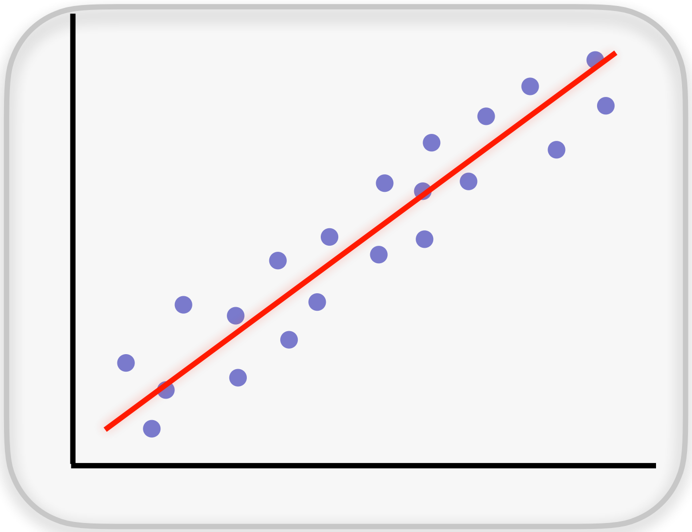
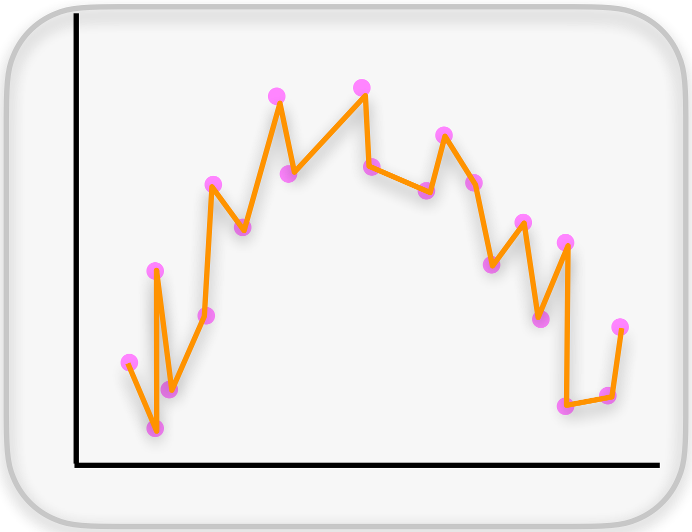
Approximation Theory

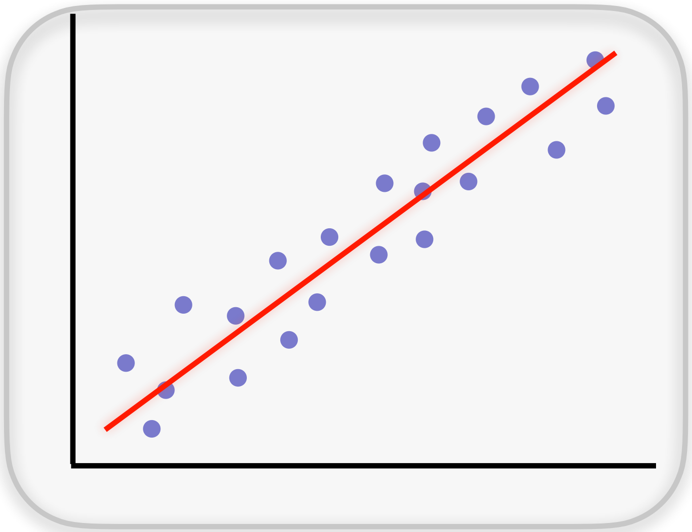
Expressivity / Representation Power





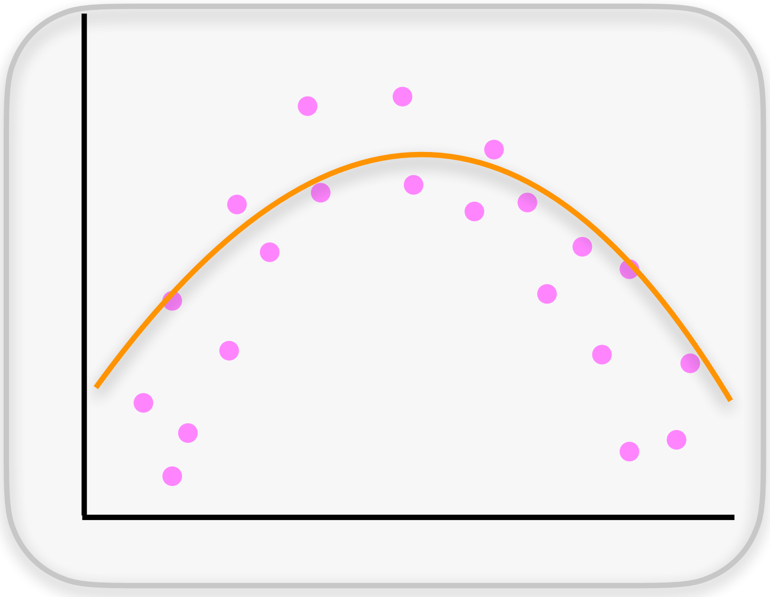
Expressive: Functions in class can represent "complicated" functions.

Linear Function



best linear fit

Review: generalized linear regression



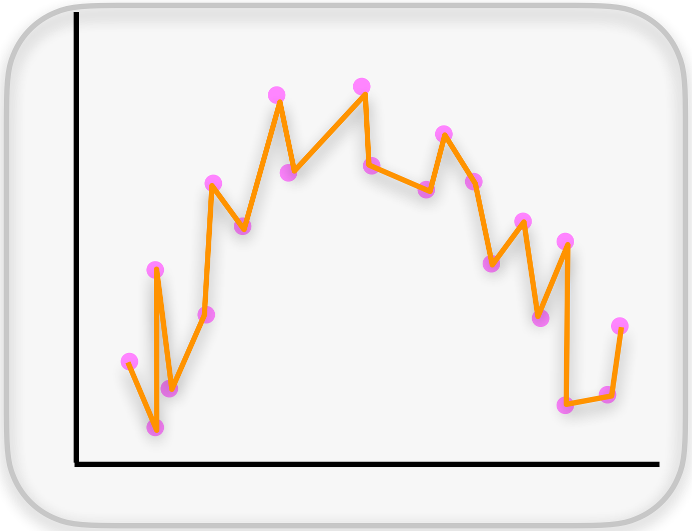
Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w$$

Review: Polynomial Regression



Approximation Theory Setup

- Goal: to show there exists a neural network that has small error on training / test set.

- Set up a natural baseline:

$$\inf_{f \in \mathcal{F}} L(f) \text{ v.s. } \inf_{g \in \text{ continuous functions}} L(g)$$

Example

Decomposition

Specific Setups

- "Average" approximation: given a distribution μ

$$\|f-g\|_{\mu} = \int_{x} |f(x)-g(x)|\, d\mu(x)$$

- "Everywhere" approximation

$$\|f-g\|_{\infty} = \sup_{x} |f(x)-g(x)| \geq \|f-g\|_{\mu}$$

Polynomial Approximation

Theorem (Stone-Weierstrass): for any function f, we can approximate it on any compact set Ω by a sufficiently high degree polynomial: for any $\epsilon > 0$, there exists a polynomial p of sufficient high degree, s.t.,

$$\max_{x \in \Omega} | f(x) - p(x) | \leq \epsilon.$$

Intuition: Taylor expansion!

Kernel Method

Polynomial kernel

Gaussian Kernel

1D Approximation

Theorem: Let $g : [0,1] \to R$, and ρ-Lipschitz. For any $\epsilon > 0, \exists$ 2-layer neural network f with $\lceil \frac{\rho}{\epsilon} \rceil$ nodes, threshold activation: $\sigma(z) : z \mapsto \mathbf{1}\{z \geq 0\}$ such that

$$\sup_{x \in [0,1]} |f(x) - g(x)| \leq \epsilon.$$

Proof of 1D Approximation

Multivariate Approximation

Theorem: Let g be a continuous function that satisfies $\|x - x'\|_\infty \leq \delta \Rightarrow |g(x) - g(x')| \leq \epsilon$ (Lipschitzness). Then there exists a 3-layer ReLU neural network with $O(\frac{1}{\delta^d})$ nodes that satisfy

$$\int_{[0,1]^d} |f(x) - g(x)| dx = \|f - g\|_1 \leq \epsilon$$

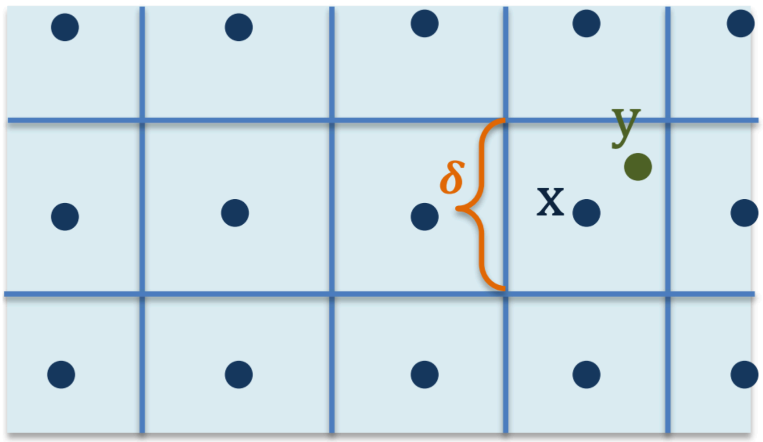
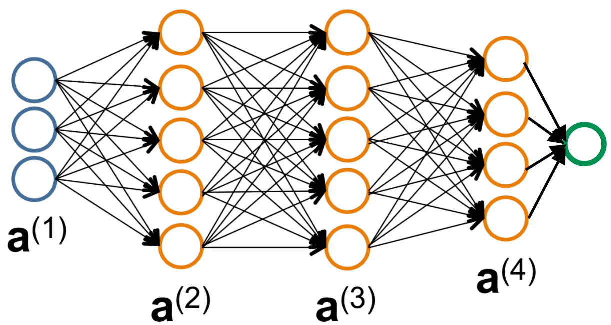


Figure credit to Andrej Risteski



Partition Lemma

Lemma: let g, δ, ϵ be given. For any partition P of $[0,1]^d$, $P = (R_1, \ldots, R_N)$ with all side length smaller than δ, there exists $(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ such that

$$\sup_{x \in [0,1]^d} |g(x) - h(x)| \leq \epsilon \text{ with } h(x) := \sum_{i=1}^{N} \alpha_i \mathbf{1}_{R_i}(x).$$

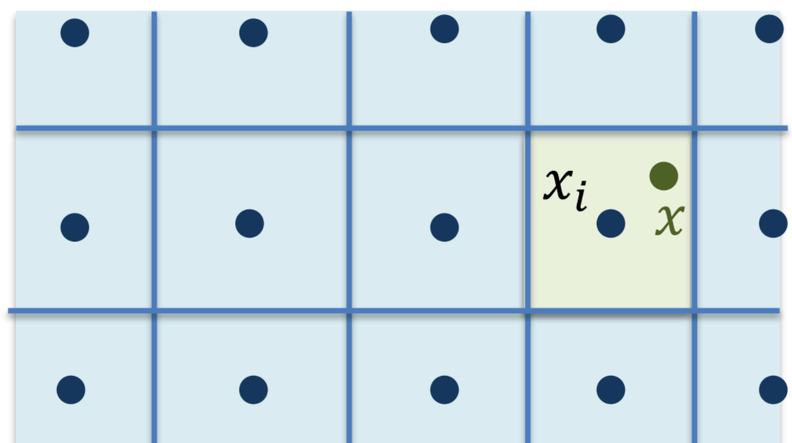


Figure credit to Andrej Risteski

Proof of Partition Lemma

Proof of Multivariate Approximation Theorem

Proof of Multivariate Approximation Theorem

Proof of Multivariate Approximation Theorem

Universal Approximation

Definition: A class of functions $\mathscr{F}$ is universal approximator over a compact set S (e.g., $[0,1]^d$), if for every continuous function g and a target accuracy $\epsilon > 0$, there exists $f \in \mathscr{F}$ such that

$$\sup_{x \in S} |f(x) - g(x)| \leq \epsilon$$

Stone-Weierstrass Theorem

Theorem: If $\mathscr{F}$ satisfies

1. Each $f \in \mathscr{F}$ is continuous.
2. $\forall x, \exists f \in \mathscr{F}, f(x) \neq 0$
3. $\forall x \neq x', \exists f \in \mathscr{F}, f(x) \neq f(x')$
4. $\mathscr{F}$ is closed under multiplication and vector space operations,

Then $\mathscr{F}$ is a universal approximator:

$$\forall g : S \to R, \epsilon > 0, \exists f \in \mathscr{F}, \|f - g\|_\infty \leq \epsilon.$$

Example: cos activation

Example: cos activation

Other Examples

Exponential activation

ReLU activation

Curse of Dimensionality

- Unavoidable in the worse case

- Barron's theory

Recent Advances in Representation Power

- Depth separation
- Analyses of different architectures
  - Graph neural network
  - Attention-based neural network
- Finite data approximation
- …