Deep neural network approximation theory

Dmytro Perekrestenko

August 2021

joint work with D. Elbrächter, P. Grohs, S. Müller, L. Eberhard and H. Bölcskei

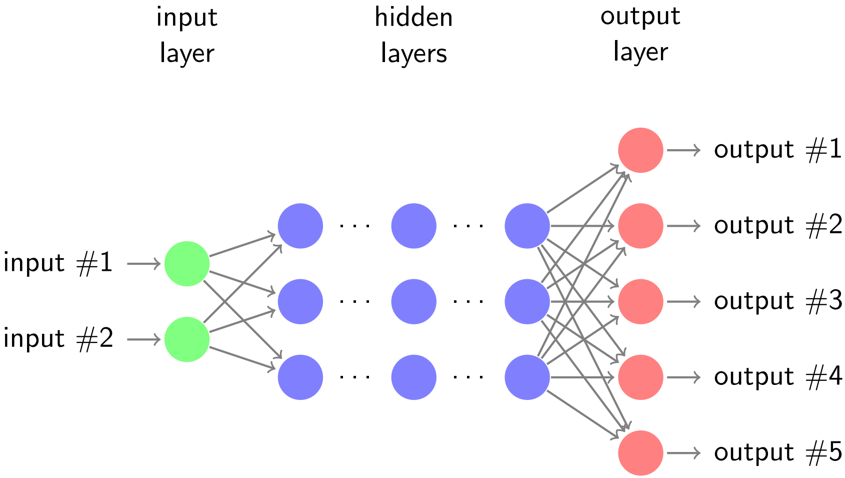
Our goals

- We study the **fundamental limits** of deep neural network learning.
- We assume an **optimal learning algorithm** and access to **infinite amounts of data**.
- We want to understand **fundamental limits** in representing functional relationships Φ (learned in practice) in the form

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

- We work in two settings: **function approximation** - $\|\Phi - f\|_\infty \leq \varepsilon$ and **probability distribution approximation** - $W(\Phi \# U, f) \leq \varepsilon$.

Neural networks



Composition of affine mappings
and non-linearities

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)**.

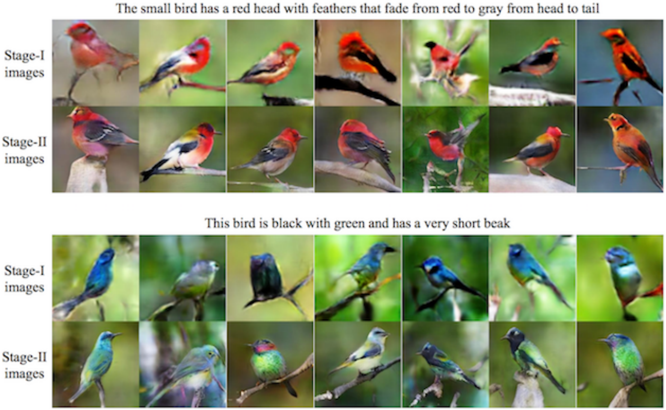
- Affine maps: $W_\ell = A_\ell x + b_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $\ell \in \{1, 2, \ldots, L\}$
- Non-linearity or activation function: $\rho : x \to \max(0, x)$ acts component-wise
- Network connectivity: $\mathcal{M}(\Phi)$ – total number of non-zero parameters in W_ℓ
- Depth of network or number of layers: $\mathcal{L}(\Phi) := L$
- Width of network: $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$

We denote by $\mathcal{N}_{d,d'}$ the set of all ReLU networks with input dimension $N_0 = d$ and output dimension $N_L = d'$. $\mathcal{N}_{1,d}$

Generation of photographs of human faces

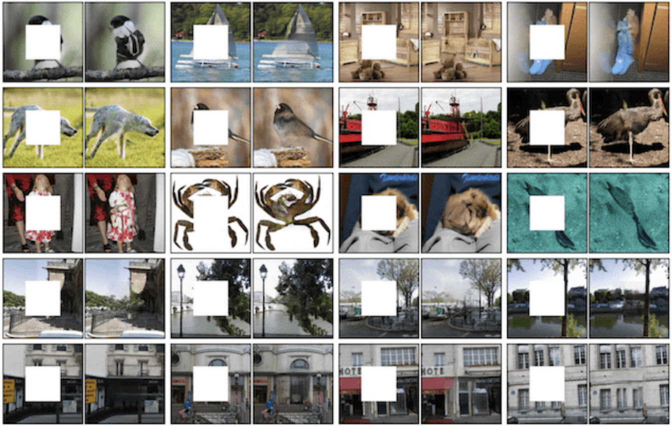
Examples of Photorealistic GAN-Generated Faces [Karras et al., 2018]

Text-to-image translation



Example of Textual Descriptions and GAN-Generated Photographs of Birds [Zhang et al., 2017]

Photo inpainting



Example of GAN-Generated Photograph Inpainting Using Context Encoders [Pathak et al., 2016]

And much more...

- Generation of Realistic Photographs
- Generation of Cartoon Characters
- Image-to-Image Translation
- Semantic-Image-to-Photo Translation
- Face Frontal View Generation
- Generate New Human Poses
- Photos to Emojis
- Photograph Editing
- Face Aging
- Photo Blending
- Super Resolution
- Clothing Translation
- Video Prediction
- 3D Object Generation

Outline of the talk

- Limits of learning functions
  - Approximation of basic functions, namely x^2, polynomials, and sinusoids
  - Approximation of function classes
  - Optimal representability
- Limits of learning distributions
  - Transporting between 1-dimensional distributions
  - Transporting to arbitrary high-dimensional distributions
  - Optimality of the generative network

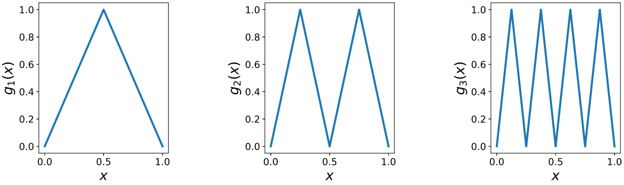
Sawtooth function

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

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

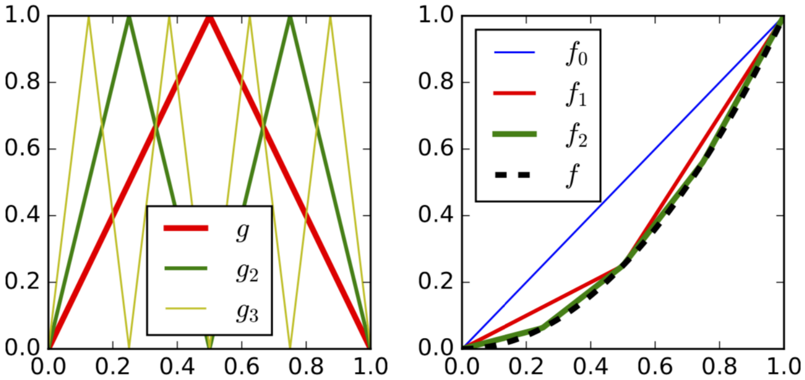
let $g_1(x) = g(x)$, and define the "sawtooth" function of order s as the s-fold composition of g with itself according to

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



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

Approximation of x^2



$$f_m(x) = x - \sum_{s=1}^{m} \frac{g_s(x)}{2^{2s}}$$

Image credit: [Yarotsky, 2017]

Follow-up results

Multiplication realized as a linear combination of squaring networks:

$$xy = \frac{1}{2}\big((x+y)^2 - x^2 - y^2\big)$$

Proposition (Polynomial approximation)

There exists a constant $C > 0$ *such that for all* $m \in \mathbb{N}$, $a = (a_i)_{i=0}^m \in \mathbb{R}^{m+1}$, $D \in \mathbb{R}_+$, *and* $\varepsilon \in (0, 1/2)$, *there is a network* $\Phi_{a,D,\varepsilon} \in \mathcal{N}_{1,1}$ *with*
$\mathcal{L}(\Phi_{a,D,\varepsilon}) \le Cm(\log(1/\varepsilon) + m\log(\lceil D\rceil) + \log(m) + \log(\lceil \|a\|_\infty \rceil))$,
$\mathcal{W}(\Phi_{a,D,\varepsilon}) \le 9$, *and satisfying*

$$\|\Phi_{a,D,\varepsilon}(x) - \sum_{i=0}^m a_i x^i\|_{L^\infty([-D,D])} \le \varepsilon.$$

Approximation of periodic functions

Main idea: Taylor series approximation of one period and periodic extension through "sawtooth" function.

Theorem (Cosine approximation)

There exists a constant $C > 0$ *such that for every* $a, D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, *there is a network* $\Psi_{a,D,\varepsilon} \in \mathcal{N}_{1,1}$ *with* $\mathcal{L}(\Psi_{a,D,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil aD \rceil))$, $\mathcal{W}(\Psi_{a,D,\varepsilon}) \leq 9$, *and satisfying*

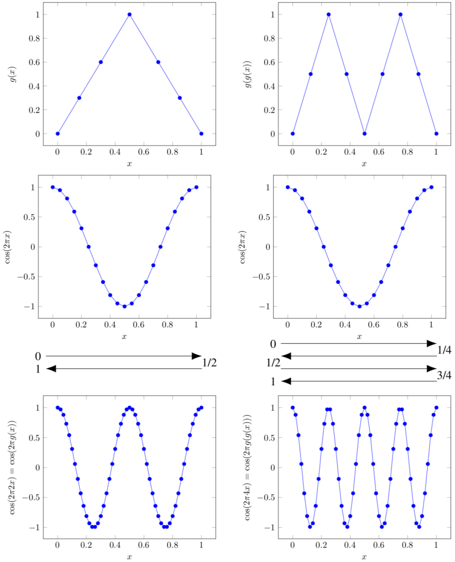
$$\|\Psi_{a,D,\varepsilon}(x) - \cos(ax)\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Approximation of periodic functions con't

$x \mapsto \cos(2\pi x)$ is 1-periodic and even. Recall the "sawtooth" functions $g_s : [0,1] \to [0,1]$ and note that

$$\cos(2\pi 2^s x) = \cos(2\pi g_s(x)).$$

This "periodization trick" avoids coefficients of exponential magnitude, coming from Taylor polynomial for $\cos(ax)$.



Exponential approximation accuracy

- Approximating network has **finite width** and **depth scaling poly-log** in $1/\varepsilon$.

- Owing to

$$\mathcal{M}(\Phi) \leq \mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi)+1),$$

we have

$$\varepsilon \leq 2^{-(\mathcal{M}(\Phi))^{1/p}}.$$

- Finite width combined with poly-log (in $1/\varepsilon$) depth yields **exponential error decay in connectivity**.

Approximation of signal classes

Definition

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and consider compact $\mathcal{C} \subset L^2(\Omega)$, to which we refer as **function class**.

Encoders and decoders:

$$\mathfrak{E}^\ell := \left\{ E : \mathcal{C} \to \{0,1\}^\ell \right\} \qquad \mathfrak{D}^\ell := \left\{ D : \{0,1\}^\ell \to L^2(\Omega) \right\}$$

Complexity is measured in **the number of bits** needed to store $\mathcal{C}$.

- Classical encoders - dictionaries (countable set of functions).
- We develop theory for neural network encoders.

Optimal exponent

Definition

Minimax code length:

$$L(\varepsilon, \mathcal{C}) := \min \Big\{ \ell \in \mathbb{N} : \exists (E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell : \sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon \Big\}$$

Optimal exponent:

$$\gamma^*(\mathcal{C}) := \sup \Big\{ \gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) \in \mathcal{O}\Big(\varepsilon^{-1/\gamma}\Big), \, \varepsilon \to 0 \Big\}$$

- $\gamma^*(\mathcal{C})$ quantifies "description complexity" of function class $\mathcal{C}$
- Larger $\gamma^*(\mathcal{C}) \Rightarrow$ smaller growth rate $\Rightarrow$ smaller memory requirements for storing signals $f \in \mathcal{C}$

Nonlinear approximation through dictionaries

For a function class $\mathcal{C} \subset L^2(\Omega)$, and a dictionary $\mathcal{D} = (\varphi_i)_{i \in I} \subset L^2(\Omega)$, $\gamma^*(\mathcal{C}, \mathcal{D})$ is defined as the supremal $\gamma > 0$ in

$$\sup_{f \in \mathcal{C}} \inf_{\substack{I_M \subseteq I, \\ \#I_M = M, (c_i)_{i \in I_M}}} \left\| f - \sum_{i \in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}), \ M \to \infty$$

- **Restrict search** for the M elements in $\mathcal{D}$ to the first $\pi(M)$ elements.
- **Require** that the **coefficients** c_i be **uniformly bounded** so that they can be quantized and stored with a finite number of bits.

If $\gamma^*(\mathcal{C}, \mathcal{D})$ satisfying these conditions is equal $\gamma^*(\mathcal{C})$, we say that the function class $\mathcal{C}$ is **optimally representable** by $\mathcal{D}$.

Function classes and their optimal exponents

Class	F	optimal dictionary	$\gamma^*(\mathcal{C})$
L^2-Sobolev	W_2^m	Fourier or Wavelet	m
L^p-Sobolev*	W_p^m	Wavelet	m/d
Hölder	C^α	Wavelet	α
Bump Algebra	$B_{1,1}^1$	Wavelet	1
Bounded Variation	BV	Haar	1
Besov**	$B_{p,q}^m$	Wavelet	m/d
Modulation***	$M_{p,p}^s$	Wilson	$\frac{1}{1/p-1/2+2s/d}$

* $p \in [1, \infty], m > d(1/p - 1/2)_+$
** $p, q \in (0, \infty], m > d(1/p - 1/2)_+$
*** $1 < p < 2, s \in \mathbb{R}_+$

Approximation with deep neural networks

- We develop the **new concept** of **best M-weight approximation** through deep neural networks
- **Neural network** interpreted as an **encoder** and its **complexity** is measured in terms of **number of bits** needed to store **network topology and quantized weights**

Best M-weight approximation

For a function class $\mathcal{C} \subseteq L^2(\Omega)$, $\gamma^*_{\mathcal{N}}(\mathcal{C})$ is defined as the supremal $\gamma > 0$ in

$$\sup_{f \in \mathcal{C}} \inf_{\substack{\Phi \in \mathcal{N}_{d,1} \\ \mathcal{M}(\Phi) \leq M}} \|f - \Phi\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}), \ M \to \infty.$$

- Infimum over **all possible network topologies**. The rate **benchmarks all learning algorithms** that map an f and an $\varepsilon > 0$ to a neural network.
- In order to encode, we additionally need **polylogarithmic depth and polynomial weight growth** in M.

If $\gamma^*_{\mathcal{N}}(\mathcal{C})$ satisfying these conditions is equal $\gamma^*(\mathcal{C})$, we say that the function class $\mathcal{C}$ is **optimally representable by neural networks.**

Transitioning from dictionaries to neural networks

- For given $\mathcal{C}$ and associated $\mathcal{D}$, we establish conditions guaranteeing the existence of a neural network with connectivity $O(M)$ that achieves the same uniform error over $\mathcal{C}$ as best M-term approximation.

- Simply put, if all elements in $\mathcal{D}$ are approximated by a network with **exponential error decay in connectivity**, then $\mathcal{D}$ is **effectively representable by neural networks.**

- Leads to a **characterization of function classes** $\mathcal{C}$ that are **optimally representable by neural networks.**

Affine dictionaries - scaling and translation

Definition (Affine dictionary)

Consider the compactly supported functions

$$g_s := \sum_{k=1}^{r} c_k^s f(\cdot - b_k), \quad s = 0, \dots, S.$$

We define the **affine dictionary** $\mathcal{D} \subset L^2(\Omega)$ corresponding to $(g_s)_{s=0}^{S}$ according to

$$\mathcal{D} := \Big\{ g_s^{j,e} := \Big(|\det(A_{s,j})|^{\frac{1}{2}} g_s(A_{s,j} \cdot - \delta e) \Big) \Big|_{\Omega} : s \in [1:S], e \in \mathbb{Z}^d, \\ j \in \mathbb{N}, \text{ and } g_s^{j,e} \neq 0 \Big\},$$

and refer to f as the **generator (function) of $\mathcal{D}$.**

Includes wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules.

Gabor dictionaries - frequency shifts

Definition (Gabor dictionaries)

Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d)$, and $x, \xi \in \mathbb{R}^d$. Define the translation operator $T_x \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ according to

$$T_x f(t) := f(t - x)$$

and the modulation operator $M_\xi \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d, \mathbb{C})$ as

$$M_\xi f(t) := e^{2\pi i \langle \xi, t \rangle} f(t) = \cos(2\pi \langle \xi, t \rangle) f(t) + i \sin(2\pi \langle \xi, t \rangle) f(t).$$

Let $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, and $g \in L^2(\mathbb{R}^d)$. The Gabor dictionaries $\mathcal{G}(g, \alpha, \beta, \Omega) \subseteq L^2(\Omega)$ is defined as

$$\mathcal{G}(g, \alpha, \beta, \Omega) := \left\{ M_\xi T_x g\big|_\Omega \colon (x, \xi) \in \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d \right\}.$$

Includes Wilson bases.

Optimality transfer

Central result

Optimality of a representation system $\mathcal{D}$ for a signal class $\mathcal{C}$ combined with effective representability of $\mathcal{D}$ by neural networks implies optimal representability of $\mathcal{C}$ by neural networks.

Optimal dictionaries

Affine dictionaries (e.g. wavelets, ridgelets, curvelets, shearlets, α-shearlets) and Gabor dictionaries are optimally representable by neural networks.

Main results - function approximation

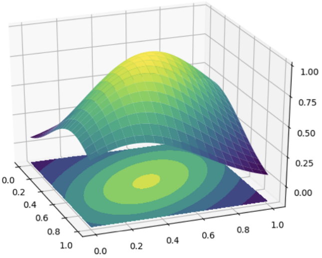
- Deep neural networks provide **exponential approximation accuracy** for a wide range of functions such as the squaring operation, multiplication, polynomials, sinusoidal functions, and even one-dimensional oscillatory textures and fractal functions.

- Deep neural networks can **learn optimally vastly different function classes** such as affine dictionaries, Gabor dictionaries, and smooth functions.

- This universality is afforded by a **concurrent invariance property of deep networks to translations, scalings, and frequency-shifts**.

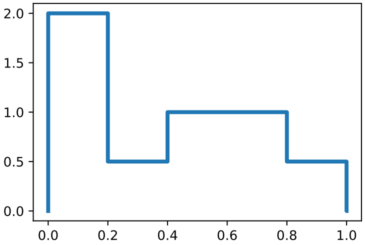
Generation of multi-dimensional distributions from $U[0,1]$

$$U[0,1] \xrightarrow{\mathcal{N}_{1,d}} p_{X,Y}$$

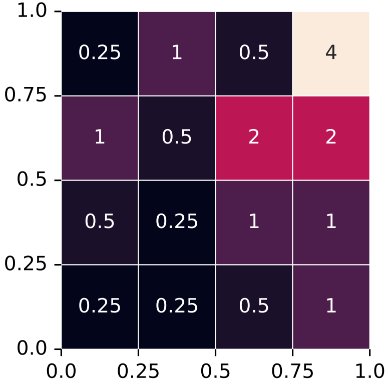


We will show that there is no fundamental limitation in going from low dimension to a higher one.

Histogram distributions



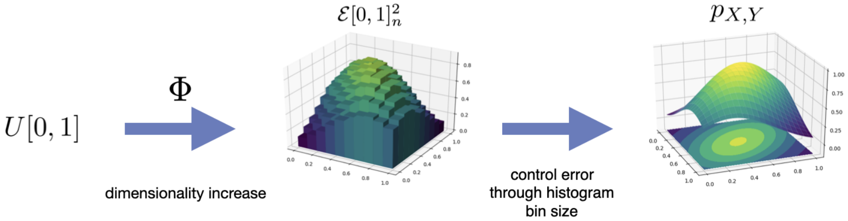
Histogram distribution $\mathcal{E}[0,1]^1_n$,
$d = 1, n = 5$.



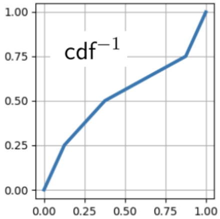
Histogram distribution $\mathcal{E}[0,1]^2_n$,
$d = 2, n = 4$.

Our goal

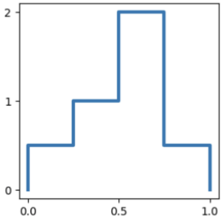
Transport $U[0,1]$ to an approximation of any given distribution supported on $[0,1]^d$. For illustration purposes we look at $d = 2$.



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}).

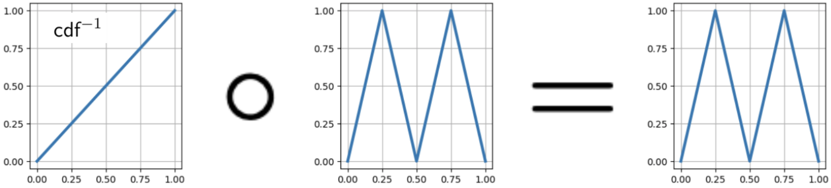
Related work

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*

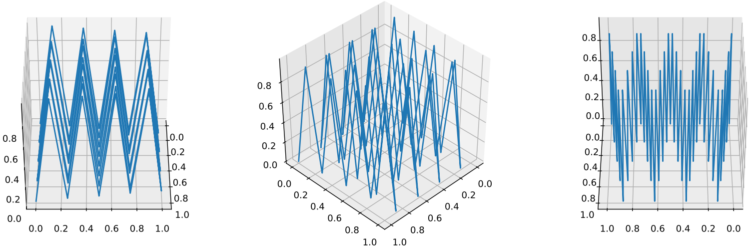
$$W(\Phi \# U[0,1], U[0,1]^2) \leq \frac{\sqrt{2}}{2^s}.$$

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



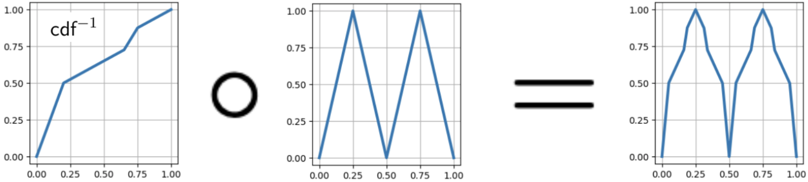
Transporting uniform distribution to higher dimensions

$$M : x \to \left(x, g_s(x), g_{2s}(x), \ldots, g_{(d-1)s}(x)\right)$$



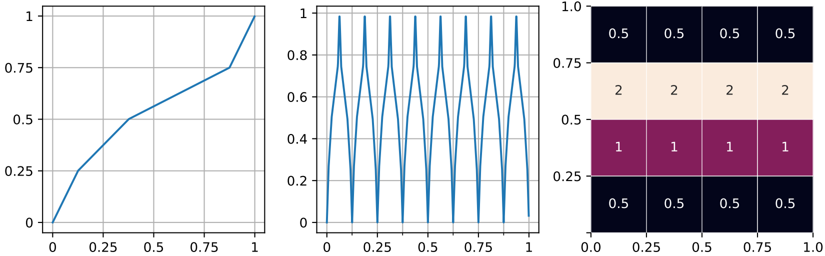
Generating a 3D uniform distribution via $x \to (x, g_3(x), g_6(x))$.

Generalization of the space-filling property



Approximating $2D$ distributions

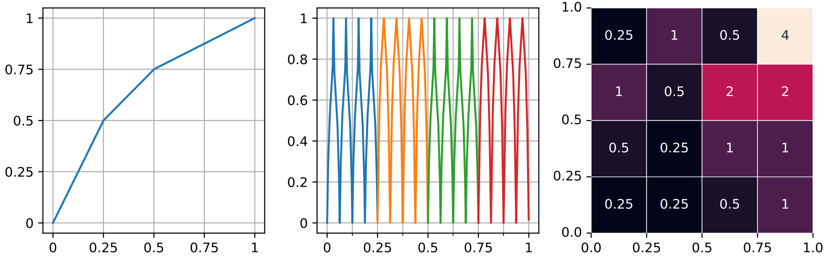
$$M : x \to (x, f(g_s(x)))$$



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

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



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

Generalization to d dimensions

Definition

Let $\mathbf{z} \in [0:(n-1)]^d$, $\mathbf{z}_i = \mathbf{z}_{|\mathbb{R}^{i-1}}$, and let $f_{X_i}^{\mathbf{z}_i}$ be the piecewise linear function that satisfies $f_{X_i}^{\mathbf{z}_i} \# U[0,1] = p_{X_i}^{\mathbf{z}_i}$, for all $i \in [1:d]$, and let for all $s \in \mathbb{N}$

$$F_0(x, \mathbf{z}_1, s) := x,$$

$$F_r(x, \mathbf{z}_{r+1}, s) := g_s\Big(n f_{X_r}^{\mathbf{z}_r}\big(F_{r-1}(x, \mathbf{z}_r, s)\big) - z_r\Big),\ 1 \leq r < d.$$

We define Z_r recursively as

$$Z_1(x, s) := f_{X_1}^{\mathbf{z}_1}(x),$$

$$Z_r(x, s) := \sum_{\mathbf{z}_r} f_{X_r}^{\mathbf{z}_r}\big(F_{r-1}(x, \mathbf{z}_r, s)\big),\ 1 < r \leq d.$$

Generalization to d dimensions con't

- $F_{r-1}(x, \mathbf{z}_r, s)$ - sth-order sawtooth function localizing mass $p_{\mathbf{X}}(\mathbf{x}_{|\mathbb{R}^r} \in c_{\mathbf{z}_r})$ on the set $c_{\mathbf{z}_r} \times [0,1]$ uniformly along the rth coordinate.
- $Z_r(x, s)$ modifies the slope per linear region of F_{r-1} to approximate the conditional distributions along the rth coordinate.

Theorem

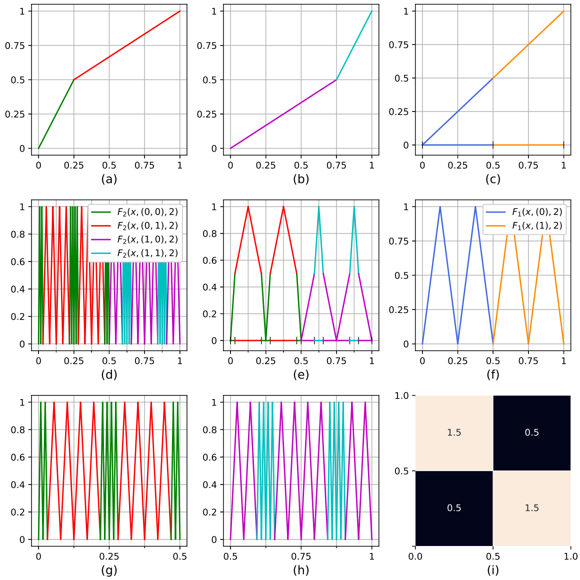
For every distribution $f_{\mathbf{X}}$ in $\mathcal{E}[0,1]^d_n$, the map

$$M : x \to (Z_1(x,s), Z_2(x,s), \ldots, Z_d(x,s))$$

satisfies

$$W(M\#U[0,1], f_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}.$$

Generalization to d dimensions con't



Generating histogram distributions with NNs

Theorem

For any $f_{\mathbf{X}} \in \mathcal{E}[0,1]^d_n$, $d > 1$, there exists a ReLU network $\Psi \in \mathcal{N}_{1,d}$ with $\mathcal{M}(\Psi) = O(n^d + sn^{d-1})$, $\mathcal{L}(\Psi) = (s+3)d - s$, such that

$$W(\Psi \# U, f_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}.$$

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

Universal approximation

Theorem

For any distribution ν *on* $[0,1]^d$*, there exists a ReLU network* $\Phi \in \mathcal{N}_{1,d}$ *with* $\mathcal{M}(\Phi) = O(n^d + sn^{d-1})$ *and* $\mathcal{L}(\Phi) = (s+3)d - s$ *such that*

$$W(\Phi \# U, \nu) \leq \frac{\sqrt{d}}{n2^s} + \frac{2\sqrt{d}}{n}.$$

Takeaway message

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

Fundamental lower bound on encoding distributions

Definition ([Graf and Luschgy, 2000])

The minimal n-term quantization error of a given distribution ν and $n \in \mathbb{N}$ is defined as $V_n(\nu) := \inf\{W(\nu, \mu) : |\mathrm{supp}(\mu)| \leq n\}$.

Theorem ([Graf and Luschgy, 2000][Theorem 6.2])

Let $X \sim \nu$ with $\mathbb{E}\|X\|^{1+\delta} < \infty$ for some $\delta > 0$, then

$$\lim_{n \to \infty} n^{1/d} V_n(\nu) = C,$$

where $C > 0$ is a constant depending only on d.

Allows to conclude that to encode a probability distribution one needs at least $d \log(\varepsilon^{-1})$ bits.

Complexity of generative networks

Lemma

Consider the class of quantized histogram distributions $\tilde{\mathcal{E}}_\delta[0,1]^d_n$ and let $\varepsilon \in (0, 1/2)$. Then, there exists a set of $\frac{\delta}{n}$-quantized ReLU networks $\Phi(\varepsilon, \cdot)$ of cardinality $2^{\ell(\varepsilon)}$, where $\ell(\varepsilon) \leq C \log(\varepsilon^{-1})$, with C a constant depending on d, δ, n, such that

$$\sup_{\nu \in \tilde{\mathcal{E}}_\delta[0,1]^d_n} W(\Phi(\varepsilon, \nu) \# U, \nu) \leq \varepsilon.$$

Complexity of generative networks con't

Lemma

Consider the class of non-singular distributions supported on $[0,1]^d$, *denoted by* $\mathcal{F}([0,1]^d)$, *and let* $\varepsilon \in (0, 1/2)$. *Then, there exists a set of quantized ReLU networks* $\Phi(\varepsilon, \cdot)$ *of cardinality* $2^{\ell(\varepsilon)}$, *where* $\ell(\varepsilon) \leq C\varepsilon^{-d} \log^2(\varepsilon^{-1})$, *with* C *a constant depending on* d, *such that*

$$\sup_{\nu \in \mathcal{F}([0,1]^d)} W(\Phi(\varepsilon, \nu)\#U, \nu) \leq \varepsilon.$$

Main results - distribution generation

- Deep neural networks are able to **generate any d-dimensional probability distribution with bounded support** without incurring a cost relative to generating the d-dimensional target distribution from d independent random variables.

- For **histogram target distributions**, the number of bits needed to uniquely encode the corresponding generative network is **close to the fundamental limit as dictated by quantization theory**.

- This is enabled by a **vast generalization of the space-filling approach** discovered recently in [Bailey and Telgarsky, 2018].

References

D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei,
Deep neural network approximation theory, IEEE Transactions on Information Theory, invited feature paper, 2021.

D. Perekrestenko, L. Eberhard, and H. Bölcskei,
High-dimensional distribution generation through deep neural networks, Partial Differential Equations and Applications, Springer, invited paper, 2021.

Other references

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.

Graf, S. and Luschgy, H. (2000).
Foundations of Quantization for Probability Distributions.
Springer-Verlag, Berlin, Heidelberg.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018).
Progressive growing of gans for improved quality, stability, and variation.
ArXiv, abs/1710.10196.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. (2016).
Context encoders: Feature learning by inpainting.