Constructive universal high-dimensional distribution generation through deep ReLU networks

Dmytro Perekrestenko

ETH zürich

July 2020

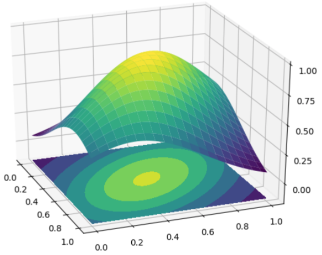
joint work with Stephan Müller and Helmut Bölcskei

Motivation

- Deep neural networks are widely used as generative models for complex data as images and natural language.
- Many generative network architectures are based on the transformation of low-dimensional distributions to high-dimensional ones, e.g., Variational Autoencoder, Wasserstein Autoencoder, etc.
- This talk answers the question of whether there exists a fundamental limitation in going from low dimension to a higher one.

Our contribution

$U[0,1]$ ➡ $p_{X,Y}$



This talk will show that there is no such limitation.

Generation of multi-dimensional distributions from $U[0,1]$

- Classical approaches - transforming distributions of the **same dimension**, e.g., the Box-Muller method [Box and Muller, 1958].
- [Bailey and Telgarsky, 2018] show that deep ReLU networks can transport $U[0,1]$ **to** $U[0,1]^d$.

Neural networks

A map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by

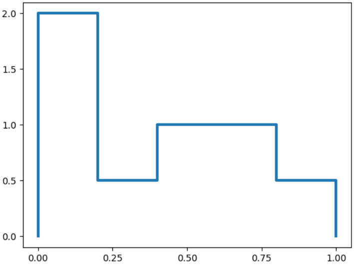
$$\Phi := W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1$$

is called a **neural network (NN)**.

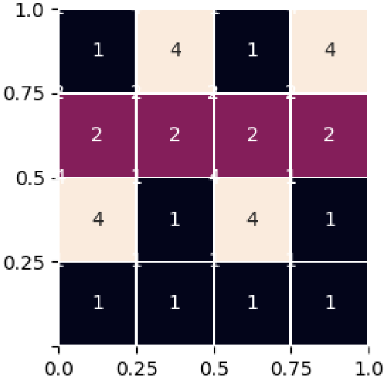
- Affine maps: $W_\ell = A_\ell x + b_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $\ell \in \{1, 2, \ldots, L\}$
- Non-linearity or activation function: ρ acts component-wise
- Network connectivity: $\mathcal{M}(\Phi)$ – total number of non-zero parameters in W_ℓ
- Depth of network or number of layers: $\mathcal{L}(\Phi) := L$

We denote by $\mathcal{N}_{d,d'}$ the set of all ReLU networks with input dimension $N_0 = d$ and output dimension $N_L = d'$.

Histogram distributions



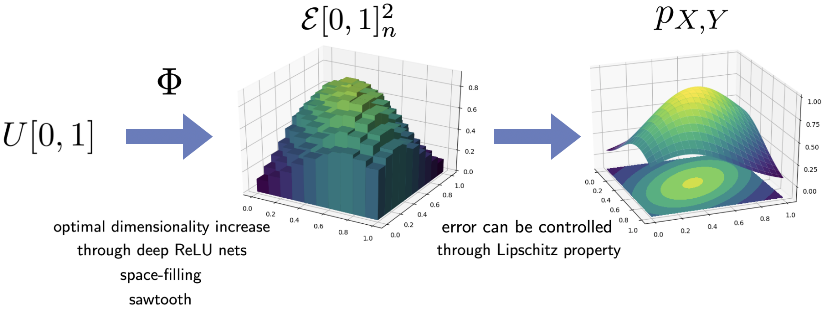
Histogram distribution $\mathcal{E}[0,1]^1_n$,
$d = 1, n = 5$.



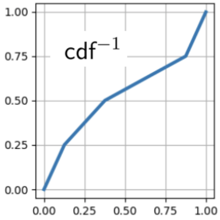
Histogram distribution $\mathcal{E}[0,1]^2_n$,
$d = 2, n = 4$.

Our goal

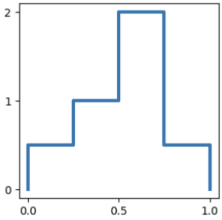
Transport $U[0,1]$ to an approximation of any given distribution supported on $[0,1]^d$. For illustration purposes we look at $d = 2$.



ReLU networks and histograms



$$\#U[0,1] =$$



Takeaway message

For any histogram distribution there exists a ReLU net that generates it from a uniform input. This net realizes an inverse cumulative distribution function (cdf^{-1}).

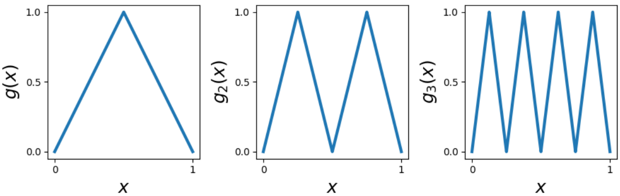
The key ingredient to dimension increase

Sawtooth function $g : [0,1] \to [0,1]$,

$$g(x) = \begin{cases} 2x, & \text{if } x < \frac{1}{2}, \\ 2(1-x), & \text{if } x \geq \frac{1}{2}, \end{cases}$$

let $g_1(x) = g(x)$, and define the "sawtooth" function of order s as the s-fold composition of g with itself according to

$$g_s := \underbrace{g \circ g \circ \cdots \circ g}_{s}, \quad s \geq 2.$$



NN realize sawtooth as $g(x) = 2\rho(x) - 4\rho(x - 1/2) + 2\rho(x-1)$.

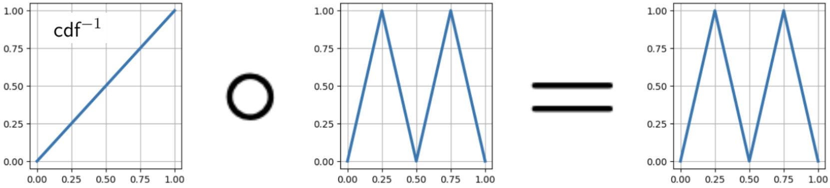
Related work

Theorem ([Bailey and Telgarsky, 2018, Th. 2.1], case $d = 2$)

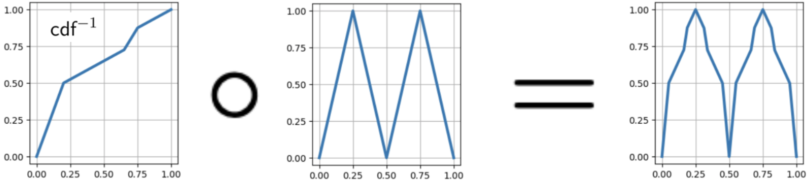
There exists a ReLU network $\Phi : x \to (x, g_s(x)), \Phi \in \mathcal{N}_{1,d}$ *with connectivity* $\mathcal{M}(\Phi) \leq Cs$ *for some constant* $C > 0$, *and of depth* $\mathcal{L}(\Phi) \leq s + 1$, *such that*

$$W(\Phi \# U[0,1], U[0,1]^2) \leq \frac{\sqrt{2}}{2^s}.$$

Main proof idea - space-filling property of sawtooth function.

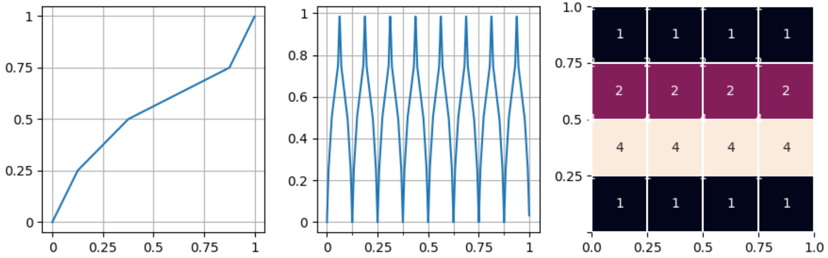


Generalization of the space-filling property



Approximating 2D distributions

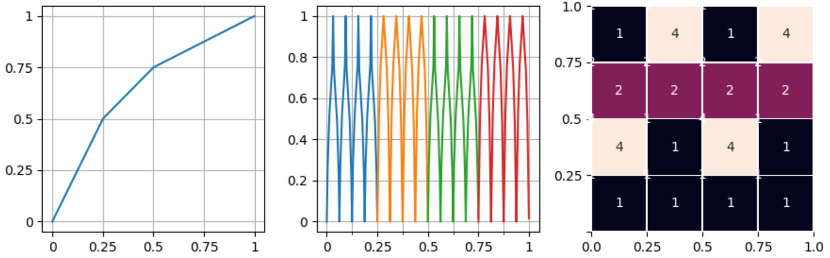
$$M : x \to (x, f(g_s(x)))$$



Generating a histogram distribution via the transport map $(x, f(g_s(x)))$. Left—the function $f(x)$, center—$f(g_4(x))$, right—a heatmap of the resulting histogram distribution.

Approximating 2D distributions con't

$$M : x \to \left(f_{\text{marg}}(x), \sum_{i=0}^{n-1} f_i(g_s(nf_{\text{marg}}(x) - i)) \right)$$



Generating a general 2-D histogram distribution. Left—the function $f_1 = f_3$, center—$\sum_{i=0}^{3} f_i\Big(g_3\Big(4x - i)\Big)\Big)$, right—a heatmap of the resulting histogram distribution. The function $f_0 = f_2$ is depicted on the left in Figure 3.

Generating histogram distributions with NNs

Theorem

For every distribution $p_{X,Y}(x,y)$ *in* $\mathcal{E}[0,1]^2_n$*, there exists a* $\Psi \in \mathcal{N}_{1,2}$ *with connectivity* $\mathcal{M}(\Psi) \leq C_1 n^2 + C_2 ns$*, for some constants* $C_1, C_2 > 0$*, and of depth* $\mathcal{L}(\Psi) \leq s+3$*, such that*

$$W(\Phi \# U[0,1], p_{X,Y}) \leq \frac{2\sqrt{2}}{n 2^s}.$$

- Error decays exponentially with depth and linearly in n
- Connectivity is in $\mathcal{O}(n^2)$ which is of the same order as the number of $\mathcal{E}[0,1]^2_n$'s parameters ($n^2 - 1$).
- Special case $n = 1$ coincides with [Bailey and Telgarsky, 2018, Th. 2.1].

Histogram approximation

Theorem

Let $p_{X,Y}$ be a 2-dimensional Lipschitz-continuous pdf of finite differential entropy on its support $[0,1]^2$. Then, for every $n > 0$, there exists a $\tilde{p}_{X,Y} \in \mathcal{E}[0,1]^2_n$ such that

$$W(p_{X,Y}, \tilde{p}_{X,Y}) \leq \frac{1}{2} \|p_{X,Y} - \tilde{p}_{X,Y}\|_{L_1([0,1]^2)} \leq \frac{L\sqrt{2}}{2n}.$$

Universal approximation

Theorem

Let $p_{X,Y}$ be an L-Lipschitz continuous pdf supported on $[0,1]^2$. Then, for every $n > 0$, there exists a $\Phi \in \mathcal{N}_{1,2}$ with connectivity $\mathcal{M}(\Phi) \leq C_1 n^2 + C_2 ns$ for some constants $C_1, C_2 > 0$, and of depth $\mathcal{L}(\Phi) \leq s + 3$, such that

$$W(\Phi \# U[0,1], p_{X,Y}) \leq \frac{L\sqrt{2}}{2n} + \frac{2\sqrt{2}}{n2^s}.$$

Takeaway message

ReLU networks have no fundamental limitation in going from low dimension to a higher one.

References I

Bailey, B. and Telgarsky, M. J. (2018).
Size-noise tradeoffs in generative networks.
In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 6489–6499. Curran Associates, Inc.

Box, G. E. P. and Muller, M. E. (1958).
A note on the generation of random normal deviates.
Ann. Math. Statist., 29(2):610–611.