Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

Breaking the curse of dimensionality
with Barron Spaces

Antonio Álvarez López

Universidad Autónoma de Madrid, Department of Mathematics

November 15, 2022

1. Supervised Learning
1.1 Problem statement
1.2 Main paradigms

2. Shallow Neural Networks
2.1 Structure
2.2 Approximation Capabilities

3. Barron Spaces
3.1 Curse of dimensionality
3.2 Definition and properties

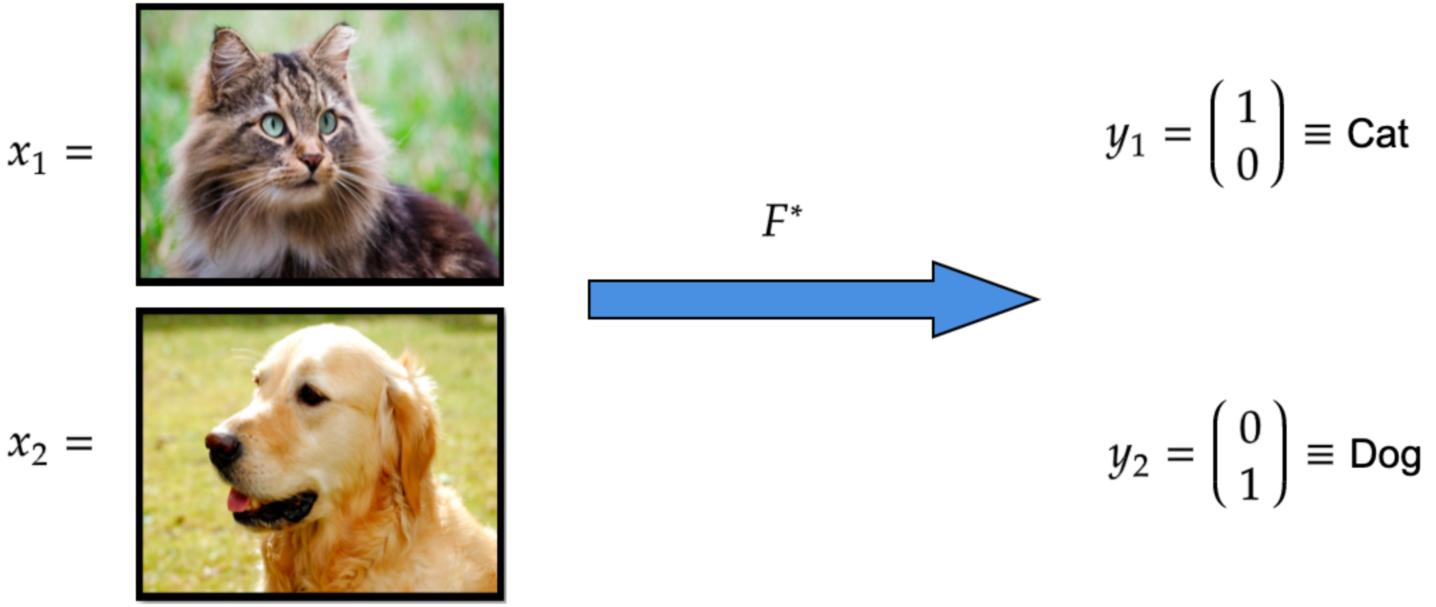
Supervised Learning

Problem statement

- *Input space* $X \subset \mathbb{R}^d$, *output space* $Y \subset \mathbb{R}^m$.
- Given dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset X \times Y$,

Goal

Approximating an ideal unknown *target function* F^* that can label any input $x \in X$ to its corresponding label $y \in Y$, using only the information contained in the dataset $\mathcal{D}$, which verifies $y_i = F^*(x_i)$ for $i = 1, \dots N$.



Supervised Learning

Learning procedure

- We construct, using $\mathcal{D}$, a predictive model $\hat{F}$ from a chosen parametric class of functions that we call the *hypothesis space* $\mathcal{H}$.
- The best possible predictor in $\mathcal{H}$ would be ideally obtained through *population risk minimization*:

$$\arg\min_{F \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mu^*} L(F(x), y),$$

where μ^* is the unknown input-output distribution and $L(\cdot, \cdot)$ is a suitable *loss function*.

Supervised Learning

Learning procedure

- We construct, using $\mathcal{D}$, a predictive model $\hat{F}$ from a chosen parametric class of functions that we call the *hypothesis space* $\mathcal{H}$.
- The best possible predictor in $\mathcal{H}$ would be ideally obtained through *population risk minimization*:

$$\arg\min_{F\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu^*} L(F(x), y),$$

  where μ^* is the unknown input-output distribution and $L(\cdot,\cdot)$ is a suitable *loss function*.
- In practice, the predictor $\hat{F}$ constructed is obtained through *empirical risk minimization*:

$$\hat{F} = \arg\min_{F\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} L(F(x_i), \underbrace{y_i}_{=F^*(x_i)}),$$

Supervised Learning

Learning procedure

- We construct, using $\mathcal{D}$, a predictive model $\hat{F}$ from a chosen parametric class of functions that we call the *hypothesis space* $\mathcal{H}$.
- The best possible predictor in $\mathcal{H}$ would be ideally obtained through *population risk minimization*:

$$\arg\min_{F\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mu^*} L(F(x), y),$$

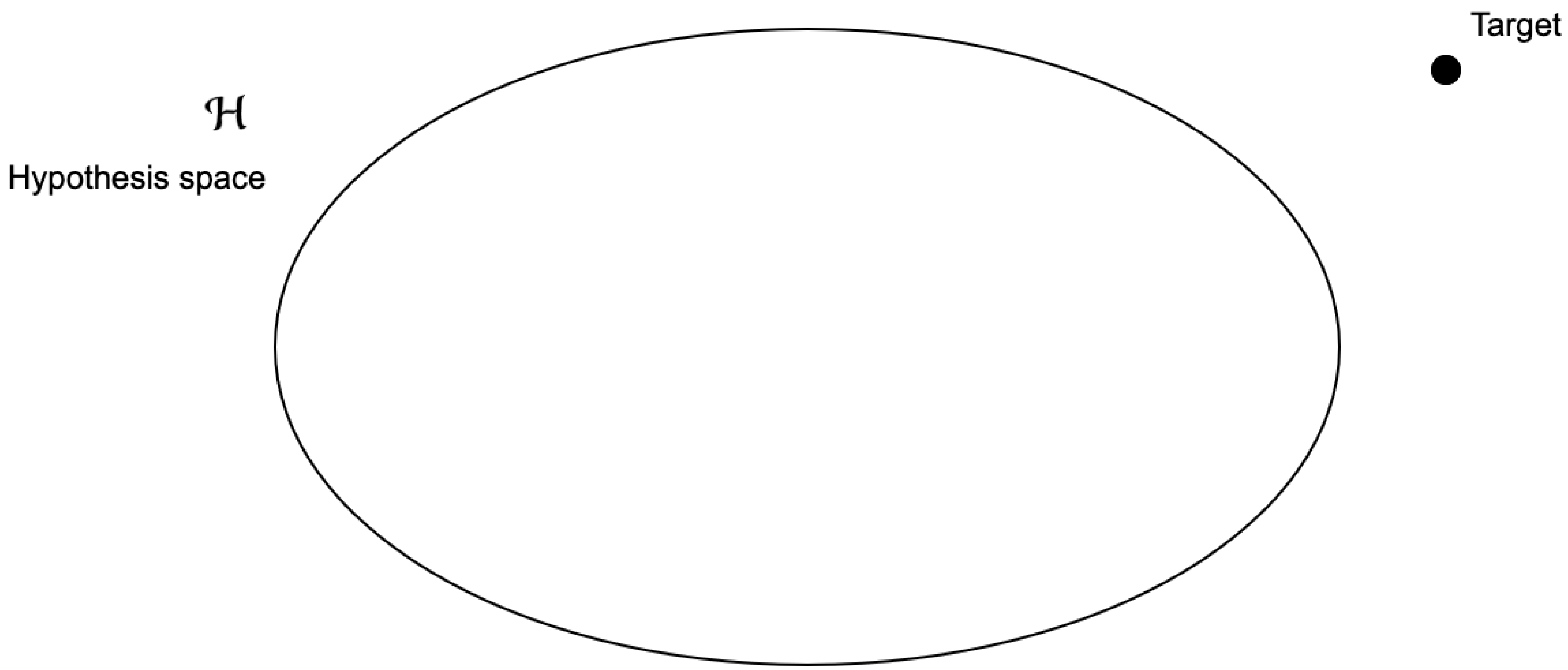
  where μ^* is the unknown input-output distribution and $L(\cdot,\cdot)$ is a suitable *loss function*.
- In practice, the predictor $\hat{F}$ constructed is obtained through *empirical risk minimization*:

$$\hat{F} = \arg\min_{F\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} L(F(x_i), \underbrace{y_i}_{=F^*(x_i)}),$$

- Main paradigms:
  - **Approximation**: How close is our hypothesis space $\mathcal{H}$ of any target function F^*?
  - **Optimization**: How can we find or get close to the best possible approximation $\hat{F} \in \mathcal{H}$ of F^*?
  - **Generalization**: Can the constructed predictor $\hat{F}$ generalize well to unseen examples?

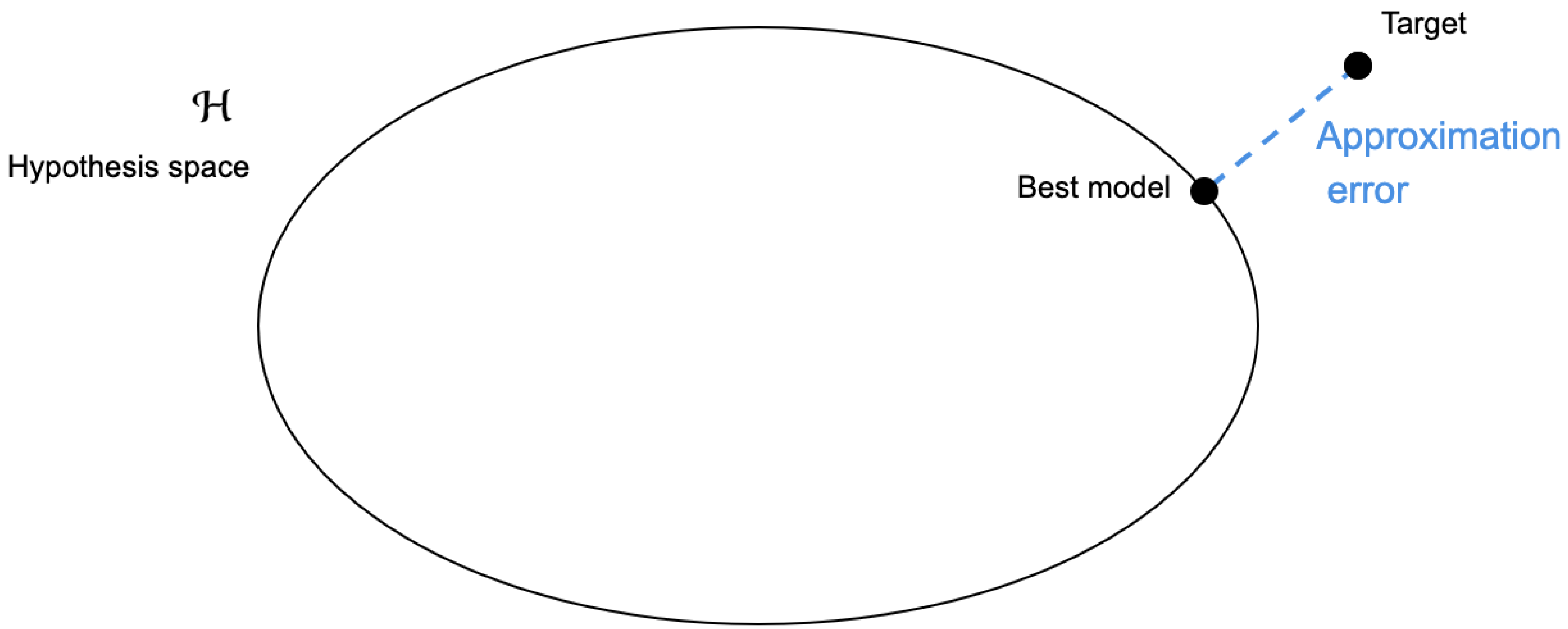
Supervised Learning

Main paradigms



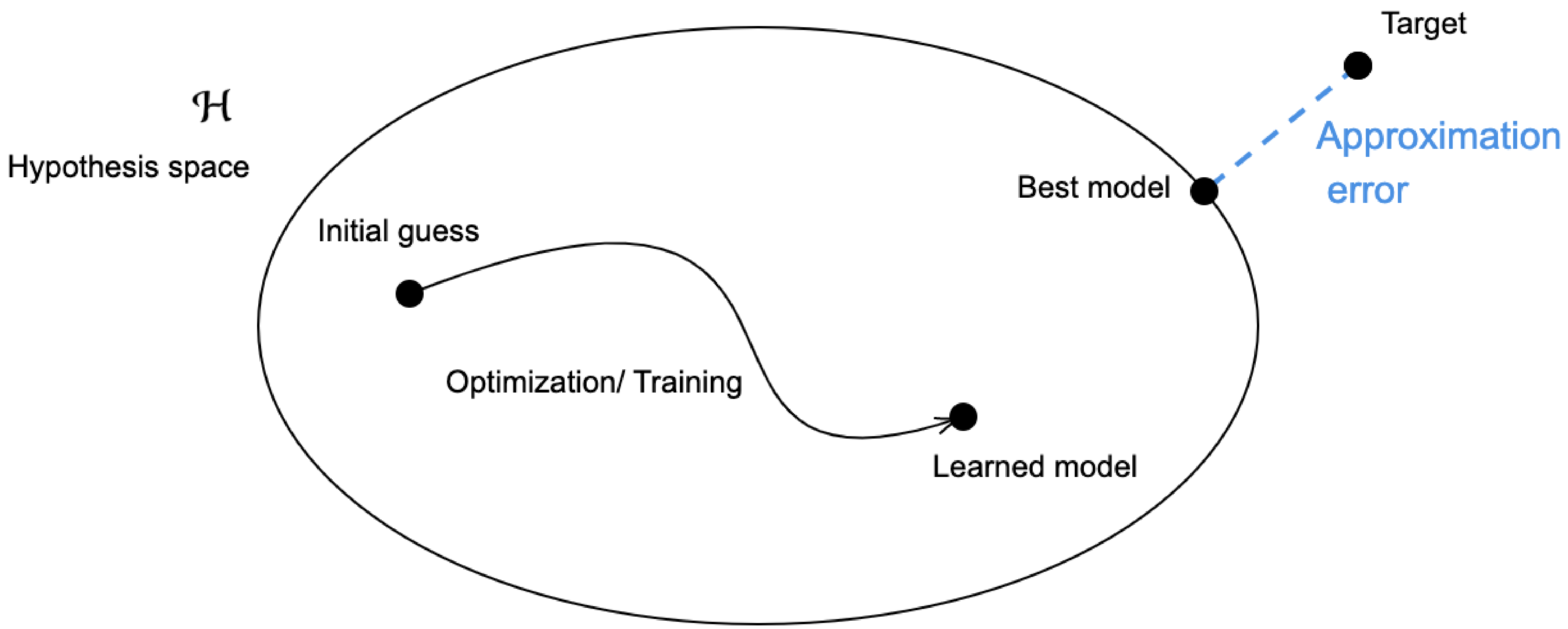
Supervised Learning

Main paradigms



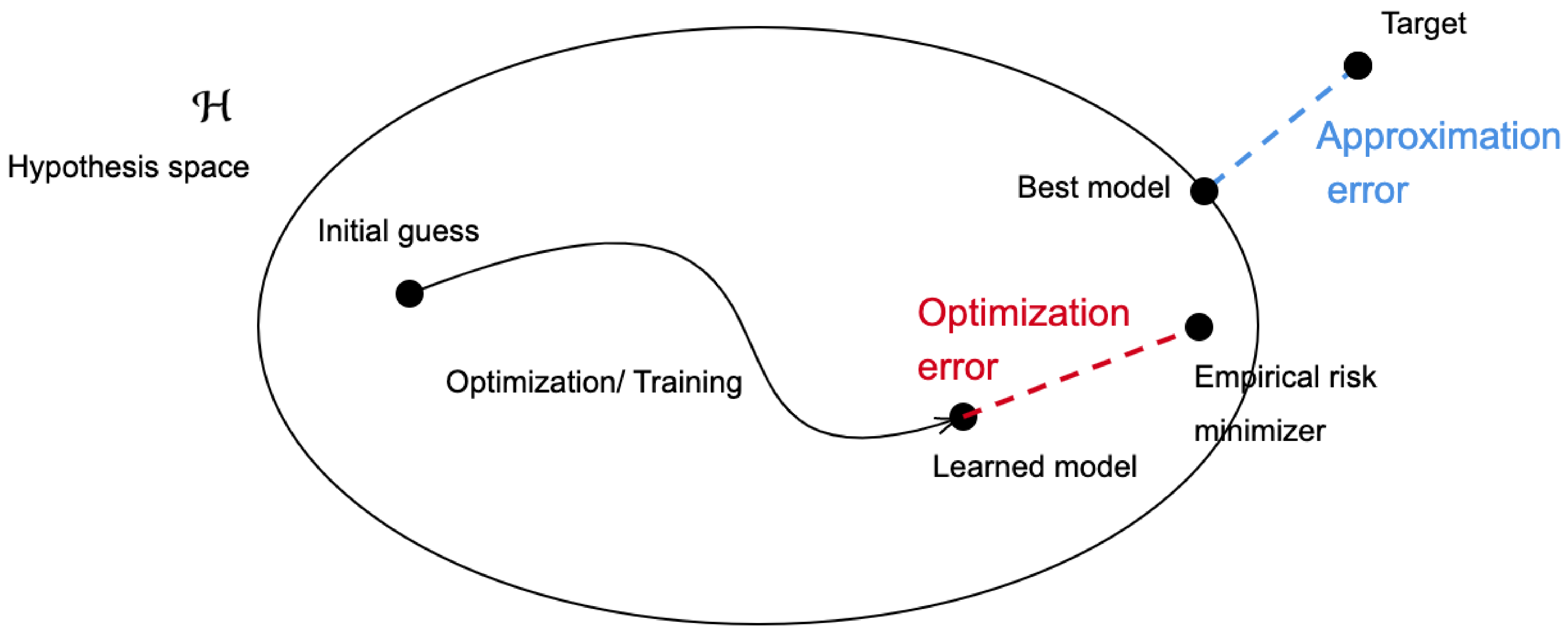
Supervised Learning

Main paradigms



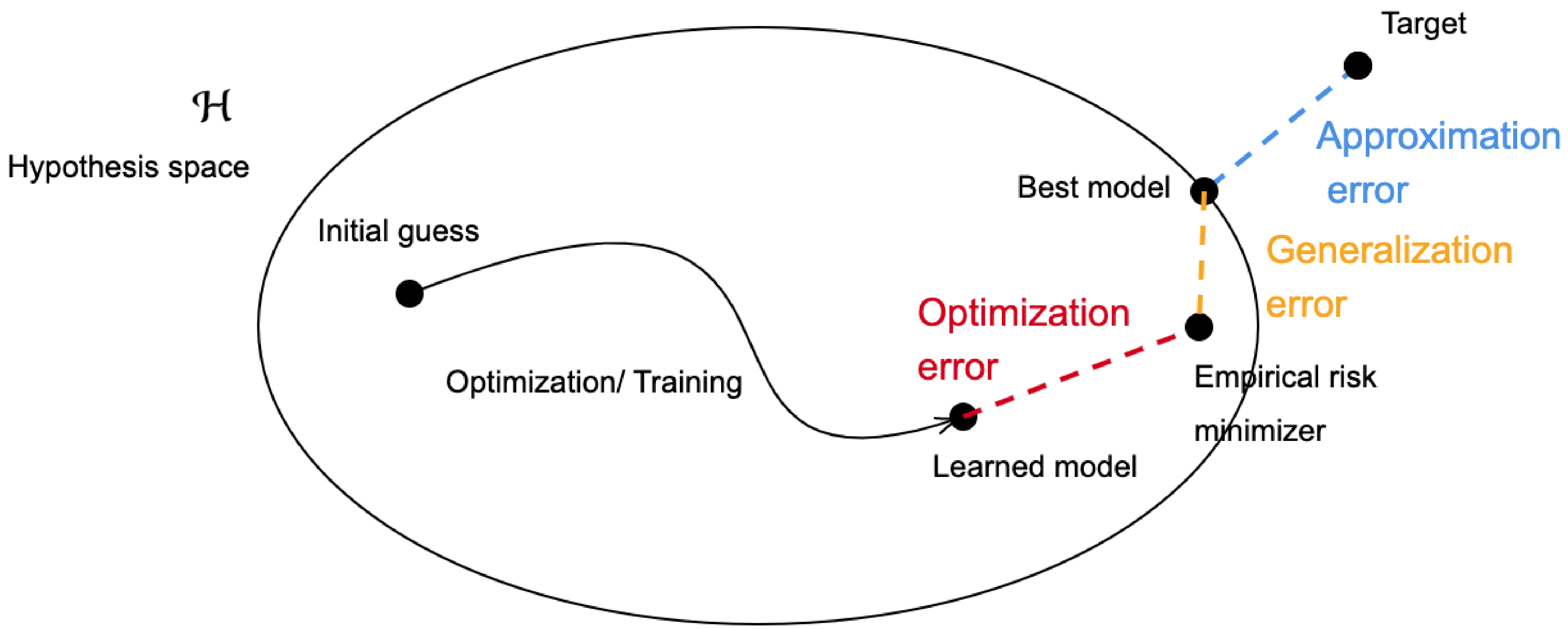
Supervised Learning

Main paradigms



Supervised Learning

Main paradigms



Supervised Learning

Examples of hypothesis spaces

- *Linear models*: $\mathcal{H} = \left\{ F : \mathbb{R}^d \to \mathbb{R} \;\middle|\; F(\mathbf{x}) = \sum_{i=0}^{M-1} w_i \phi_i(\mathbf{x}), w_i \in \mathbb{R} \right\}$, where $\phi_i : \mathbb{R}^d \to \mathbb{R}$ are a prefixed set of *basis functions* or *feature maps*:

$$\phi_j(\mathbf{x}) = \mathbf{x}^j, \qquad \phi_j(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{m_j})^2}{2d^2}\right), \qquad \phi_j(\mathbf{x}) = \text{sigm}\left(\frac{\mathbf{x}-\mathbf{m_j}}{d}\right), \text{ with } \text{sigm}(b) = \frac{1}{1+e^{-b}}.$$

Supervised Learning

Examples of hypothesis spaces

- *Linear models*: $\mathcal{H} = \left\{ F : \mathbb{R}^d \to \mathbb{R} \;\middle|\; F(\mathbf{x}) = \sum_{i=0}^{M-1} w_i \phi_i(\mathbf{x}), w_i \in \mathbb{R} \right\}$, where $\phi_i : \mathbb{R}^d \to \mathbb{R}$ are a prefixed set of *basis functions* or *feature maps*:

$$\phi_j(\mathbf{x}) = \mathbf{x}^j, \qquad \phi_j(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{m_j})^2}{2d^2}\right), \qquad \phi_j(\mathbf{x}) = \text{sigm}\left(\frac{\mathbf{x}-\mathbf{m_j}}{d}\right), \text{ with sigm}(b) = \frac{1}{1+e^{-b}}.$$

- *Shallow Neural Networks (SNNs)*: $\mathcal{H} = \left\{ F_M : F_M(\mathbf{x}) = \sum_{i=1}^{M} w_i \sigma(\mathbf{a_i}^T \cdot \mathbf{x} + b_i),\; w_i \in \mathbb{R}, \mathbf{a_i} \in \mathbb{R}^d, b_i \in \mathbb{R}, M \in \mathbb{N} \right\}$.
  - σ is the *activation function*.

$$\text{ReLU: } \sigma(z) = \max(0, z), \qquad \sigma(z) = \tanh(z), \qquad \sigma(z) = \text{sigm}(z), \ldots$$

  - M is the *width*, which controls the complexity of the model.

Supervised Learning

Examples of hypothesis spaces

- *Linear models*: $\mathcal{H} = \left\{ F : \mathbb{R}^d \to \mathbb{R} \;\middle|\; F(\mathbf{x}) = \sum_{i=0}^{M-1} w_i \phi_i(\mathbf{x}), w_i \in \mathbb{R} \right\}$, where $\phi_i : \mathbb{R}^d \to \mathbb{R}$ are a prefixed set of *basis functions* or *feature maps*:

$$\phi_j(\mathbf{x}) = \mathbf{x}^j, \qquad \phi_j(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{m_j})^2}{2d^2}\right), \qquad \phi_j(\mathbf{x}) = \text{sigm}\left(\frac{\mathbf{x}-\mathbf{m_j}}{d}\right), \text{ with sigm}(b) = \frac{1}{1+e^{-b}}.$$

- *Shallow Neural Networks (SNNs)*: $\mathcal{H} = \left\{ F_M : F_M(\mathbf{x}) = \sum_{i=1}^{M} w_i \sigma(\mathbf{a_i}^T \cdot \mathbf{x} + b_i),\; w_i \in \mathbb{R}, \mathbf{a_i} \in \mathbb{R}^d, b_i \in \mathbb{R}, M \in \mathbb{N} \right\}$.
  - σ is the *activation function*.

$$\text{ReLU: } \sigma(z) = \max(0,z), \qquad \sigma(z) = \tanh(z), \qquad \sigma(z) = \text{sigm}(z), \ldots$$

  - M is the *width*, which controls the complexity of the model.
- *Deep Neural Networks* (DNNs):
  - *Multilayer Perceptron* (MLP) of *depth* K: $\mathcal{H}_K = \left\{ F_K : F_K(\mathbf{x}) = \mathbf{w^T x}(K),\; \mathbf{w} \in \mathbb{R}^{d_K} \right\}$, with
    - $\mathbf{x}(k+1) = \mathbf{w}(k)\sigma\left(\mathbf{a}(k)^T \mathbf{x}(k) + b(k)\right), \qquad \mathbf{w}(k) \in \mathbb{R}^{d_{k+1}}, \qquad \mathbf{a}(k) \in \mathbb{R}^{d_k}, \qquad b(k) \in \mathbb{R}, \qquad k = 0, \ldots, K-1.$
    - $d_k \in \mathbb{N}$ for all k, and $d_0 = d$, $\mathbf{x}(0) = \mathbf{x}$.
  - *Residual Networks* (ResNets): take MLP redefining $\mathbf{x}(k+1) = \mathbf{x}(k) + \mathbf{w}(\mathbf{k})\sigma\left(\mathbf{a}(k)^T \mathbf{x}(k) + b(k)\right)$.

1. Supervised Learning

1.1 Problem statement

1.2 Main paradigms

2. Shallow Neural Networks

2.1 Structure

2.2 Approximation Capabilities

3. Barron Spaces

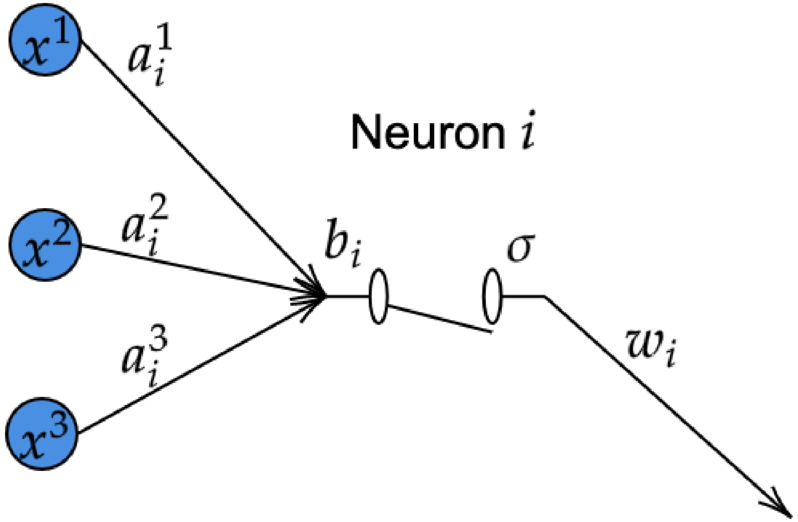
3.1 Curse of dimensionality

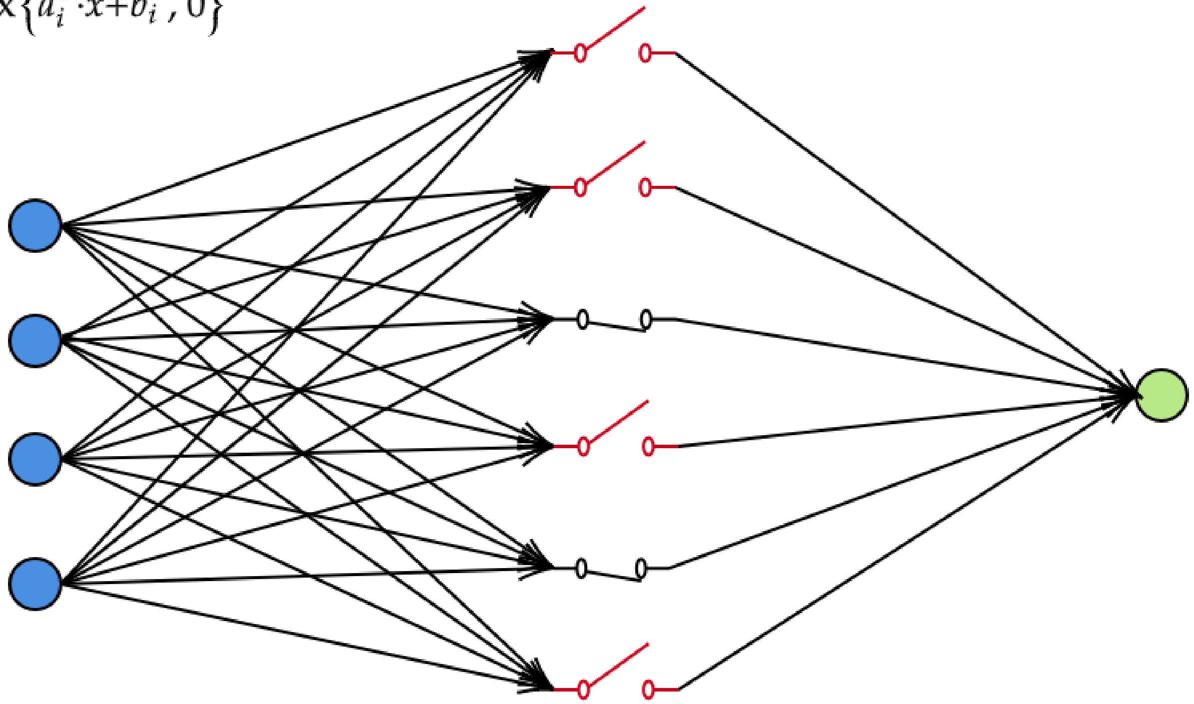
3.2 Definition and properties

Shallow Neural Networks

Structure

$$\sigma \equiv ReLU: \quad x \in \mathbb{R}^d \quad \mapsto \quad \Big[\underbrace{\sigma\big(a_i^T \cdot x + b_i\big)}_{=\max\left\{a_i^T \cdot x + b_i\,,\, 0\right\}} \Big]_{i=1}^{m} \quad \mapsto \quad F(x) = \sum_{i=1}^{m} w_i \sigma\big(a_i^T \cdot x + b_i\big) \in \mathbb{R}$$





Shallow Neural Networks

Approximation Properties

Universal Approximation Theorem for SNNs [Cyb89]

Let $\Omega \subset \mathbb{R}^d$ be a compact set, and $F^* \in C(\Omega)$. Assume that the activation function σ is continuous and *sigmoidal*, i.e. $\lim_{z\to\infty} \sigma(z) = 1$, $\lim_{z\to-\infty} \sigma(z) = 0$. Then, for every $\epsilon > 0$ there exists $F_M \in \mathcal{H}$ such that

$$\|F_M - F^*\|_{C(K)} = \max_{\mathbf{x}\in K} |F_M(\mathbf{x}) - F^*(\mathbf{x})| < \epsilon. \tag{1}$$

Universal Approximation Theorem for SNNs [Pin99]

Let $\Omega \subset \mathbb{R}^d$ be a compact set and σ a continuous activation function. Then, $\mathcal{H}$ is dense in $C(\Omega)$ in the topology of uniform convergence if and only if σ is non-polynomial.

Also for Deep Neural Networks:

Universal Approximation Theorem for Deep-Narrow NNs [KL20]

Let $\Omega \subset \mathbb{R}^d$ be a compact set and σ a nonaffine continuous activation function. Assume further that σ is continuously differentiable with nonzero derivative at least at one point. Then, $\mathcal{H}_K$ is dense in $C(K; \mathbb{R}^{d_K})$ in the topology of uniform convergence if $K = d + d_K + 2$.

1. Supervised Learning
1.1 Problem statement
1.2 Main paradigms

2. Shallow Neural Networks
2.1 Structure
2.2 Approximation Capabilities

3. Barron Spaces
3.1 Curse of dimensionality
3.2 Definition and properties

Necessity: Curse of dimensionality

- Is the space of continuous functions enough?

Necessity: Curse of dimensionality

- Is the space of continuous functions enough?
- No $\rightarrow$ *Curse of dimensionality (CoD)*: Complexity of the models required for better estimates increases exponentially with the dimension of the ambient space

$$\|F_M - F^*\|_{L^p(\Omega)} \leq C \frac{\|F^*\|_{W^{s,p}}}{M^{\alpha(s)/d}}.$$

We need $M > \epsilon^{-d}$ to achieve an approximation error of ϵ.

Necessity: Curse of dimensionality

- Is the space of continuous functions enough?
- No $\rightarrow$ *Curse of dimensionality (CoD)*: Complexity of the models required for better estimates increases exponentially with the dimension of the ambient space

$$\|F_M - F^*\|_{L^p(\Omega)} \leq C\frac{\|F^*\|_{W^{s,p}}}{M^{\alpha(s)/d}}.$$

  We need $M > \epsilon^{-d}$ to achieve an approximation error of ϵ.
- The curse of dimensionality is intrinsic for high dimensional spaces.
- The model works *without* curse of dimensionality when the complexity depends at most polynomially on d for fixed ϵ.
- We can only avoid it by considering a smaller set of problems $\rightarrow$ Find the “right” space of target functions to approximate.

Example

Classical numerical analysis

- Theory of splines and theory of finite element methods approximate functions using piecewise polynomials.
- One starts from a function that lies in a Sobolev/Besov space and proceeds to derive optimal error estimates.
- These estimates depend on the function norm and the regularity encoded in the function space, as well as the approximation scheme.
- Sobolev/Besov spaces are the right ones for these classical theories:
  - Direct and inverse approximation theorems: a function can be approximated by piecewise polynomials with certain convergence rate if and only if the function is in a certain Sobolev/Besov space
  - The functions that we are interested in (e.g. solutions of PDEs) lie on these spaces.

Barron Spaces

Definition

- Let $\Omega \subset \mathbb{R}^d$ be a compact set. We will work with $\sigma \equiv$ ReLU.
- Consider functions $f : \Omega \to \mathbb{R}$ that admit the representation

$$f(\mathbf{x}) = \int_{\Theta} w\sigma(\mathbf{a}^T\mathbf{x} + c)\rho(dw, \mathbf{da}, db) = \mathbb{E}_{\rho}[w\sigma(\mathbf{a}^T\mathbf{x} + b)], \qquad \mathbf{x} \in \Omega, \tag{2}$$

where $\Theta = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ space of parameters and ρ is a probability distribution on $(\Theta, \Sigma_{\Theta})$, being Σ_{Θ} a Borel σ-algebra on Θ.

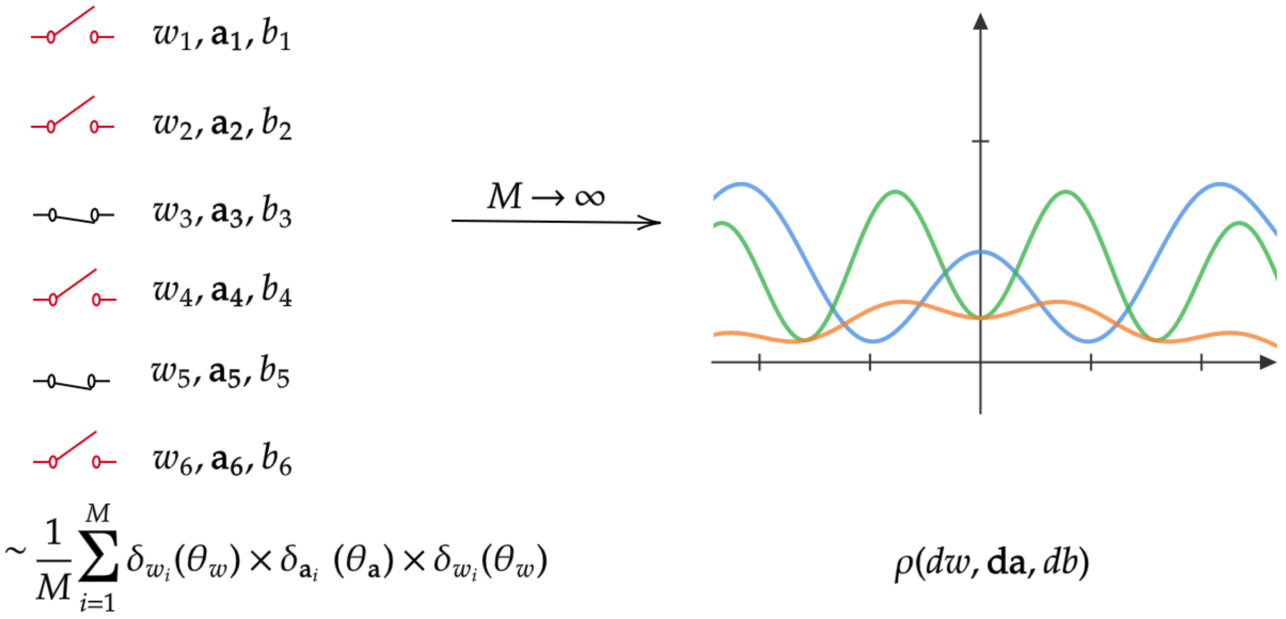
Barron Spaces

Definition

- Let $\Omega \subset \mathbb{R}^d$ be a compact set. We will work with $\sigma \equiv$ ReLU.
- Consider functions $f : \Omega \to \mathbb{R}$ that admit the representation

$$f(\mathbf{x}) = \int_{\Theta} w\sigma(\mathbf{a}^T\mathbf{x} + c)\rho(dw, \mathbf{da}, db) = \mathbb{E}_\rho[w\sigma(\mathbf{a}^T\mathbf{x} + b)], \qquad \mathbf{x} \in \Omega, \tag{2}$$

where $\Theta = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ space of parameters and ρ is a probability distribution on (Θ, Σ_Θ), being Σ_Θ a Borel σ-algebra on Θ.



Barron Spaces

Definition

- In general, the ρ's for which (2) holds are not unique. For a function that admits this representation, we define its *Barron norm*
$$\|f\|_{\mathcal{B}_p} = \inf_{\rho} \left(\mathbb{E}_{\rho}[|w|^p(\|\mathbf{a}\|_1 + |b|)^p]\right)^{1/p}, \qquad 1 \leq p \leq \infty, \tag{3}$$
where the infimum is taken over all ρ for which (2) holds for all $\mathbf{x} \in \Omega$.
- Barron spaces $\mathcal{B}_p$ are defined as
$$\{f \in C(\Omega) : f \text{ admits a representation (2)}, \|f\|_{\mathcal{B}_p} < \infty\}.$$

Barron Spaces

Definition

- In general, the ρ's for which (2) holds are not unique. For a function that admits this representation, we define its *Barron norm*

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho} \left(\mathbb{E}_\rho[|w|^p(\|\mathbf{a}\|_1 + |b|)^p]\right)^{1/p}, \qquad 1 \leq p \leq \infty, \tag{3}$$

  where the infimum is taken over all ρ for which (2) holds for all $\mathbf{x} \in \Omega$.
- Barron spaces $\mathcal{B}_p$ are defined as

$$\{f \in C(\Omega) : f \text{ admits a representation (2)}, \|f\|_{\mathcal{B}_p} < \infty\}.$$

- By Hölder's inequality, we have

$$\mathcal{B}_\infty \subset \cdots \subset \mathcal{B}_2 \subset \mathcal{B}_1 :$$

- The opposite is also true:

Proposition [MW+22]

For any $f \in \mathcal{B}_1$, we have $f \in \mathcal{B}_\infty$ and

$$\|f\|_{\mathcal{B}_1} = \|f\|_{\mathcal{B}_\infty}.$$

- As a consequence, there is just one Barron space and one Barron norm that we denote by $\mathcal{B}$ and $\|\cdot\|_{\mathcal{B}}$, respectively.

What functions belong to $\mathcal{B}$?

Early study

- Recall:
  - *Fourier transform* of $f : \mathbb{R}^d \to \mathbb{R}$:
  $$\hat{f}(\xi) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-i\xi\cdot\mathbf{x}} \mathbf{dx}.$$
  - *Fourier inversion formula*:
  $$f(\mathbf{x}) = \int_{\mathbb{R}^d} \hat{f}(\mathbf{x}) e^{i\xi\cdot\mathbf{x}} d\xi.$$
  - $\widehat{Df}(\xi) = i\xi\hat{f}(\xi)$

What functions belong to $\mathcal{B}$?

Early study

- Recall:
  - *Fourier transform* of $f : \mathbb{R}^d \to \mathbb{R}$:
$$\hat{f}(\xi) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-i\xi\cdot\mathbf{x}} \mathbf{dx}.$$
  - *Fourier inversion formula*:
$$f(\mathbf{x}) = \int_{\mathbb{R}^d} \hat{f}(\mathbf{x}) e^{i\xi\cdot\mathbf{x}} d\xi.$$
  - $\widehat{Df}(\xi) = i\xi\hat{f}(\xi)$

Barron's Theorem [Bar93]

For a function $F^* : \Omega \to \mathbb{R}$, let $\hat{F}^*$ be the Fourier transform of any extension of F^* to $\mathbb{R}^d$. Then, if

$$\gamma(F^*) := \inf_{\hat{F}} \int_{\mathbb{R}^d} \|\xi\|_1^2 |\hat{F}^*(\xi)| d\xi = \|\widehat{D^2 F^*}\|_1 < +\infty,$$

for any $M > 0$ there exists a SNN $F_M(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^M w_i \sigma(\mathbf{a_i}^T\mathbf{x} + b_i)$ satisfying

$$\|F_M - F^*\|_{L^2(\Omega)}^2 \leq \frac{3\gamma(F^*)^2}{M},$$

and $\sum_{i=1}^M |w_i|(\|\mathbf{a_i}\|_1 + |b_i|) \leq 2\gamma(F^*)$.

What functions belong to $\mathcal{B}$?

Theorem [Bar93]

Let $F^* \in C(\Omega)$ and assume that F^* satisfies $\gamma(F^*) < \infty$. Then F^* admits an integral representation (2). Moreover,

$$\|F^*\|_{\mathcal{B}} \leq 2\gamma(F^*) + 2\|\nabla F^*(0)\|_1 + 2|F^*(0)|.$$

- To achieve $\gamma(F^*) < \infty$:
  - Necessary condition: All first order partial derivatives are bounded.
  - Sufficient condition: All partial derivatives of order less or equal than s belong to $L^2(\mathbb{R}^d)$, being $s = \lceil 1 + d/2 \rceil$.
- Not enough to generally avoid CoD ($\gamma(F^*)$ involves a $d-$dimensional integral), but there are many examples for which $\gamma(F^*)$ is only moderately large, e.g., $O(d)$ or $O(d^2)$.

Corollary

All **gaussian** functions, **positive definite** functions, **linear** functions and **radial** functions belong to $\mathcal{B}$.

Theorem of direct approximation

- Define the *path norm* as

$$\|\theta\|_{\mathcal{P}} := \frac{1}{M}\sum_{i=1}^{M} |w_i|(\|\mathbf{a_i}\|_1 + |b_i|),$$

where θ denotes a specific set of parameters $\{(w_i, \mathbf{a}_i, b_i)\}_{i=1}^{M}$.

Theorem of direct approximation [MW+22]

For any $F^* \in \mathcal{B}$ and $M > 0$, there exists a SNN $F_M(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} w_i\sigma(\mathbf{a_i}^T\mathbf{x} + b_i)$ satisfying

$$\|F_M(\cdot;\theta) - F^*(\cdot)\|_{L^2(\Omega)}^2 \leq \frac{3\|F^*\|_{\mathcal{B}}^2}{M}.$$

Furthermore, we have $\|\theta\|_{\mathcal{P}} \leq 2\|F^*\|_{\mathcal{B}}$.

- $\mathcal{B}$ can be seen as the closure of $\mathcal{H}$ with respect to the path norm.

Inverse approximation theorem

- Define $\mathcal{N}_Q := \{F_M(\mathbf{x};\theta) : \ \|\theta\|_{\mathcal{P}} \leq Q, m \in \mathbb{N}^+\}$.

Theorem of inverse approximation [MW+22]

Let $F^* \in C(\Omega)$. Assume there exists a constant Q and a sequence of functions $(F_M) \subset \mathcal{N}_Q$ such that

$$\lim_{M\to\infty} F_M(\mathbf{x}) = F^*(\mathbf{x}),$$

for all $\mathbf{x} \in \Omega$. Then $F^* \in \mathcal{B}$ and $\|F^*\|_{\mathcal{B}} \leq Q$.

- Idea of the proof:

  Assume $F_M(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^M w_i^{(M)}\sigma\left(\mathbf{a}_i^{(M)}\mathbf{x} + b_i^{(M)}\right)$. The Theorem's hypothesis implies that the sequence (ρ_M) defined by

$$\rho_M(w,\mathbf{a},b) = \frac{1}{M}\sum_{i=1}^M \delta(w - w_i^{(M)})\delta(\mathbf{a} - \mathbf{a}_i^{(M)})\delta(b - b_i^{(M)})$$

  is tight. By Prokhorov's Theorem, there exists a subsequence (ρ_{M_k}) and a probability measure ρ^* such that ρ_{M_k} converges weakly to ρ^*.

Consequences/Conclusions

- The Barron space catches all the functions that can be approximated by Shallow Neural Networks with bounded path norm, and the approximation error does not suffer from CoD.
- The Barron space is the largest function set which is well approximated by Shallow Neural Networks, and the Barron norm is the natural norm associated with it. Target functions outside $\mathcal{B}$ may be increasingly difficult to approximate by SNNs as dimension increases.

Consequences/Conclusions

- The Barron space catches all the functions that can be approximated by Shallow Neural Networks with bounded path norm, and the approximation error does not suffer from CoD.
- The Barron space is the largest function set which is well approximated by Shallow Neural Networks, and the Barron norm is the natural norm associated with it. Target functions outside $\mathcal{B}$ may be increasingly difficult to approximate by SNNs as dimension increases.

Open questions

- More specific descriptions of the functions that belong to $\mathcal{B}$.
- Extension to Deep Neural Networks? For ResNets $\Rightarrow$ *Compositional function spaces* [MW+22]

References

References

[Bar93] A. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". In: *IEEE Transactions on Information Theory* 39.3 (1993), pp. 930–945. DOI: 10.1109/18.256500.

[Cyb89] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[KL20] P. Kidger and T. Lyons. "Universal Approximation with Deep Narrow Networks". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 2306–2327.

[MW+22] C. Ma, L. Wu, et al. "The Barron space and the flow-induced function spaces for neural network models". In: *Constructive Approximation* 55.1 (2022), pp. 369–406.

[Pin99] A. Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta Numerica* 8 (1999), pp. 143–195.