Mathematics of Deep Learning, Summer Term 2020

Week 4

Kolmogorov–Arnold Representation

Philipp Harms Lars Niemann

University of Freiburg

Overview of Week 4

1. Hilbert's 13th Problem
2. Kolmogorov–Arnold Representation
3. Approximate Hashing for Specific Functions
4. Approximate Hashing for Generic Functions
5. Proof of the Kolmogorov–Arnold Theorem
6. Approximation by Networks of Bounded Size
7. Wrapup

Acknowledgement of Sources

Sources for this lecture:

- Arnold (1958): On the representation of functions of several variables
- Torbjörn Hedberg: The Kolmogorov Superposition Theorem. In Shapiro (1971): Topics in Approximation Theory
- Philipp Christian Petersen (Faculty of Mathematics, University of Vienna): Course on Neural Network Theory.

Mathematics of Deep Learning, Summer Term 2020

Week 4, Video 1

Hilbert's 13th Problem

Philipp Harms Lars Niemann

University of Freiburg

Hilbert's 13th Problem

Hilbert's 13th problem

Can the roots of the equation

$$x^7 + ax^3 + bx^2 + cx + 1 = 0$$

be represented as superpositions of continuous functions of two variables?

Remark:

- This is the general form of a septic equation after some algebraic transformations. The roots are functions of (a, b, c).
- A single superposition is $w\big(u(a,b), v(b,c)\big)$, and a double superposition is $w\Big(u\big(p(a,b), q(b,c)\big), v\big(r(b,c), s(c,a)\big)\Big)$.
- More generally, the question becomes: Do functions of three variables exist at all, or can they be represented as superpositions of functions of less than three variables?

Hilbert's Conjecture

Conjecture: Hilbert conjectured that such reductions to smaller numbers of variables are impossible. The strongest supporting evidence is:

Theorem (Vitushkin 1955)

There is a polynomial such that neither the polynomial itself nor any function sufficiently close to it (in the sense of uniform convergence) can be decomposed into a simple superposition of continuous functions of two variables in any region or in any system of coordinates.

Dimension theory

Remark: Kolmogorov interpreted Hilbert's problem using dimension theory:

- Let $N(\epsilon)$ be the smallest number of ϵ-balls needed to cover a metric space X.
- On $X = [0,1]^n$ one has $\dim(X) := \liminf_{\epsilon \to 0} \frac{-\log N(\epsilon)}{\log \epsilon} = n$.
- On $X = C^s([0,1]^n)$ one has $\dim(X) := \liminf_{\epsilon \to 0} \frac{-\log \log N(\epsilon)}{\log \epsilon} = n/s$.
- In this sense, Hölder functions of 3 variables are strictly richer than Hölder functions of 2 variables.
- However, as we will see, this argument does not generalize to continuous functions.

Reduction to three variables

Theorem (Kolmogorov 1956)

Any continuous function f of $n \in \mathbb{N}$ variables can be represented as a finite number of superpositions of functions of 3 variables. For instance, for $n = 4$ one has

$$f(x_1, x_2, x_3, x_4) = \sum_{i=1}^{4} g^i\big(u(x_1, x_2, x_3), v(x_1, x_2, x_3), x_4\big)$$

for some continuous functions $g^i, u, v \colon \mathbb{R}^3 \to \mathbb{R}$.

Sketch of Proof: Reduction to three variables

Sketch of Proof:

- The level sets (aka. contour lines) of a continuous function form a tree (Kronrod, Menger):

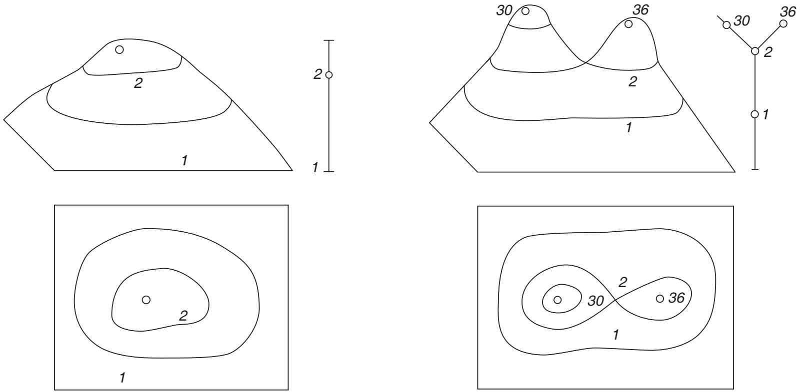


Figure: Figure from Arnold (1956)

Sketch of Proof: Reduction to three variables (cont.)

- Any continuous function of n variables can be written as a sum of $n+1$ continuous functions with standard trees, i.e., trees which do not depend on the given function (Kolmogorov):

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n+1} f^i(x_1, \ldots, x_n).$$

- Each of function f_i can be written as a one-parameter family of functions of $n-1$ variables:

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n+1} f^i_{x_n}(x_1, \ldots, x_{n-1})$$

Sketch of Proof: Reduction to three variables (cont.)

- Each of the functions $f^i_{x_n}$ factors through a function on the corresponding standard tree:

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n+1} g^i_{x_n}(\ell^i(x_1, \ldots, x_{n-1})).$$

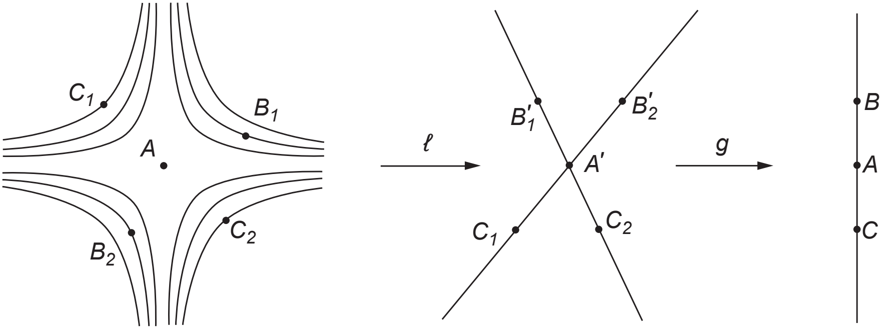


Figure: Figure from Arnold (1956)

Sketch of Proof: Reduction to three variables (cont.)

- Embedding the trees in a plane with a two-dimensional coordinate system (u, v) transforms this into:

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n+1} g^i_{x_n}\big(u^i(x_1, \ldots, x_{n-1}), v^i(x_1, \ldots, x_{n-1})\big).$$

- This yields 3-variate functions g_i and $(n-1)$-variate functions u^i, v^i:

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n+1} g^i\big(u^i(x_1, \ldots, x_{n-1}), v^i(x_1, \ldots, x_{n-1}), x_n\big).$$

- Applying this construction iteratively to u^i and v^i yields the reduction to superpositions of functions of 3 variables. $\square$

Questions to Answer for Yourself / Discuss with Friends

- Repetition: State Hilbert's 13th problem and describe how Kolmogorov cast it in the frameworks of dimension and graph theory.
- Check: What happens to Hilbert's problem when continuous functions are replaced by measurable or arbitrary functions?
- Background: Find out about generalizations, limitations, and open problems related to Hilbert's thirteenth problem.

Mathematics of Deep Learning, Summer Term 2020

Week 4, Video 2

Kolmogorov–Arnold Representation

Philipp Harms Lars Niemann

University of Freiburg

Kolmogorov–Arnold Representation

Theorem (Kolmogorov–Arnold 1956–1957)

For every $n \in \mathbb{N}_{\geq 2}$, there exist $\varphi_{i,j} \in C([0,1])$ such that any $f \in C([0,1]^n)$ can be represented as

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{2n+1} g_i \left(\sum_{j=1}^{n} \varphi_{i,j}(x_j) \right),$$

for some $g_i \in C(\mathbb{R})$.

Remark:

- This disproves Hilbert's conjecture and shows that "the only" multivariate function is a sum.
- The inner functions $\varphi_{i,j}$ are universal, i.e., they do not depend on f.
- The outer functions g_i can be learned by linear regression.

Sprecher's Refinement: Universal Inner Function

Theorem (Sprecher 1965, Köppen 2002)

For every $n \in \mathbb{N}_{\geq 2}$, there exists a continuous function $\varphi : \mathbb{R} \to \mathbb{R}$ and constants $a, \lambda_j \in \mathbb{R}$ such that any $f \in C([0,1]^n)$ can be represented as

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{2n+1} g_i \left(\sum_{j=1}^{n} \lambda_j \varphi(x_j + ia) \right),$$

for some $g_i \in C(\mathbb{R})$.

Remark:

- The function φ and the constants λ_j and a can be constructed explicitly and are universal, i.e., independent of f.
- Sprecher's representation can be interpreted as a neural network.
- There are many further versions of the Kolmogorov–Arnold theorem with varying regularity and structural assumptions.

Sprecher's Refinement: Universal Inner Function

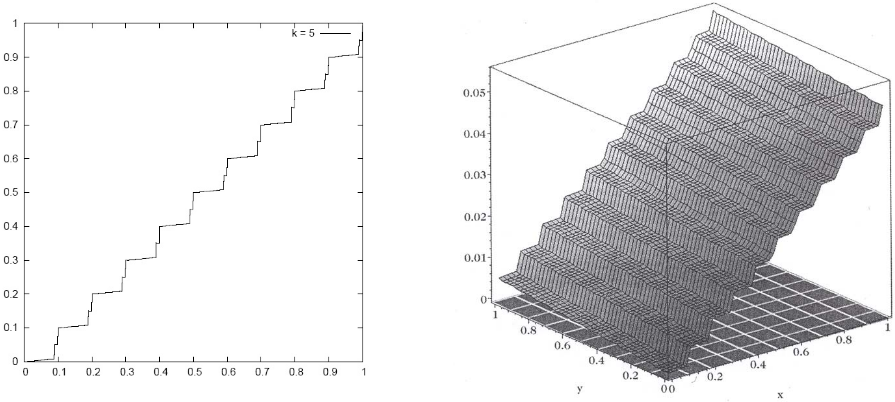


Figure: Sprecher's universal inner functions φ (left) and ψ_1 (right), where $\psi_i(x_1, x_2) := \lambda_1\varphi(x_1 + ia) + \lambda_2\varphi(x_2 + ia)$ for some constants λ_1, λ_2, a. [Leni Fougerolle Truchetet 2008]

Hashing

Remark:

- The inner functions in the Kolmogorov–Arnold representation theorem can be interpreted as hash functions.

Background:

- Hash functions are widely used in computer science for array indexing operations.
- They map high-dimensional/unstructured/variable-length data to scalar hash values.
- Hash functions should be fast to compute and should be “nearly” injective, i.e., minimize duplication of output values.

Hashing and Kolmogorov–Arnold Representation

Lemma

For each $i \in \{1, \ldots, 2n+1\}$, Sprecher's inner function

$$\psi_i \colon [0,1]^n \ni (x_1, \ldots, x_n) \mapsto \sum_{j=1}^{n} \lambda_j \varphi(x_j + ia) \in \mathbb{R}$$

is injective on a countable dense subset $D \subseteq [0,1]^n$.

Remark:

- It is sufficient to establish injectivity of $\psi(x) := \sum_j \lambda_j \varphi(x_j)$ on D.
- This follows from the following two facts: ϕ takes rational values on D, and the coefficients λ_j are independent over the rational numbers.
- Of course, ψ is not injective everywhere; otherwise the Kolmogorov–Arnold theorem would be trivial.

Space-filling curves

- Intuitively, the inverse of a hash function $[0,1]^n \to [0,1]$ is a space-filling curve, i.e., a surjective continuous map $[0,1] \to [0,1]^n$.
- For Sprecher's hash function, this is made precise as follows: By carefully examining the properties of ψ, one may construct an "inverse" map $\lambda : [0,1] \to [0,1]^n$ with the following properties:

Lemma

1. *The map $\lambda : [0,1] \to [0,1]^n$ is a space-filling curve.*
2. *Its image may be approximated by discrete curves Λ_k as $k \to \infty$.*

Remark:

- By the Hahn–Mazurkiewicz theorem, a non-empty Hausdorff topological space is a continuous image of the unit interval if and only if it is compact, connected, locally connected, and second-countable.

Space-filling curves

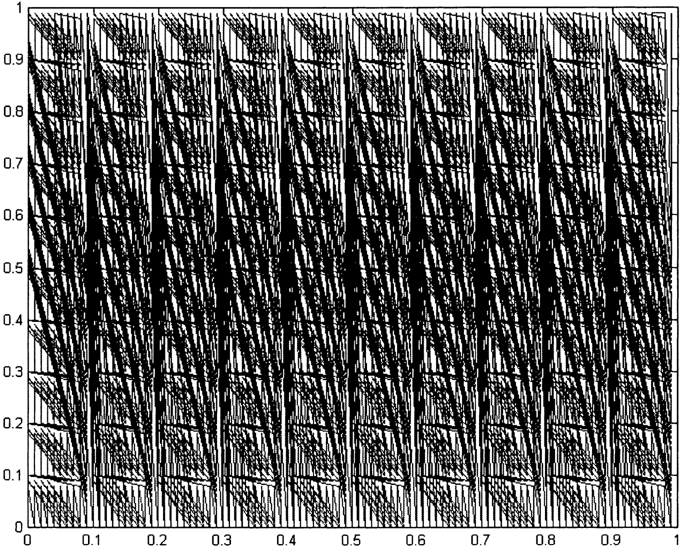


Figure: An approximation Λ_k of the space-filling curve λ. [Sprecher Draghici 2002]

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Recall and compare the presented versions of the Kolmogorov–Arnold Theorem.
- Check: Why exactly does the Kolmogorov–Arnold representation theorem disprove Hilbert's conjecture?
- Check: Show that there is no continuous bijection $[0,1]^n \to [0,1]$ for any $n \geq 2$.
- Discussion: How would you implement Sprecher's theorem using neural networks? Do you think this could work well for supervised learning?

Mathematics of Deep Learning, Summer Term 2020

Week 4, Video 3

Approximate Hashing for Specific Functions

Philipp Harms   Lars Niemann

University of Freiburg

Hashing rational numbers

Lemma

There exists a linear map $\ell\colon \mathbb{R}^n \to \mathbb{R}$ *whose restriction to rational numbers is injective.*

Proof:

- $n = 2$: Set $\ell(x, y) = x + \lambda y$ for any irrational number λ.
- $n \geq 2$: Set $\ell(x_1, \ldots, x_n) := \lambda_1 x_1 + \cdots + \lambda_n x_n$, where λ_i are independent over $\mathbb{Q}$, e.g. $\lambda_i = \pi^{i-1}$ or some other powers of any transcendental number. $\square$

Remark:

- Thus, any $f\colon \mathbb{Q}^n \to \mathbb{R}$ can be written as $f = g \circ \ell$, where ℓ is the above linear hashing function. However, g cannot be chosen continuously, and the approximation error cannot be controlled on non-rational numbers—a more elaborate construction is needed.
- We fix an irrational number $\lambda \in \mathbb{R} \setminus \mathbb{Q}$ throughout this section.

Approximate Hashing for a Specific Function

Remark:

- The key step in the proof of the Kolmogorov–Arnold theorem is the construction of approximate hashing functions.
- This is done here for a given specific function and in the next section for generic functions.
- We restrict ourselves to bivariate functions.

Definition (Approximate hashing functions, specific f)

A function $\varphi \in C([0,1], \mathbb{R}^5)$ is called approximate hashing function for $f \in C([0,1]^2)$ if there exists $g \in C(\mathbb{R})$ such that

$$\sup_{t \in \mathbb{R}} |g(t)| \leq 1/7, \qquad \sup_{x,y \in [0,1]} \left| f(x,y) - \sum_{i=1}^{5} g\big(\varphi_i(x) + \lambda \varphi_i(y)\big) \right| < 7/8.$$

Approximate Hashing for a Specific Function

Lemma

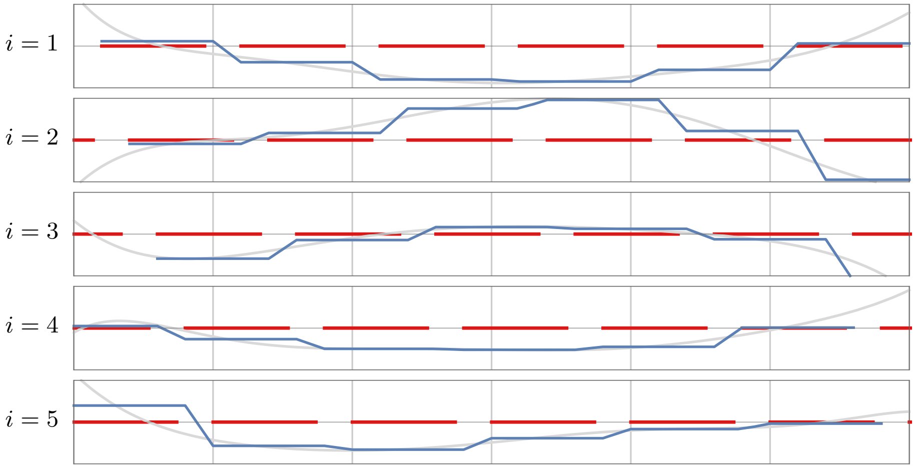
For any $f \in C^2([0,1]^2)$ with $\|f\|_\infty \leq 1$, the set of approximate hashing functions for f is open and dense in $C([0,1], \mathbb{R}^5)$.

Proof:

- The set is open, since if g works for a particular φ, it does so for every nearby φ.
- It remains to show that the set is dense in $C([0,1], \mathbb{R}^5)$.
- Thus, given $\epsilon > 0$ and $\chi \in C([0,1], \mathbb{R}^5)$, we have to find an approximate hashing function φ for f such that $\|\varphi - \chi\| \leq \epsilon$.

Proof: Approximate Hashing for a Specific Function

- Divide $[0,1]$ into $N \in \mathbb{N}$ intervals, cut out the i-th fifth of each interval, and color all remaining intervals red.
- Approximate χ_i (gray) by functions φ_i (blue), which are constant on red intervals of type i.



Proof: Approximate Hashing for a Specific Function

- It can be arranged that each function φ_i assumes distinct rational numbers on each of the red intervals, and that these numbers are distinct for different i.
- Moreover, for sufficiently large N, $\|\varphi - \chi\| \leq \epsilon$, as desired.
- Furthermore, by the uniform continuity of f on $[0,1]^2$, we can make N even larger to get

$$\left|f(x,y) - f(x',y')\right| \leq 1/7 \text{ whenever } \max\{\left|x - x'\right|, \left|y - y'\right|\} \leq 4/N.$$

Proof: Approximate Hashing for a Specific Function

- The function $\psi_i(x, y) := \varphi_i(x) + \lambda\varphi_i(y)$ is constant on red rectangles of type i, which are defined as products of red intervals of type i.
- The irrational numbers, which the functions ψ_i assume on rectangles of type i, are all distinct for different rectangles and/or different i.
- Thus, there is $g \in C(\mathbb{R})$ such that $g(\psi_i(x, y)) = \pm 1/7$ if (x, y) belongs to a red rectangle of type i where $f \gtrless 0$.
- Without loss of generality, $\|g\| \leq 1/7$.
- Intuitively, g tracks the sign of f on each rectangle.

Proof: Approximate Hashing for a Specific Function

- For any point (x, y), consider the approximation error

$$\left| f(x,y) - \sum_{i=1}^{5} g\big(\psi_i(x,y)\big) \right|. \qquad (*)$$

- If $f(x,y) \gtrless 1/7$, then $f \gtrless 0$ on each red rectangle containing (x,y).
- There are at least 3 such rectangles because out of 5 types, one may fail on the x-axis and another one on the y-axis.
- Thus, the majority of the summands in $(*)$ tracks the sign of f correctly, and the approximation error is bounded by $6/7$.
- If $|f(x,y)| \le 1/7$, the approximation error is again bounded by $6/7$, regardless of correct or incorrect tracking.
- As $6/7 < 7/8$, we have shown that φ is an approximate hashing function, which is ϵ-close to χ. $\square$

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Recall the definition of and main result on approximate hashing.
- Background: Refresh your memory of algebraic closures and the definition of algebraic and transcendental numbers, if necessary.
- Check: Draw the red rectangles of types 1 to 5 and verify that each point is contained in at least three rectangles.
- Check: What is the role of the numbers 5 and $1/7$ in the lemma? Can they be altered?

Mathematics of Deep Learning, Summer Term 2020

Week 4, Video 4

Approximate Hashing for Generic Functions

Philipp Harms   Lars Niemann

University of Freiburg

Approximate Hashing for Generic Functions

Remark:

- As before, we fix an irrational number $\lambda \in \mathbb{R} \setminus \mathbb{Q}$.

Definition (Approximate hashing functions)

A function $\varphi \in C([0,1], \mathbb{R}^5)$ is called approximate hashing function if for any $f \in C([0,1]^2)$, there exists $g \in C(\mathbb{R})$ such that

$$\|g\|_\infty \leq \frac{1}{7} \|f\|_\infty, \qquad \left\| f - \sum_{i=1}^{5} g \circ \psi_i \right\|_\infty \leq \frac{8}{9} \|f\|_\infty,$$

where $\psi_i(x,y) = \varphi_i(x) + \lambda \varphi_i(y)$.

Remark:

- Compared to hashing for specific functions f, this definition imposes the hashing property simultaneously for all f and with a slightly worse error bound.

Approximate Hashing for Generic Functions

Lemma

The set of approximate hashing functions is dense in $C([0,1],\mathbb{R}^5)$.

Proof:

- Let U_k be the sets of approximate hashing functions of f_k, for some dense sequence $(f_k)_{k\in\mathbb{N}}$ in the unit sphere of $C([0,1]^2)$.
- The sets U_k are open and dense. By Baire's category theorem, its intersection U is dense.
- Any function $\varphi \in U$ is an approximate hashing function: for any f with $\|f\|_\infty \leq 1$, there exists f_k and g such that

$$\begin{aligned}\Big\|f - \sum_i g\circ\psi_i\Big\|_\infty &\leq \|f - f_k\|_\infty + \Big\|f_k - \sum_i g\circ\psi_i\Big\|_\infty \\ &\leq \left(\tfrac{8}{9} - \tfrac{7}{8}\right) + \tfrac{7}{8} = \tfrac{8}{9}.\end{aligned}$$

- Extend to general f by scaling. □

Questions to Answer for Yourself / Discuss with Friends

- Repetition: What is the difference between hashing for specific versus generic functions, and how does the former imply the latter?
- Background: Refresh your memory of the Baire category theorem if necessary.
- Discussion: Can you strengthen the proof to get monotonically increasing approximate hashing functions?

Mathematics of Deep Learning, Summer Term 2020

Week 4, Video 5

Proof of the Kolmogorov–Arnold Theorem

Philipp Harms Lars Niemann

University of Freiburg

Kolmogorov–Arnold Representation, Refined Version

Remark: The approximate hashing results imply the following refined version of the Kolmogorov–Arnold representation theorem:

Theorem (Kolmogorov–Arnold representation, refined version)

For any $n \in \mathbb{N}_{\geq 2}$, there exist $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and $\varphi_1, \ldots, \varphi_{2n+1} \in C([0,1])$ such that any $f \in C([0,1]^n)$ admits a representation

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{2n+1} g\big(\lambda_1 \varphi_i(x_1) + \cdots + \lambda_n \varphi_i(x_n)\big)$$

for some continuous function g.

Remark: The difference to Kolmogorov's original result is that g does not depend on i.

Proof: Kolmogorov–Arnold Representation for $n = 2$

Proof: Iterative improvement of the approximate hashing representation

- Let $\varphi \in C([0,1], \mathbb{R}^5)$ be an approximate hashing function, define $\psi_i(x,y) = \lambda_1 \varphi_i(x) + \lambda_2 \varphi_i(y)$ for $\lambda_1 := 1$ and λ_2 irrational, and define $Tg := \sum_{i=1}^{5} g \circ \psi_i$.
- Set $f_1 := f$ and find g_1 with $\|g_1\|_\infty \leq \frac{1}{7}\|f_1\|_\infty$ and $\|f_1 - Tg_1\|_\infty \leq \frac{7}{8}\|f_1\|_\infty$.
- Set $f_2 := f_1 - Tg_1$ and find g_2 with $\|g_2\|_\infty \leq \frac{1}{7}\|f_2\|_\infty$ and $\|f_2 - Tg_2\|_\infty \leq \frac{7}{8}\|f_2\|_\infty$.
- Continue to eternity. When done, set $g = \sum_k g_k$ and note that $f = Tg$ as required. $\square$

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Recall the proof of the Kolmogorov–Arnold theorem via the construction of approximate hashing functions.
- Discussion: How does the proof work in higher dimensions?

Mathematics of Deep Learning, Summer Term 2020

Week 4, Video 6

Approximation by Networks of Bounded Size

Philipp Harms   Lars Niemann

University of Freiburg

Approximation by Networks of Bounded Size

Theorem

There exists a continuous, piece-wise polynomial activation function $\rho\colon \mathbb{R} \to \mathbb{R}$ *which allows one to approximate continuous multivariate functions by realizations of neural networks with bounded size, that is, for all* $n \in \mathbb{N}$ *there exists a constant* $C = C(n)$ *such that*

$$\forall \epsilon > 0 \; \forall f \in C([0,1]^n) \; \exists \Phi : \quad \mathrm{L}(\Phi) = 3, \, \mathrm{M}(\Phi) \leq C(n), \, \|f - \mathrm{R}(\Phi)\|_\infty \leq \epsilon.$$

Remark:

- This theorem is in a sense "too good" because it provides an approximate representation of continuous functions by finitely many real numbers.
- It highlights the influence of the choice of activation function on the resulting approximation theory.
- It also points to the importance of asking for bounded weights.

Approximation by Networks of Bounded Size

Lemma (Univariate case)

The theorem holds in the univariate case $n = 1$*: there exists a continuous, piecewise polynomial activation function* $\rho\colon \mathbb{R} \to \mathbb{R}$ *such that*

$$\forall \epsilon > 0 \; \forall f \in C([0,1]) \; \exists \Phi : \quad \mathrm{L}(\Phi) = 2, \quad \mathrm{M}(\Phi) \leq 3, \quad \|f - \mathrm{R}(\Phi)\|_\infty \leq \epsilon \, .$$

Remark: By translation and scaling, this extends to continuous functions f on every closed interval $[a,b] \subseteq \mathbb{R}$.

Proof: Approximation by Networks of Bounded Size

Proof of the lemma:

- Recall that the set Π of polynomials with rational coefficients is dense in the Polish space $C([0,1])$, and let $(\pi_i)_{i\in\mathbb{Z}}$ be an enumeration of Π.
- Define the activation function ρ by

$$\rho(x) := \begin{cases} \pi_i(x-2i), & x \in [2i, 2i+1] \\ \pi_i(1)(2i+2-x) + \pi_{i+1}(0)(x-2i-1), & x \in (2i+1, 2i+2) \end{cases}$$

- Note that, by the very definition of ρ, one has $\rho(x+2i) = \pi_i(x)$ for $x \in [0,1]$.
- Hence, the neural network $\Phi := ((1,2i),(1,0))$ has the desired properties. $\square$

Proof: Approximation by Networks of Bounded Size

Proof of the theorem:

- By the Kolmogorov–Arnold theorem (refined version),

$$f = \sum_{i=1}^{2n+1} g \circ \psi_i, \quad \psi_i(x_1, \dots, x_n) = \lambda_1 \varphi_i(x_1) + \dots + \lambda_n \varphi_i(x_n).$$

  for some $g \in C(\mathbb{R})$, $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ and $\varphi_1, \dots, \varphi_{2n+1} \in C([0,1])$.
- By the previous lemma, $\varphi_i \approx \mathrm{R}(\Phi_i) \in C([0,1])$ for some networks Φ_i and a piece-wise polynomial activation function ρ, where $\approx$ denotes approximation up to arbitrary accuracy.
- Then $\psi_i \approx \mathrm{R}(\Psi_i) \in C([0,1]^n)$ for each $i \in \{1, \dots, 2n+1\}$, where

$$\Psi_i = (((\lambda_1, \dots, \lambda_n), 0)) \bullet \mathrm{FP}(\Phi_i, \dots, \Phi_i).$$

Proof: Approximation of Multivariate Functions (cont.)

- By the previous lemma, $g \approx \mathrm{R}(\Xi) \in C([-K,K])$, where K is sufficiently large such that $\psi_i([0,1]^n) \subseteq [-K,K]$.
- Then the network

$$\Phi := (((1,\dots,1),0)) \bullet \mathrm{FP}(\Xi,\dots,\Xi) \bullet \mathrm{P}(\Psi_1,\dots,\Psi_{2n+1}).$$

  has the desired number of layers and weights.
- Moreover, $f \approx \mathrm{R}(\Phi)$ thanks to the estimate

$$\begin{aligned}\|f - \mathrm{R}(\Phi)\| &\leq \sum_i \|\mathrm{R}(\Xi) \circ \mathrm{R}(\Psi_i) - g \circ \psi_i\| \\ &\leq \sum_i \|\mathrm{R}(\Xi) \circ \mathrm{R}(\Psi_i) - \mathrm{R}(\Xi) \circ \psi_i\| + \|\mathrm{R}(\Xi) \circ \psi_i - g \circ \psi_i\|,\end{aligned}$$

  and thanks to the uniform continuity of $\mathrm{R}(\Xi)$ on $[-K,K]$. $\square$

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Recall the approximation of univariate and multivariate functions by networks of bounded size.
- Check: Verify that the activation function ρ constructed in the univariate case is continuous.
- Discussion: What are theoretical implications to approximation theory and practical implications to supervised learning?

Mathematics of Deep Learning, Summer Term 2020

Week 4, Video 7

Wrapup

Philipp Harms Lars Niemann

University of Freiburg

Outlook on this week's discussion and reading session

- Reading:
  - Arnold (1958): On the representation of functions of several variables
  - Bar-Natan (2009): Hilberts 13th problem, in full color
  - Hecht-Nielsen (1987): Kolmogorov's mapping neural network existence theorem

Summary by learning goals

Having heard this lecture, . . .

- You can describe the Kolmogorov–Arnold representation theorem and its proof.
- You can appreciate the fundamental distinction between inner and outer network layers.
- You are aware that different choices of activation functions may lead to very different approximation theories.