Foundations of Deep Learning

Lecture 03

Complexity Theory

Aurelien Lucchi

Fall 2024

Section 1

Complexity and Escaping the Curse of Dimensionality

Complexity: Two Questions

Function classes represented by neural networks are rich (enough), but ...

- 1. How many units are required to obtain a desired approximation accuracy?
- 2. Is there an advantage of compositionality (multiple layers)?

Fourier Transform

For any absolutely integrable f, i.e. $f \in L^1$ ($\int_{\mathbb{R}^d} |f(\mathbf{x})| d\mathbf{x} < \infty$), define the Fourier transform of f as

$$\boxed{\widehat{f}(\boldsymbol{\omega}) = \mathcal{F}f(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-2\pi i \boldsymbol{\omega}\cdot\mathbf{x}} f(\mathbf{x}) d\mathbf{x}, \quad \widehat{f} : \mathbb{R}^d \to \mathbb{C}}$$

- $\widehat{f}(\boldsymbol{\omega}) =$ result of the Fourier transform. It represents the function in the frequency domain: it tells us how much of each spatial frequency $\boldsymbol{\omega}$ is present in $f(\cdot)$.
- $e^{-2\pi i \boldsymbol{\omega}\cdot\mathbf{x}}$: This part essentially tests how well each spatial frequency component fits with the original function at each spatial position.

Fourier Transform

Convolution Let $r(x) = \{g * h\}(x) \triangleq \int_{-\infty}^{\infty} g(\tau)h(x-\tau)\,d\tau$.

Theorem 1 (Convolution theorem)

$$\hat{r}(\boldsymbol{\omega}) = \hat{g}(\boldsymbol{\omega})\hat{h}(\boldsymbol{\omega}) \quad \textit{and} \quad r(\mathbf{x}) = \mathcal{F}^{-1}(\hat{g}(\boldsymbol{\omega})\hat{h}(\boldsymbol{\omega})),$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform.

Regularity Class

Regularity condition on Fourier transform $\widehat{g}$ of a function g

$$\boxed{C_g := \int \|\boldsymbol{\omega}\| \, |\widehat{g}(\boldsymbol{\omega})| d\boldsymbol{\omega} < \infty}$$

$C_g < \infty$: Fourier transformation of gradient function has to be **absolutely integrable**.

If g is differentiable, the Fourier transform of ∇g is given by

$$\Rightarrow \boxed{\widehat{\nabla g}(\boldsymbol{\omega}) = \boldsymbol{\omega}\widehat{g}(\boldsymbol{\omega}) \, .}$$

Barron's Construction for Infinite-width

The main idea is very simple. It simply start from the inverse Fourier transform

$$f(\mathbf{x}) = \int \exp(2\pi i \boldsymbol{\omega}^\top \mathbf{x}) \widehat{f}(\boldsymbol{\omega}) d\boldsymbol{\omega}.$$

- We have written $f(\mathbf{x})$ as an infinite integral which we can interpret as an infinite-width neural network with a rather **strange complex activation function**
- Barron's construction will consists into converting these activations into **threshold activation functions**

Barron's Construction for Infinite-width

Theorem 2 (Infinite-width representation)

Assume $\int \|\widehat{\nabla f}(\boldsymbol{\omega})\| d\boldsymbol{\omega} < \infty$, $f \in L^1, \widehat{f} \in L^1$. *Then, for bounded* $\|\boldsymbol{\omega}\|$ *and* $\|\mathbf{x}\| \leq 1$, *we have the following* **infinite representation of f with threshold nodes***:*

$$
\begin{aligned}
& f(\mathbf{x}) - f(0) \\
&= \int \frac{\cos(2\pi\boldsymbol{\omega}^\top \mathbf{x} + 2\pi\theta(\boldsymbol{\omega})) - \cos(2\pi\theta(\boldsymbol{\omega}))}{2\pi\|\boldsymbol{\omega}\|} \|\nabla \widehat{f}(\boldsymbol{\omega})\| d\boldsymbol{\omega} \\
&= -2\pi \int \int_0^{\|\boldsymbol{\omega}\|} \mathbf{1}[\boldsymbol{\omega}^\top \mathbf{x} - b \geq 0] \sin(2\pi b + 2\pi\theta(\boldsymbol{\omega})) |\widehat{f}(\boldsymbol{\omega})| db d\boldsymbol{\omega} \\
&+ 2\pi \int \int_{-\|\boldsymbol{\omega}\|}^0 \mathbf{1}[-\boldsymbol{\omega}^\top \mathbf{x} + b \geq 0] \sin(2\pi b + 2\pi\theta(\boldsymbol{\omega})) |\widehat{f}(\boldsymbol{\omega})| db d\boldsymbol{\omega}.
\end{aligned}
$$

Proof

Barron's Theorem (1993)

Condition on σ: bounded (measurable) and monotonic function σ such that $\sigma(t) \overset{t\to\infty}{\longrightarrow} 1$ and $\sigma(t) \overset{t\to-\infty}{\longrightarrow} 0$.

Theorem 3

For every $g : \mathbb{R}^d \to \mathbb{R}$ with **finite C_g** *and any $r > 0$, there is a sequence of MLP functions $g_k(\mathbf{x})$ of the form*

$$g_k(\mathbf{x}) = \sum_{j=1}^{k} \beta_j \sigma(\boldsymbol{\theta}_j \cdot \mathbf{x} + b_j) + b_0$$

such that

$$\int_{r\mathbb{B}} (g(\mathbf{x}) - g_k(\mathbf{x}))^2 \mu(\mathbf{dx}) \leq \mathcal{O}\left(\frac{1}{k}\right)$$

where $r\mathbb{B} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq r\}$ and μ **is any probability measure** *on $r\mathbb{B}$.*

Interpretation

Main points of the theorem:

1. Lack of dependency on d. MLPs **do not suffer from the curse of dimensionality** when approximating (certain) functions.
2. Freedom in the choice of measure (data distribution)
3. Remarkable approximation error rate $\propto 1/m$.
4. Additional bounds and constraints on the parameters
5. Proof uses iterative construction: add units to fit residuals

Re:first point: linear combination of m basis functions has lower approximation error bound $(1/m)^{2/d}$ (much worse).

$\implies$ **Adaptivity of feature extraction is key!**

Barron's Theorem: proof preliminaries

Let $X = \mathbb{E}[V]$, where the random variable V is supported on set S (in a Hilbert space H)

How can we compute an estimate of the mean?

- Sample a set of random variables $\{V_1, \ldots, V_k\}$ and compute the empirical mean $\hat{X} = \frac{1}{k}\sum_{i=1}^{k} V_i$
- Want to show that $\hat{X}$ gets "closer" (in terms of norm) to X as we increase the number of samples k

Barron's Theorem: proof preliminaries

Lemma 4 (Maurey [Pisier(1981)])

Let $X = \mathbb{E}V$ *be given, with* V *supported on* $S \subset H$*, and let* $V_1, \dots V_k$ *be iid draws from the same distribution. Then*

$$\mathbb{E}_{V_1,\dots V_k} \left\| X - \frac{1}{k} \sum_{i=1}^{k} V_i \right\|^2 \leq \frac{\mathbb{E}\|V\|^2}{k} \leq \frac{\sup_{U \in S} \|U\|^2}{k}.$$

Moreover there exist $(U_1, \dots U_k)$ *in* S *so that*

$$\left\| X - \frac{1}{k} \sum_{i=1}^{k} U_i \right\|^2 \leq \mathbb{E}_V \left\| X - \frac{1}{k} \sum_{i=1}^{k} V_i \right\|^2 .$$

Proof: see exercise sheet.

Barron's Theorem: proof preliminaries

Can we extend our sampling lemma to **functions of random variables**?

- Need to define a valid Hilbert space: consider L^2 space for which the inner product is defined as follows: $\forall f, g \in \mathcal{F}, \langle f, g \rangle = \int f(x)g(x)dP(x)$ for some probability measure P on x
- Corresponding norm is $\|f\|^2_{L^2(P)} = \int f(x)^2 dP(x)$.

Barron's Theorem: proof preliminaries

Lemma 5 (Maurey for signed measure [Pisier(1981)])

Let μ denote a nonzero signed measure supported on $S \subseteq \mathbb{R}^p$, and $g(\mathbf{x}) = \int g(\mathbf{x},\boldsymbol{\omega})d\mu(\boldsymbol{\omega})$. Let $\tilde{\boldsymbol{\omega}}_1, \ldots, \tilde{\boldsymbol{\omega}}_k$ be i.i.d. draws from the corresponding $\tilde{\mu}$ and let P be a probability measure on x. Define $\tilde{g}$ such that $g = \mathbb{E}_{\tilde{\mu}}\tilde{g}$. Then

$$\mathbb{E}_{\tilde{\boldsymbol{\omega}}_1,\ldots\tilde{\boldsymbol{\omega}}_k}\left\|g(\cdot) - \frac{1}{k}\sum_{i=1}^{k}\tilde{g}(\cdot,\tilde{\boldsymbol{\omega}}_i)\right\|_{L^2}^2 \le \frac{\mathbb{E}\|\tilde{g}(\cdot,\tilde{\boldsymbol{\omega}})\|_{L^2}^2}{k}.$$

Moreover there exist $(\boldsymbol{\omega}_1, \ldots \boldsymbol{\omega}_k)$ and $s \in \{\pm 1\}^k$ in S s.t.

$$\left\|g(\cdot) - \frac{1}{k}\sum_{i=1}^{k}\tilde{g}(\cdot,\boldsymbol{\omega}_i, s_i)\right\|_{L^2}^2 \le \mathbb{E}_{\tilde{\boldsymbol{\omega}}_1,\ldots\tilde{\boldsymbol{\omega}}_k}\left\|g(\cdot) - \frac{1}{k}\sum_{i=1}^{k}\tilde{g}(\cdot,\tilde{\boldsymbol{\omega}}_i)\right\|_{L^2}^2$$

Barron's Theorem: proof idea

General idea: convert the infinite-size construction introduced in Theorem 2 on to a finite-size one.

To do so, we sample from the integral $\int \sigma(\boldsymbol{\omega}^\top \mathbf{x}) p(\boldsymbol{\omega}) d\boldsymbol{\omega}$ by using a finite estimate $\sum_{j=1}^m s_j \tilde{\sigma}(\boldsymbol{\omega}_j^\top \mathbf{x})$ with:

- $\tilde{\sigma}(z) = \sigma(z) \int |p(\boldsymbol{\omega})| d\boldsymbol{\omega}$
- $\boldsymbol{\omega}_j \sim \frac{|p(\boldsymbol{\omega})|}{\int |p(\boldsymbol{\omega})| d\boldsymbol{\omega}}$
- $s_j = \mathrm{sign}(p(\boldsymbol{\omega}_j))$.

Next: we will give a proof for the case where σ is a threshold node, i.e. $z \mapsto \mathbf{1}[z \geq 0]$ and where $\|x\| \leq 1$.

Barron's Theorem: proof

Section 2

Benefits of Depth

Benefits of Depth

- Consistent empirical evidence: deeper network yield better approximations
- Classical results: focus on strength of shallow models (e.g. single hidden layer)
- **Do deep networks offer representational benefits?**
- No comprehensive theory exists of why deeper models are preferred and when.
- We will see some interesting pieces of the puzzle: e.g. **paradigmatic example of a function that is much easier to approximate with 2 hidden layers than 1.**

Subsection 1

Separation between Shallow and Deep Networks

Main result

Theorem 6 ([Telgarsky(2016)])

Let any integer $L \geq 1$ *be given. There exists a ReLU neural network* $f : \mathbb{R} \to \mathbb{R}$ *with* $3L^2 + 6$ *nodes and* $2L^2 + 4$ *layers that can not be approximated by any ReLU network* g *with* $\leq 2^L$ *nodes and* $\leq L$ *layers such that*

$$\int_{[0,1]} |f(x) - g(x)| \, dx \geq \frac{1}{32}.$$

Measuring complexity

In order to prove this theorem, we will...

- define a notion of complexity that depends on the number of oscillations in the function implemented by the neural network,
- show that this complexity measure grows polynomially in width, but exponentially in depth.

How do we measure oscillations in a function?
$\longrightarrow$ simply count the number of affine pieces

Formally, let $\mathcal{F}$ be the set of piecewise univariate linear mappings on R. Given a function $f \in \mathcal{F}$, we denote by $\delta_A(f)$ the number of affine pieces of f.

Properties of $\delta_A(f)$

In order to study the number of oscillations in a neural network function, we will need the following properties of $\delta_A(f)$.

Proposition 7

1. *For all* $f \in \mathcal{F}$, *and* $\sigma(x) = \max(0, x)$, $\delta_A(\sigma(f)) \leq 2\delta_A(f)$,
2. *For all* $f_1, \ldots f_m \in \mathcal{F}$, $\delta_A\left(\sum_{i=1}^m f_i\right) \leq \sum_{i=1}^m \delta_A(f_i)$.

Properties of $\delta_A(f)$: Proof i)

Any linear piece of f that does not cross the 0-axis either stays a linear piece (if above the 0-axis) or is mapped to zero. Only pieces that cross the 0-axis are separated into two pieces.

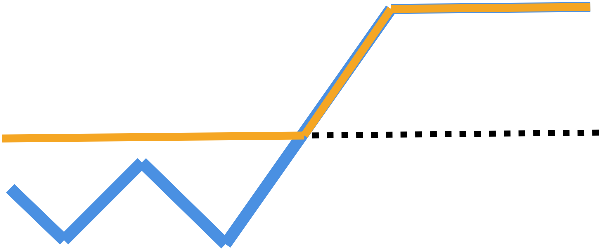
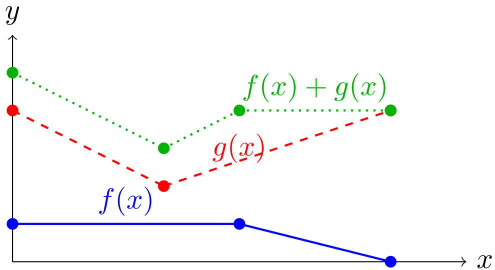


Figure: Illustration showing the effect of applying a ReLU function on a piecewize linear function. The dotted black line is the 0-axis. The piecewize linear function is shown in blue and the output of the ReLU function is shown in orange.

Properties of $\delta_A(f)$: Proof ii)

ii) When summing two piecewize functions $f(x) + g(x)$, there can only be a change in the slope at x if one of the functions f or g also had a change of slope at x.



Properties of $\delta_A(f)$

Lemma 8

Let $f : \mathbb{R} \to \mathbb{R}$ be a ReLU network with L layers of widths $(m_1, \ldots, m_L)$ such that $m = \sum_{i=1}^{L} m_i$. Let $g : \mathbb{R} \to \mathbb{R}$ denote the output of some node in layer i as a function of the input. Then the number of affine pieces $\delta_A(g)$ satisfies

$$\delta_A(g) \leq 2^i \prod_{j<i} m_j.$$

The number of affine pieces in f satisfies $\delta_A(f) \leq \left(\frac{2m}{L}\right)^L$.

Proof.

See exercise session. $\square$

Proof of main theorem

Main idea: We will create a highly oscillatory function f which we will approximate with a function g with few oscillations.

How? In order to create f, we will compose the following Δ function with itself such that the result of the composition increases its complexity (number of pieces).

$$\Delta(x) = 2\sigma_{\mathrm{r}}(x) - 4\sigma_{\mathrm{r}}(x-1/2) + 2\sigma_{\mathrm{r}}(x-1) = \begin{cases} 2x & x \in [0, 1/2), \\ 2 - 2x & x \in [1/2, 1), \\ 0 & \text{otherwise}, \end{cases}$$

where

$$\sigma_r(x) = \max(0, x).$$

Proof of main theorem

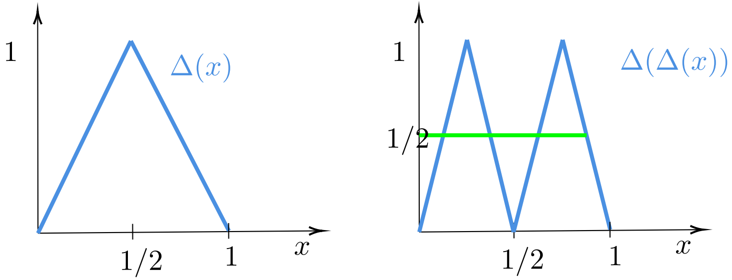


Figure: Illustration of the $\Delta(x)$ function as well as $\Delta(\Delta(x))$.

Try to compose the function with itself, $\Delta \circ \Delta$, what does the resulting function look like?
$\longrightarrow$ If you repeat this composition, you will see that Δ^L has 2^{L-1} copies of it self.

Proof of main theorem

Consider the highly oscillatory blue function $f(x) = \Delta^{L^2+2}(x)$.

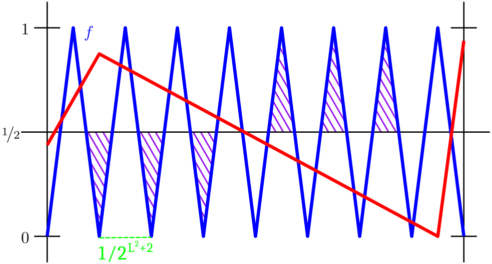


Figure: Source: [Telgarsky(2021)]

There are 2^{L^2+1} copies of $\Delta \implies 2^{L^2+2} - 1$ (half-)triangles since we get two triangles for each Δ but one lost on the boundary of $[0, 1]$.

Proof of main theorem

Goal: approximate f with the red function $g \in \mathcal{F}$ which has few oscillations.

$$\int_{[0,1]} |f-g| \geq [\text{number surviving triangles}] \cdot [\text{area of triangle}]$$

- $g \in \mathcal{F}$ crosses the axis $x = \frac{1}{2}$ at most $\delta_A(g)$ times
- Number of half-triangles on one side of the $x = \frac{1}{2}$ axis is larger than $2^{L^2+2} - 1 - 2\delta_A(g)$
- By Lemma 8 (with $m \leq 2^L$): $\delta_A(g) \leq (2 \cdot 2^L/L)^L \leq 2^{L^2}$
- Area of each triangle is $\frac{1}{4} \cdot \frac{1}{2^{L^2+2}} = 2^{-L^2-4}$

Properties of $\delta_A(f)$

We obtain the following bound

$$\begin{aligned}
\int_{[0,1]} |f-g| &\geq [\text{number surviving triangles}] \cdot [\text{area of triangle}] \\
&\geq \frac{1}{2}\left[2^{L^2+2}-1-2\cdot 2^{L^2}\right]\cdot\left[2^{-L^2-4}\right] \\
&= \frac{1}{2}\left[2^{L^2+1}-1\right]\cdot\left[2^{-L^2-4}\right] \\
&\geq \frac{1}{32}.
\end{aligned}$$

Subsection 2

Paradigmatic Example: when two Layers are Better than One

Idea

The key idea is very simple:

- Define a **target radial function** $g(\mathbf{x}) = \psi(\|\mathbf{x}\|)$
- ... that can be naturally approximated by first approximating the norm (via the span of the first hidden layer) and then approximating ψ (via the span of the second hidden layer).
- ... assuming that the norm can be approximated by span$(\mathcal{G}_\sigma^n)$ in an n-efficient manner and that this is not true for g.

Proof sketch, I: move to Fourier space

We are interested in the L_2 loss between f and a **target** g with regard to density ϕ^2 (i.e. $\int \phi^2(\mathbf{x})d\mathbf{x} = 1$)

$$\begin{aligned}\ell^\phi(f,g) :=& \int (f(\mathbf{x}) - g(\mathbf{x}))^2 \phi^2(\mathbf{x}) d\mathbf{x} \\ =& \int (f(\mathbf{x})\phi(\mathbf{x}) - g(\mathbf{x})\phi(\mathbf{x}))^2 d\mathbf{x} \\ =& \|f\phi - g\phi\|_{L^2}^2 \overset{(1)}{=} \|\widehat{f\phi} - \widehat{g\phi}\|_{L^2}^2 \overset{(2)}{=} \|\widehat{f} \star \widehat{\phi} - \widehat{g} \star \widehat{\phi}\|_{L^2}^2\end{aligned}$$

Here $\widehat{h}$ denotes the (generalized) Fourier transforms of h.

step (1): Parseval identity
step (2): convolution theorem.

Goal: chose ϕ and g to separate 1 vs. 2 hidden layer MLPs

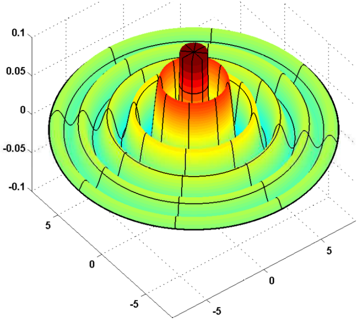
Proof sketch, II: Spotting the Weakness

Interest: support of $\widehat{f} \star \widehat{\phi}$.

$\rightsquigarrow$ Indicates what functions we can approximate.

Design choice: ϕ s.t. $\widehat{\phi} = \mathbf{1}[\mathbb{B}^n]$ (i.e. indicator on the unit ball)

$\implies$ Consequence: ϕ is isotropic and bandlimited



Proof sketch, II: Spotting the Weakness

- Single ridge function: $\sigma(\mathbf{x}) = \sigma(\mathbf{x} \cdot \boldsymbol{\theta})$

$$\Rightarrow \boxed{\mathrm{supp}\,(\widehat{\sigma}) = \mathrm{span}\{\boldsymbol{\theta}\}}$$

- Convolved ridge function:

$$\Rightarrow \boxed{\mathrm{supp}\left(\widehat{\sigma} \star \widehat{\phi}\right) = \mathrm{span}\{\boldsymbol{\theta}\} + \mathbb{B}}$$

- MLP with one hidden layer of width m implements a function

$$f \in \mathrm{span}\{\sigma_j(\mathbf{x}) := \sigma(\boldsymbol{\theta}_j \cdot \mathbf{x}),\, 1 \le j \le m\} \subset \mathrm{span}(\mathcal{G}_\sigma^n)$$

Linear combination of convolved ridge functions:

$$\Rightarrow \boxed{\mathrm{supp}(\widehat{f} \star \widehat{\phi}) = \bigcup_j (\mathrm{span}\{\boldsymbol{\theta}_j\} + \mathbb{B})}$$

Proof sketch, III: Covering the space and Curse of Dimensionality

Frequency components of $\widehat{f} \star \widehat{\phi}$ for $f \in \text{span}(\mathcal{G}_\sigma^n)$ have a peculiar structure: **union of unit width tubes**.

Full frequency support: m large enough s.t. $\text{supp}(\widehat{f} \star \widehat{\phi}) \supseteq r\mathbb{B}$ as r grows.

Because: if $\text{supp}(\widehat{f} \star \widehat{\phi}) \not\supseteq r\mathbb{B} \implies \exists\, \boldsymbol{\omega} \in r\mathbb{B}$ representing oscillations that $\widehat{f} \star \widehat{\phi}$ cannot capture.

In fact one can show the following volume ratio formula as $n \to \infty$

$$\frac{\mathbb{V}\left(\text{supp}(\widehat{f} \star \widehat{\phi}) \cap r\mathbb{B}\right)}{\mathbb{V}(r\mathbb{B})} \lessapprox m e^{-n}$$

Designing the Target

Target: Radial function $g = \psi \circ \| \cdot \|$

Construction is technically involved. High level idea: random sign indicator functions of thin shells.

Assume $\|\mathbf{x}\| \leq R$ and chose an N-partition $\{\Delta_i\}$ of $[0; R]$. Then define

$$\psi(z) = \sum_{i=1}^{N} \epsilon_i \psi_i(z), \quad \psi_i(z) = \mathbf{1}\{\Delta_i\}, \quad \epsilon_i \in \{-1, 1\}$$

- Sign flips generate oscillations

Designing the Target

Theorem

Theorem 9 (Eldan & Shamir, 2016)

For $n \geq C$ there exists a probability measure μ with density ϕ^2 and a function g with the following properties:

1. *g is bounded in $[-2;2]$ supported on $\{\mathbf{x} : \|\mathbf{x}\| \leq C\sqrt{n}\}$ and expressible by a 2 hidden layer network with width $Ccn^{19/4}$.*
2. *Every function f implemented by a one-hidden layer network with width $m \leq ce^{cn}$ satisfies*

$$\mathbf{E}_{\mathbf{x}\sim\mu}\left(f(\mathbf{x}) - g(\mathbf{x})\right)^2 \geq c$$

Gilles Pisier.
Remarques sur un résultat non publié de b. maurey.
Séminaire Analyse fonctionnelle (dit, pages 1–12, 1981.

Matus Telgarsky.
Benefits of depth in neural networks.
In **Conference on learning theory**, pages 1517–1539. PMLR, 2016.

Matus Telgarsky.
Deep learning theory lecture notes.
`https://mjt.cs.illinois.edu/dlt/`, 2021.
Version: 2021-10-27 v0.0-e7150f2d (alpha).