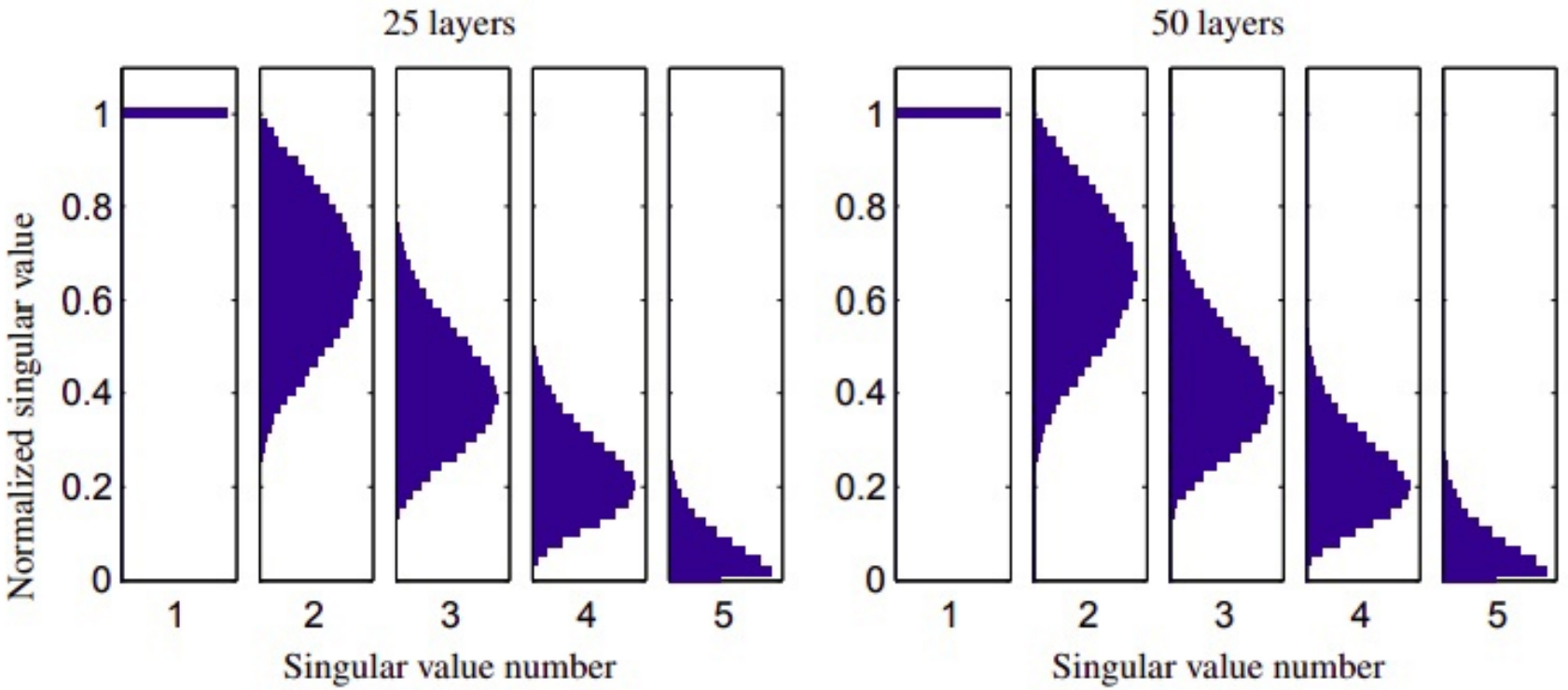
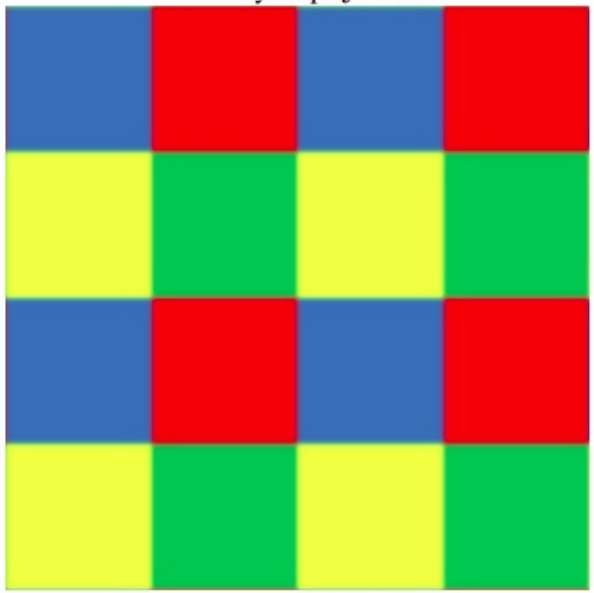


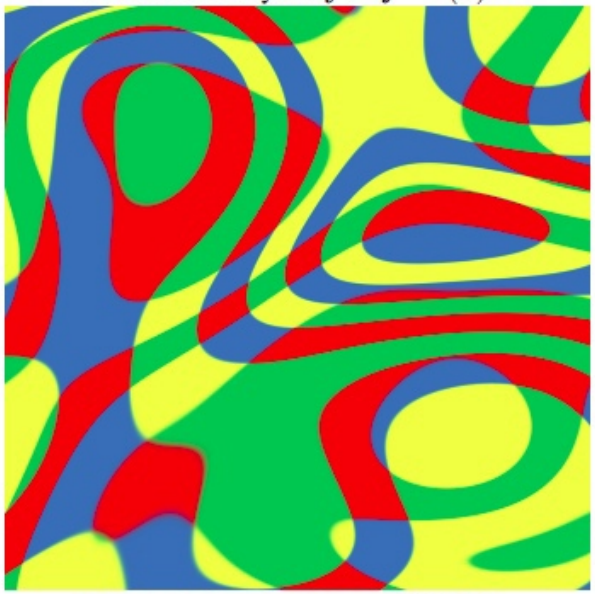
Figure 11: The distribution of singular values drawn from 5-dimensional input-connected deep GP priors, 25 layers deep (*Left*) and 50 layers deep (*Right*). Compared to the standard architecture, the singular values are more likely to remain the same size as one another, meaning that the model outputs are more often sensitive to several directions of variation in the input.

Пример распределения

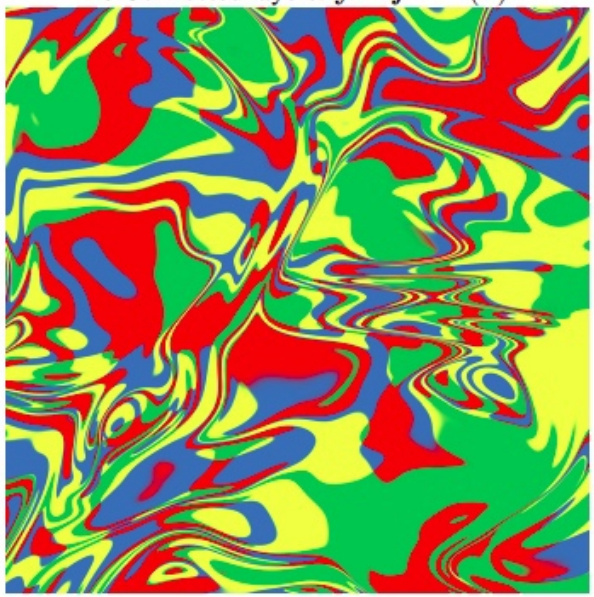
Identity map: $\mathbf{y} = \mathbf{x}$



2 Connected layers: $\mathbf{y} = \boldsymbol{f}^{(1:2)}(\mathbf{x})$



10 Connected layers: $\mathbf{y} = \boldsymbol{f}^{(1:10)}(\mathbf{x})$



20 Connected layers: $\mathbf{y} = \boldsymbol{f}^{(1:20)}(\mathbf{x})$