The 3 problems of Machine Learning

1. **Expressivity**
   What is the complexity of the functions my model can represent?
2. **Trainability**
   How easy is training of my model (i.e. solving the optimization problem)?
3. **Generalization**
   How does my model behave on unseen data?
   In presence of a shift in distributions?

(After Eric Jang & Jascha Sohl-Dickstein)

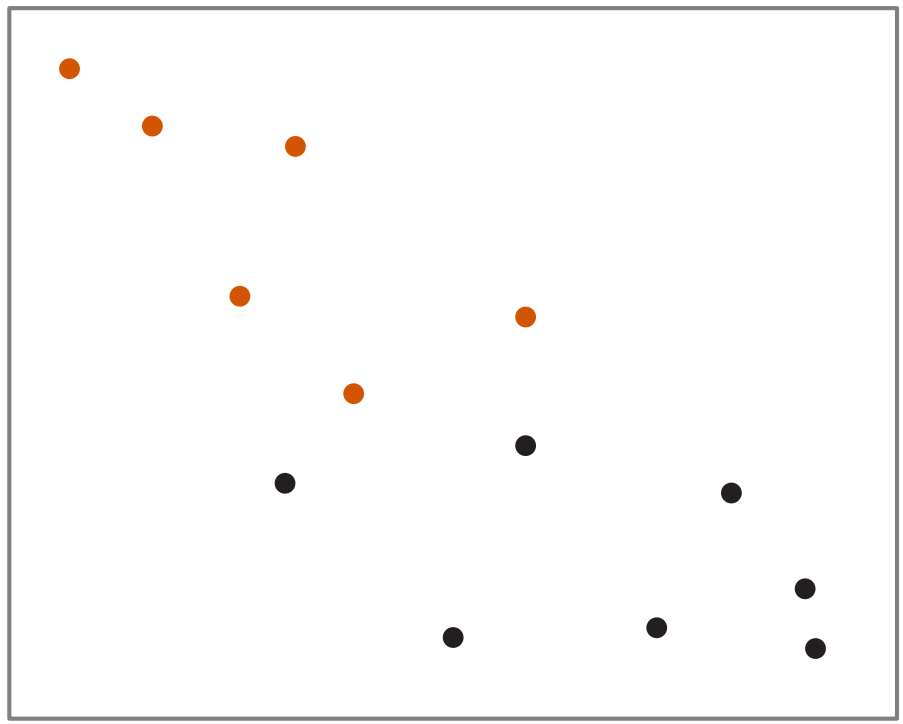
Toy problem

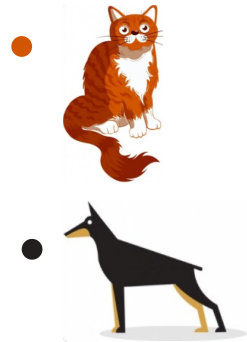
We would like to classify between two classes, cats and dogs.

We have access to two measurements:

x_1 — the weight, in kg.

x_2 — a sound, in Hz.





Supervised Learning:

We have access to a training set of labeled examples.

What is the predicted class (cat or dog) for a new unseen example?

(Separable) Learning problems

Inputs / *features* are from a **domain** set $\mathcal{X}$.
There is a **distribution** $\mathcal{D}$ over the domain $\mathcal{X}$.

Labels are from the label set $\mathcal{Y} = \{0, 1\}$ (binary case).

True labelling function $f : \mathcal{X} \to \mathcal{Y}, y_i = f(x_i)$
Predictor (**"hypothesis"**) $h : \mathcal{X} \to \mathcal{Y}, \hat{y}_i = h(x_i)$

Training data $S = \{(x_i, y_i)\}_{i=1\ldots m}$

For the moment we focus on the $0-1$ Loss / error, i.e.:

$$\ell_{0-1}(h, (x, y)) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases}$$

Empirical risk minimization

We would like to estimate the "**true**" error (risk), which is NOT available:

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\})$$

Instead, we have access to the **empirical** error (risk) on the trainig set:

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

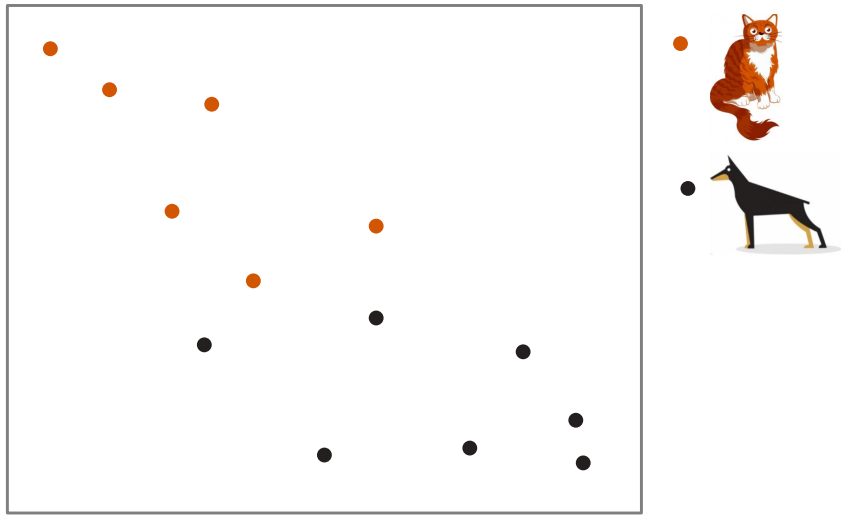
Empirical risk minimization optimizes this risk:

$$h_S = \min_h L_S(h)$$

Overfitting

Let us define an "*inefficient*" classifier ... with training error 0:

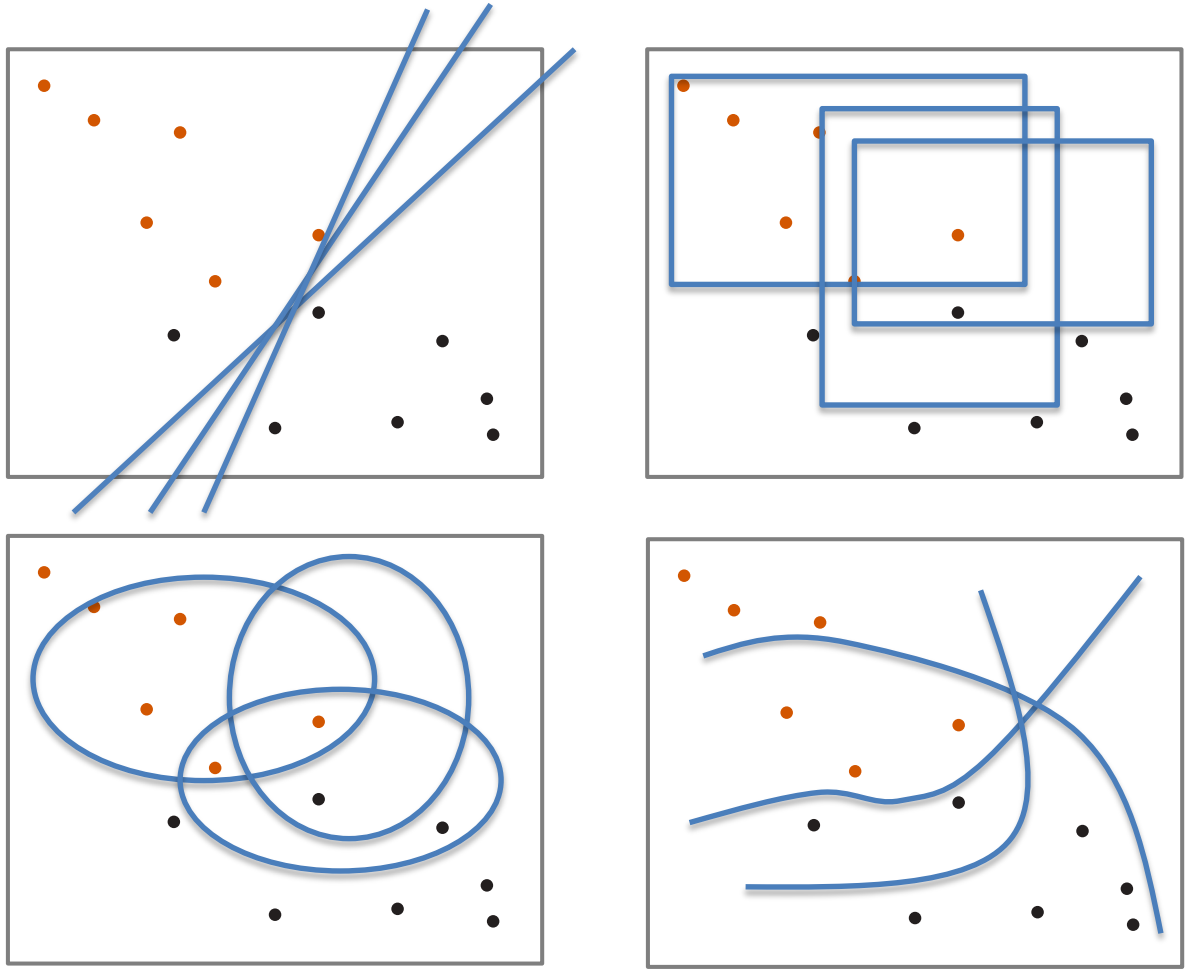
$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{otherwise} \end{cases}$$



Among all possible classifiers with zero training error, which ones is most likely to generalize best?
Can we find theoretical bounds on generalization?

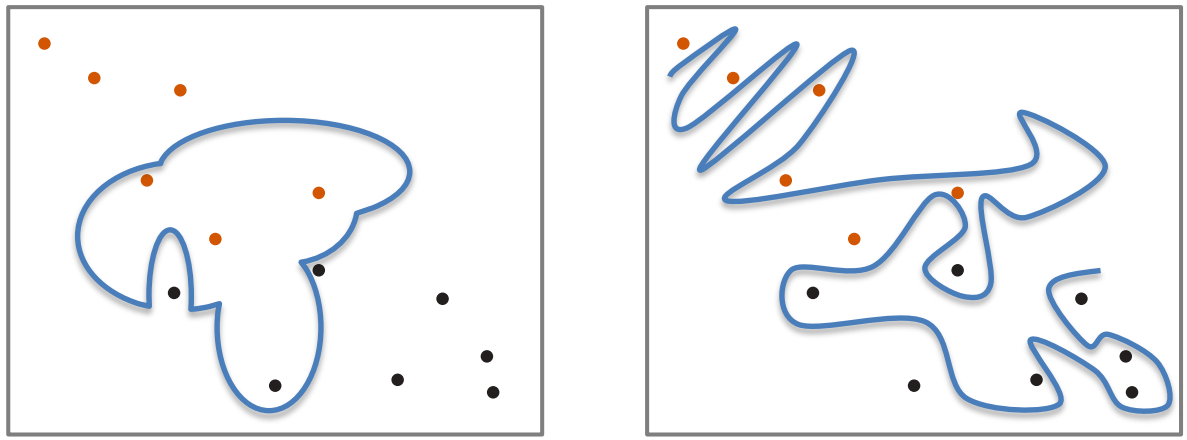
Hypothesis classes

Can we limit the predictor h ("*hypothesis*") to a given class $\mathcal{H}$?



Hypothesis classes

Intuitively, if the decision frontiers are arbitrarily complex, the generalization gap can become arbitrarily large.



Ensuring generalization

Proposition: if the complexity of the mathematical function representing h is limited *(before looking at the data!)*, we can limit the gap between empirical risk and true risk.

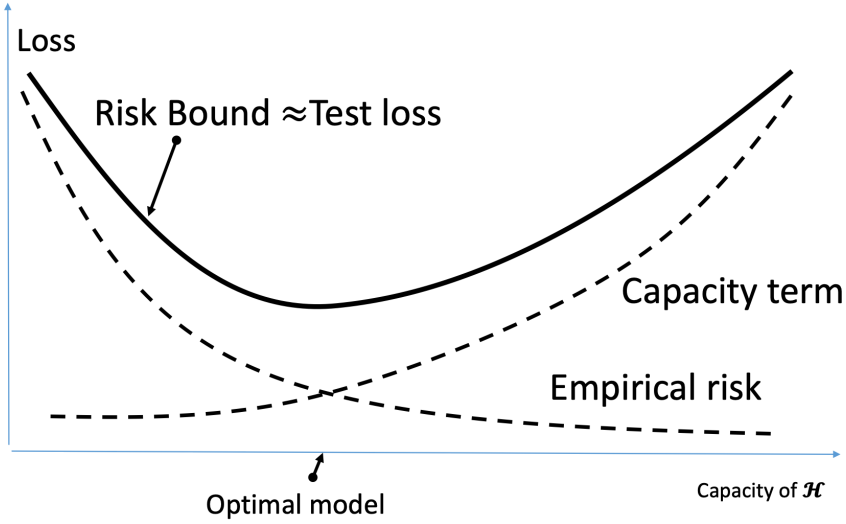
How can we measure the complexity of a function?
$\longrightarrow$ We need some measure of **capacity**.

A large family of bounds uses some form of capacity:

$$\forall h \in \mathcal{H} \quad \underbrace{L_{\mathcal{D},f}(h)}_{\text{expected risk}} - \underbrace{L_S(h)}_{\text{empirical risk}} < \underbrace{O^*\left(\sqrt{\frac{\text{cap}(\mathcal{H})}{m}}\right)}_{\text{capacity term}}$$

m ... number of training samples.

The classical U-curve (bias/complexity)



(Figure reproduced from M. Belkin, *Fit without fear ...*, 2021.)

$$\forall h \in \mathcal{H} \quad \underbrace{L_{\mathcal{D},f}(h)}_{\text{expected risk}} - \underbrace{L_S(h)}_{\text{empirical risk}} < \underbrace{O^*\left(\sqrt{\frac{\text{cap}(\mathcal{H})}{m}}\right)}_{\text{capacity term}}$$

Some types of bounds

Finite hypothesis classes

The number of hypotheses is finite,
Separable problem.

VC-Dimension

Capacity is described on how many data points a hypothesis can separate.

...

The next slides are largely inspired by: Shai-Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning*, 2014.)

Some types of bounds

Finite hypothesis classes

The number of hypotheses is finite,
Separable problem.

VC-Dimension

Capacity is described on how many data points a hypothesis can separate.

...

The next slides are largely inspired by: Shai-Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning*, 2014.)

Finite hypothesis classes

Let us study an easy case: finite hypothesis classes.
Examples:

- All straight lines with parameters a, b being discretized
- All axis aligned rectangles on a discretized grid
- All classification programs expressed in computer code stored in no more than N bits
- ...

Proposition: the generalization gap of predictors learned from finite hypothesis classes is bounded.

i.i.d. Samples

The **i.i.d. assumption**:

> *The samples in the training set S are* **independently** *and* **identically** *sampled, i.e. $S \sim \mathcal{D}$.*

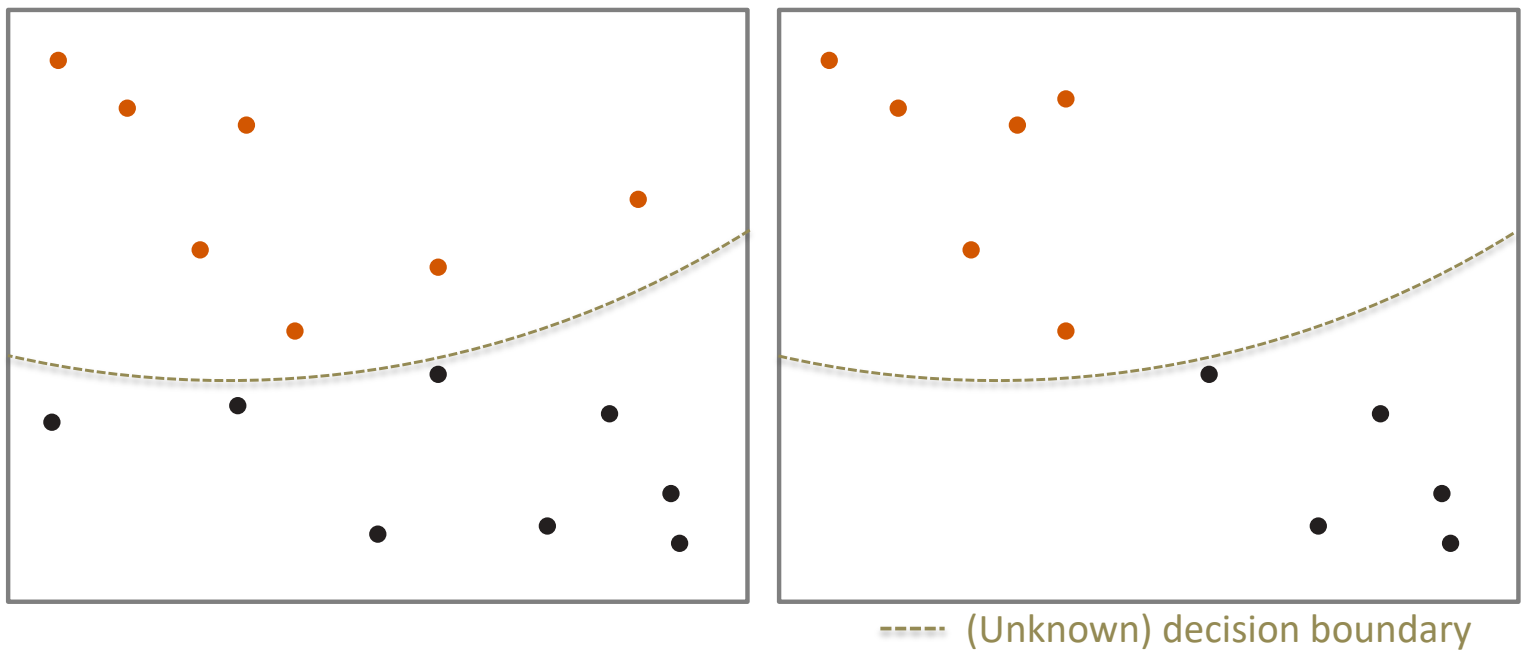
Without this assumption, mathematical analysis of generalization error is diffult.

Predictors are random

The hypothesis $h_S : \mathcal{X} \to \mathcal{Y}$ is a random function, subject to randomness induced by the selection of the training set.

There is a non-zero probability, that

- the training set is non representative, or even
- that the same point is sampled again and again.



Problem more stringent in higher dimensions!

Bounding generalization error

We suppose the problem is realizable (=separable), i.e.

$$\exists h^* \in \mathcal{H} : \; L_{(\mathcal{D},f)}(h^*) = 0$$

We define the failure of h_S as true risk being higher than ϵ:

$$L_{(\mathcal{D},f)}\left(h_S\right) > \epsilon$$

Probability of failure, over sampling the training set $S|_x = (x_1, \ldots, x_m)$:

$$\mathcal{D}^m\left(\left\{ S|_x : L_{(\mathcal{D},f)}\left(h_S\right) > \epsilon \right\}\right)$$

Our goal is to **upper bound** this probability.

Misleading samples

Bad hypotheses:

$$\mathcal{H}_B = \left\{ h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon \right\}$$

The set of **misleading samples**: each set $S|_x$ contains at least one bad hypothesis with zero training error.

$$M = \{ S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0 \}$$

Can be rewritten as

$$M = \bigcup_{h \in \mathcal{H}_B} \{ S|_x : L_S(h) = 0 \}$$

Since the problem is separable, failure can only happen if the sample is misleading, i.e.

$$\left\{ S|_x : L_{(\mathcal{D},f)}\left(h_S\right) > \epsilon \right\} \subseteq M$$

Misleading samples

The probability of failure is therefore smaller than the probability of misleading samples:

$$\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}\left(h_S\right) > \epsilon\right\}\right) \leq \mathcal{D}^m(M)$$

$$= \mathcal{D}^m\left(\cup_{h\in\mathcal{H}_B}\left\{S|_x : L_S(h) = 0\right\}\right)$$

$$\leq \sum_{h\in\mathcal{H}_B} \mathcal{D}^m\left(\left\{S|_x : L_S(h) = 0\right\}\right)$$

(where we used the union bound $\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$)

Individual hypotheses h and samples i

For each single bad hypothesis h, the samples are i.i.d., therefore

$$\begin{aligned}\mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right) &= \mathcal{D}^m\left(\{S|_x : \forall i, h\left(x_i\right) = f\left(x_i\right)\}\right) \\ &= \prod_{i=1}^{m} \mathcal{D}\left(\{x_i : h\left(x_i\right) = f\left(x_i\right)\}\right).\end{aligned}$$

Since h is a bad hypothesis, we can bound individual errors:

$$\begin{aligned}\mathcal{D}\left(\{x_i : h\left(x_i\right) = y_i\}\right) &= 1 - L_{(\mathcal{D},f)}(h) \\ &\leq 1 - \epsilon\end{aligned}$$

Wrapping up . . .

Over all samples, we get:

$$\begin{aligned}\mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right) &\leq (1-\epsilon)^m \\ &\leq e^{-\epsilon m}\end{aligned}$$

Integrating all bad hypotheses, we get:

$$\begin{aligned}\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}\left(h_S\right) > \epsilon\right\}\right) &\leq |\mathcal{H}_B|\, e^{-\epsilon m} \\ &\leq |\mathcal{H}| e^{-\epsilon m}\end{aligned}$$

Illustration: union bound over bad hypotheses

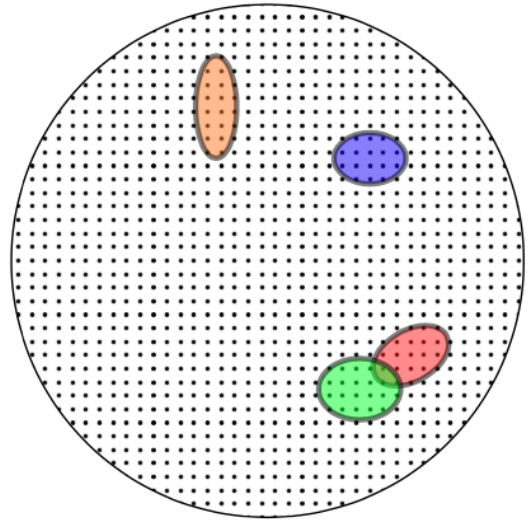


Figure 2.1 Each point in the large circle represents a possible m-tuple of instances. Each colored oval represents the set of "misleading" m-tuple of instances for some "bad" predictor $h \in \mathcal{H}_B$. The ERM can potentially overfit whenever it gets a misleading training set S. That is, for some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$.

(Figure reproduced from Shai-Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning*, 2014.)

Sample Complexity of Finite Hypothesis classes

Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$ and let m be an integer that satisfies

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}.$$

Then, for any labeling function, f, and for any distribution, $\mathcal{D}$, for which the realizability assumption holds (that is, for some $h \in \mathcal{H}, L_{(\mathcal{D},f)}(h) = 0$), with probability of at least $1–\delta$ over the choice of an i.i.d. sample S of size m, we have that for every ERM hypothesis, h_S, it holds that

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon.$$

PAC-Learning

A formal definition of **PAC-Learning**:
"Probably Approximately Correct Learning"

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to N$ and a learning algorithm with the following property:

For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to (0,1)$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_H(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

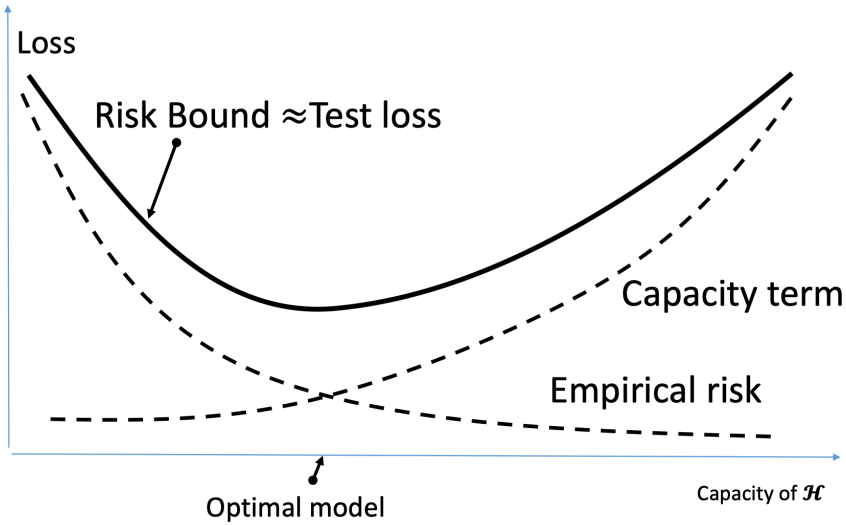
Sample Complexity of Finite Hypothesis classes (2)

We rephrase learnability of finite $\mathcal{H}$:

Finite hypothesis classes $\mathcal{H}$ are PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left| \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right|$$

Bias/complexity: finite $\mathcal{H}$, separable problems



(Figure reproduced from M. Belkin, *Fit without fear ...*, 2021.)

Non-realizable (non-separable) problems

Reminder: for **realizable problems**,

- we specify a marginal distribution $\mathcal{D}$ over the domain $\mathcal{X}$,
- the (unknown) labelling function assigns a unique label to each sample, $f : \mathcal{X} \to (0, 1)$.

For classification, we call these problems **separable**.

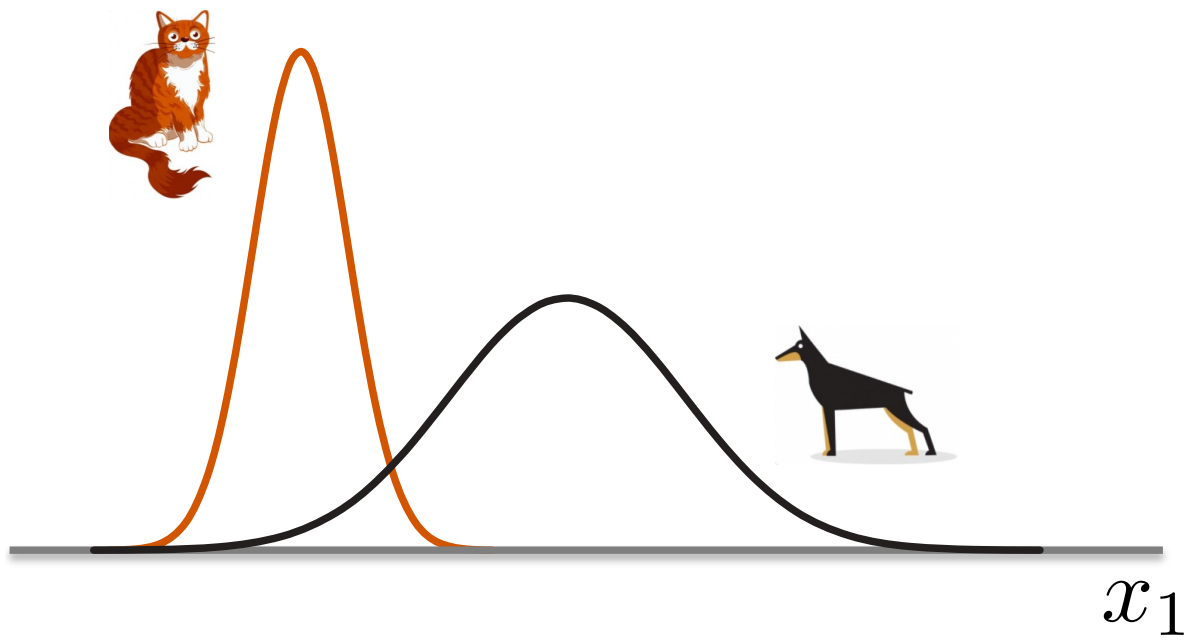
We now remove this constraint and address more practicable problems:

- we integrate noise and uncertainty, there is no clear unique label for a given sample
- we specify the full joint distribution $\mathcal{D}$ over domain and labels, $\mathcal{X} \times \mathcal{Y}$.

Non-separable problems: example

We try to distinguish between cats and dogs, but with a single scalar input value x_1 (the weight in kg).

Let's assume the following **class conditional probabilities**:
$\mathbb{P}(x_1|y=1)$ // cat
$\mathbb{P}(x_1|y=2)$ // dog

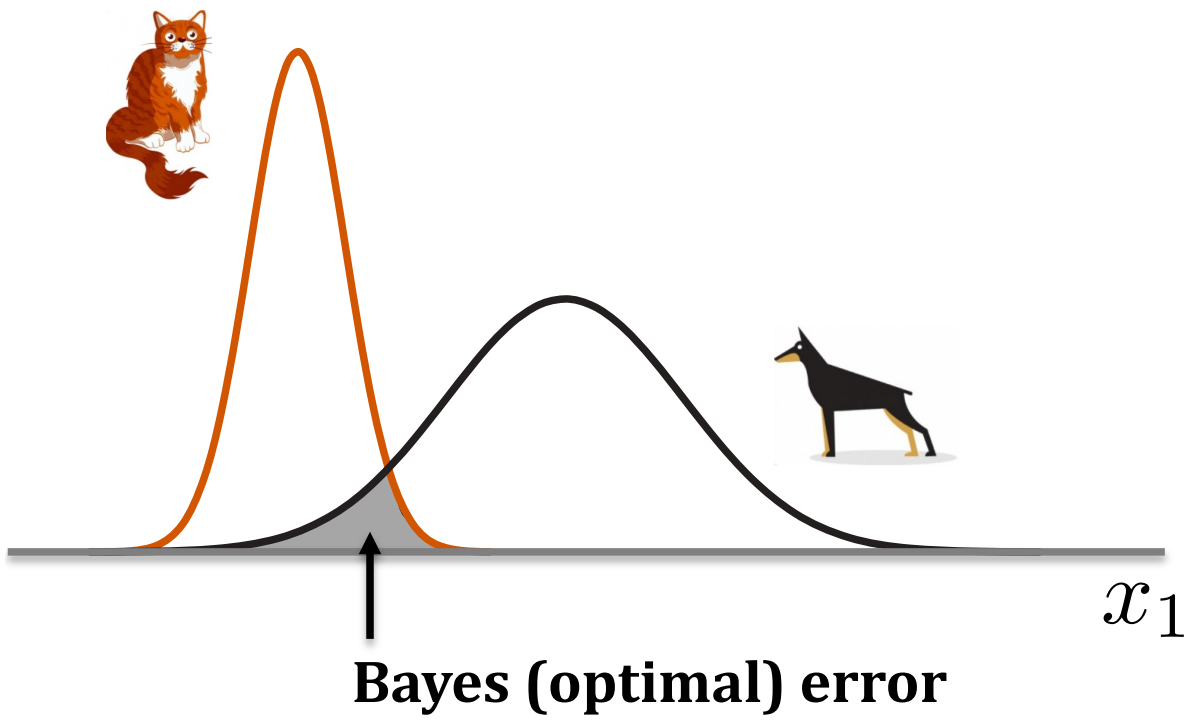


They overlap!

Non-separable problems: example

We try to distinguish between cats and dogs, but with a single scalar input value x_1 (the weight in kg).

Let's assume the following **class conditional probabilities**:
$\mathbb{P}(x_1|y=1)$ // cat
$\mathbb{P}(x_1|y=2)$ // dog



The Bayes optimal classifier

The **true error** is the error obtained by a classifier h when the labels are sampled according to the distribution $\mathcal{D}$:

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \stackrel{\text{def}}{=} \mathcal{D}(\{(x,y) : h(x) \neq y\})$$

The **Bayes optimal classifier** maximizes the **posterior probability**:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}[y = 1 \mid x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Some types of bounds

Finite hypothesis classes

The number of hypotheses is finite,
Separable problem.

VC-Dimension

Capacity is described on how many data points a hypothesis can separate.

...

VC-Dimension

Finite hypothesis classes are learnable.

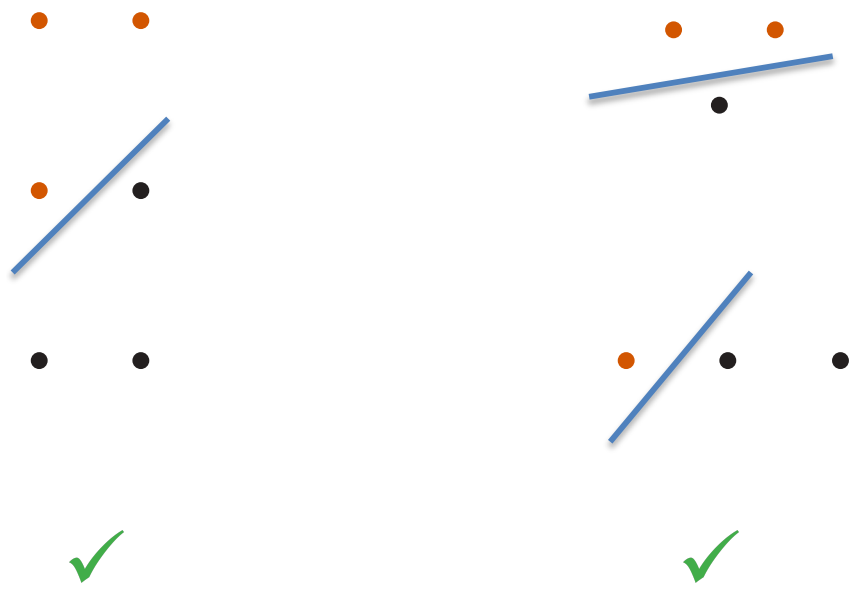
Infinite hypothesis classes are not learnable, if they are completely unrestricted (No free lunch theorem).

However, infinite hypothesis classes can be learnable too, if the capacity is somehow limited.
How do we measure this limitation (=**capacity**)?

Shattering of linear classifiers

Linear classifiers can separate 2 or 3 points,

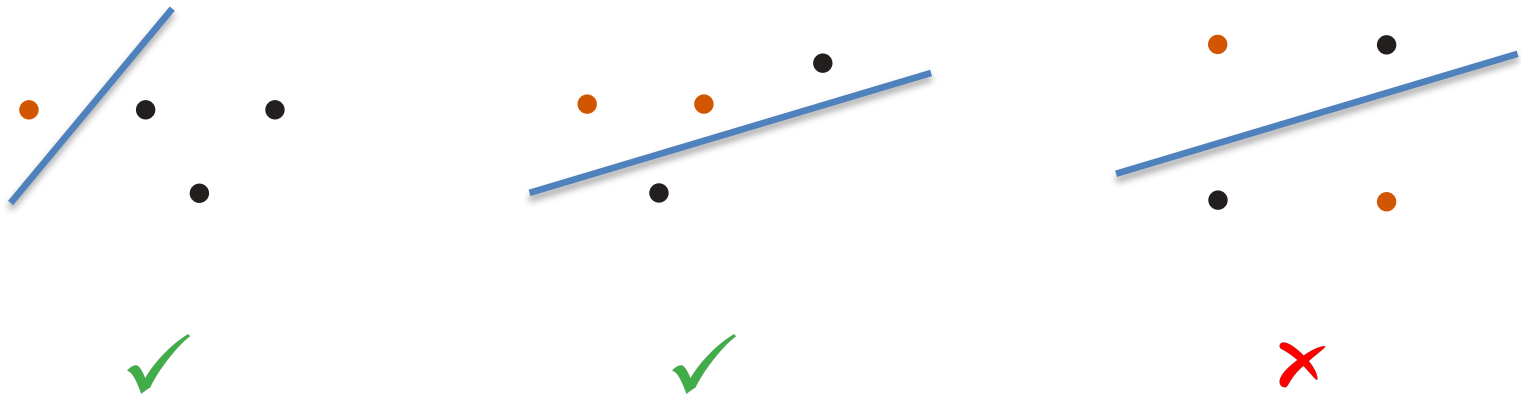
- whatever their labelling y_i,
- whatever their position x_i.



Shattering of linear classifiers

For a set of 4 points or more:

- some configurations $\{(x_i, y_i)\}_{i=1}^{4}$ can be separated by a linear classifier,
- some can not!

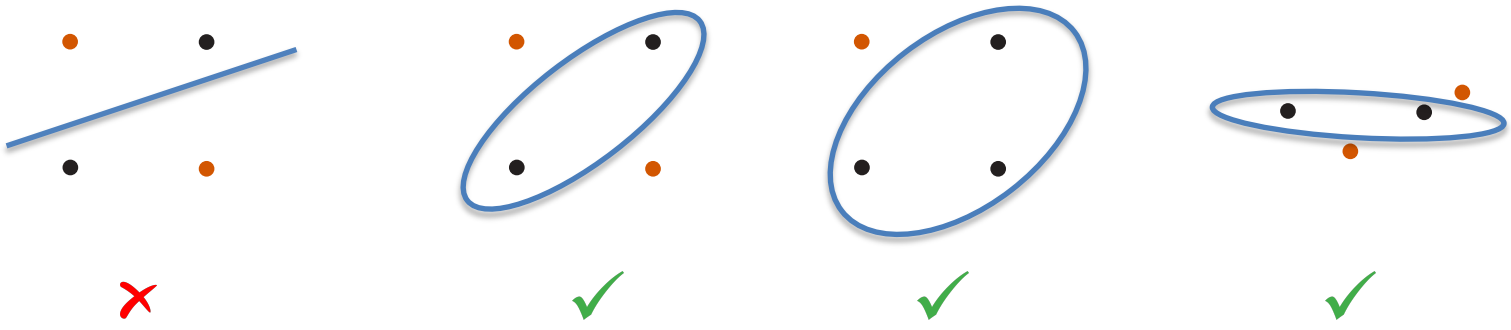


(XOR brought the AI winter . . .)

Shattering ...

A classifier with elliptical decision boundaries can separate 4 points,

- whatever their labelling y_i,
- whatever their position x_i.



Shattering: more formally

Reminder: a hypothesis class $\mathcal{H}$ is the class of functions $h : \mathcal{X} \to (0, 1)$.

A hypothesis class $\mathcal{H}$ shatters a finite set $C = \{c_1, c_2, \ldots, c_m\} \subset \mathcal{X}$ if the set of functions from C to $(0, 1)$ that can be derived from $\mathcal{H}$, i.e.

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) : h \in \mathcal{H}\}$$

is equal to the set of all functions from C to $(0, 1)$.

(Each element in $(\mathcal{H}_C)$ is a sequence of bits $0/1$)

VC-dimension

The VC-dimension of a hypothesis class $\mathcal{H}$ is the maximum size of a set $C \subset \mathcal{X}$ such that C can be shattered by $\mathcal{H}$.
(Not all sets C!)

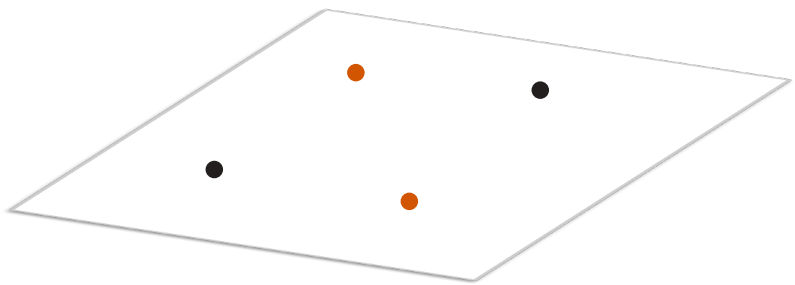
Examples:

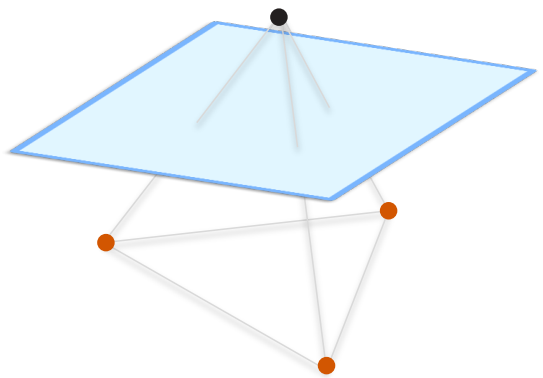
Linear classifiers in 2D have a VC dimension of 3.
Generally, in D dimensions, linear classifiers have dimension D+1.

$\Rightarrow$ there exist sets of 4 points in 3D space, which can be separated by hyperplanes whatever their labelling.

Example: linear separators in 3D

$\Rightarrow$ there exist sets of 4 points in 3D space, which can be separated by hyperplanes whatever their labelling.





The fundamental theorem of PAC Learning

The Fundamental Theorem of Statistical Learning
Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $(0, 1)$ and let the loss function be the $0 - 1$ loss. Then, the following are equivalent:

- $\mathcal{H}$ has a finite VC-dimension.
- $\mathcal{H}$ is PAC learnable.
- Any ERM rule is a successful PAC learner for $\mathcal{H}$.

(Other equivalencies exist and are out of scope of this lecture)

Proof in: Shai-Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning*, 2014.)

The fundamental theorem of PAC Learning

The Fundamental Theorem of Statistical Learning — Quantitative Version

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $(0, 1)$ and let the loss function be the $0 - 1$ loss. Assume that it's VC-dimension is $\mathcal{H} = d < \infty$. Then, there are absolute constants C_1, C_2 such that:

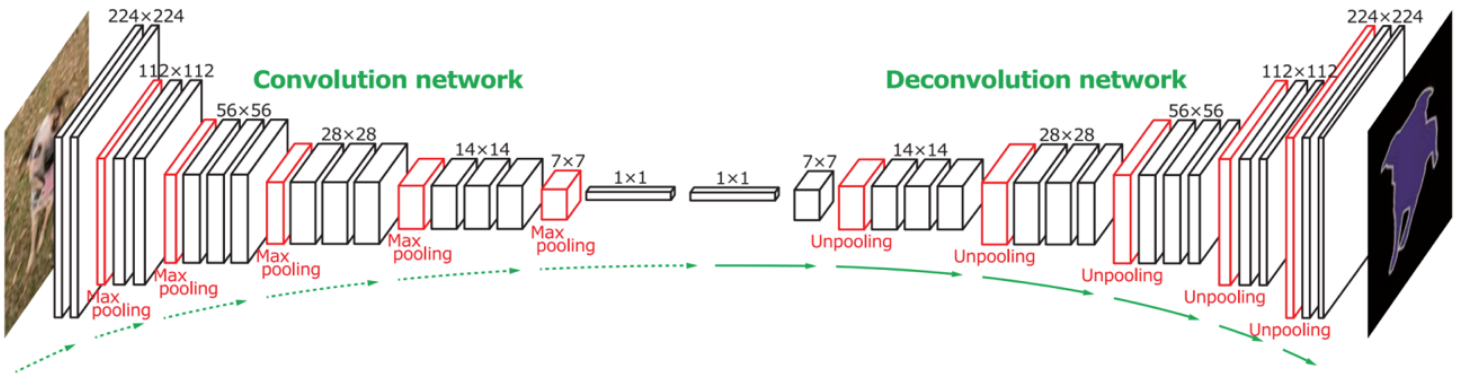
$\mathcal{H}$ is PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

(Other properties are implied ... and are out of scope of this lecture)

Proof in: Shai-Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning*, 2014.)

Then Deep Learning came along ...

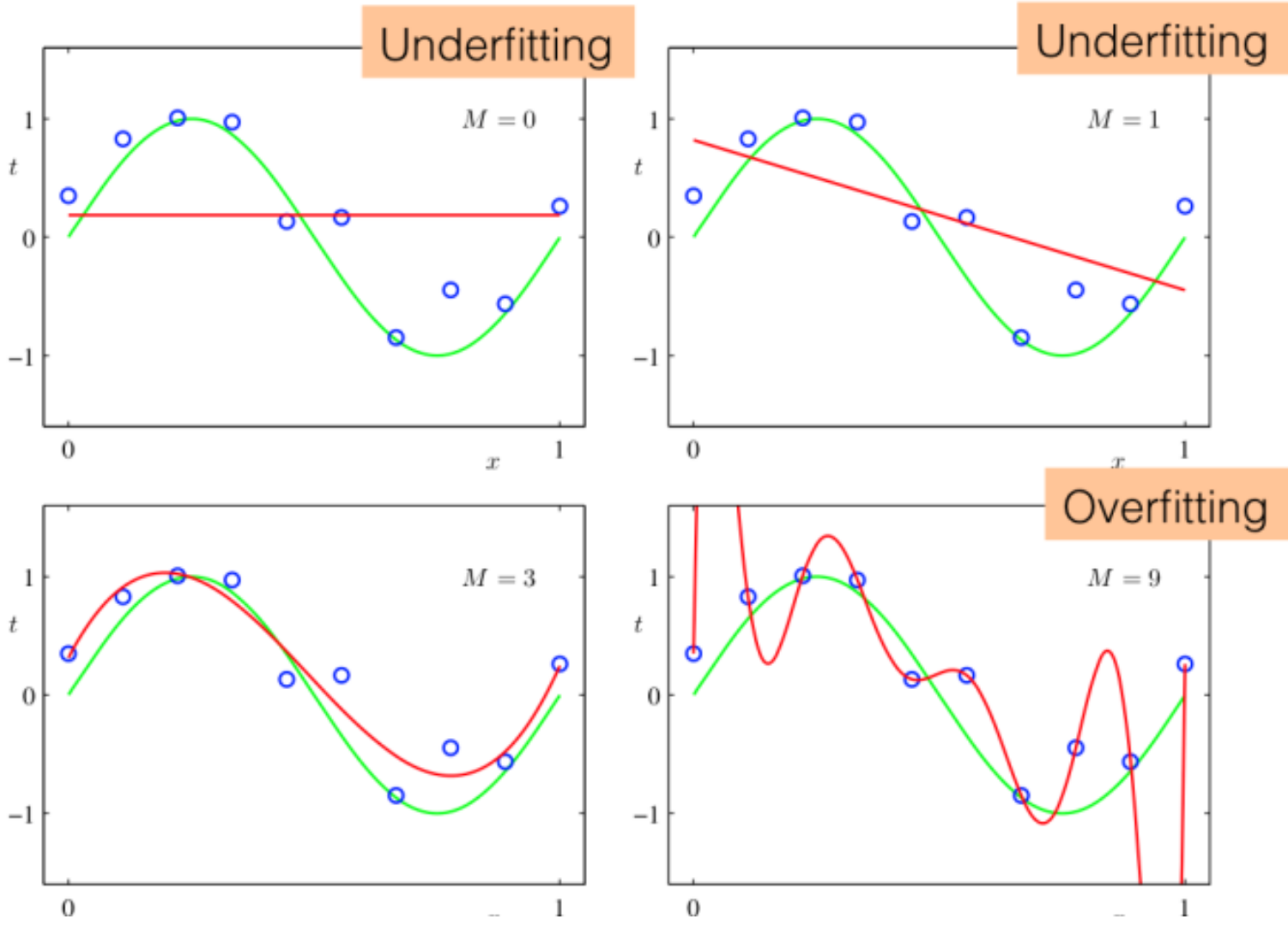


(Noh et al., 2015)

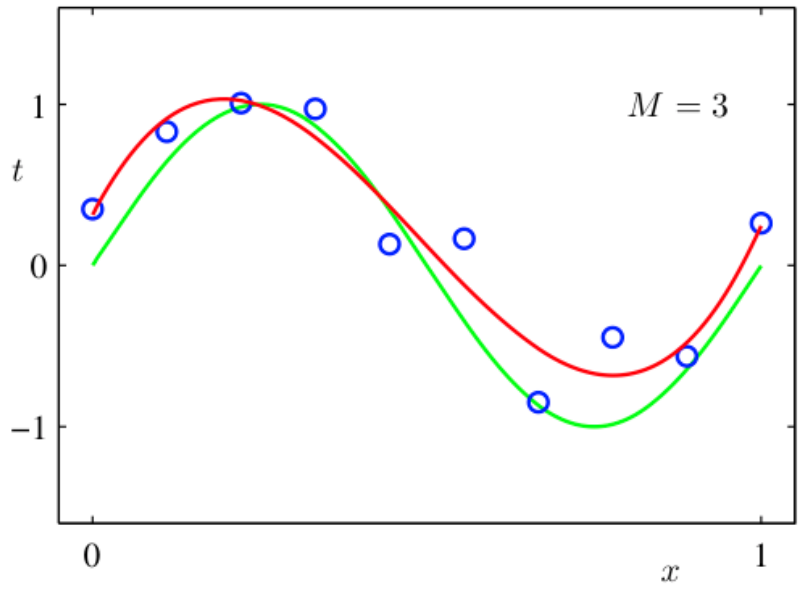
Rethinking generalization

C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, Understanding deep learning requires rethinking generalization, ICLR 2017 (best paper)

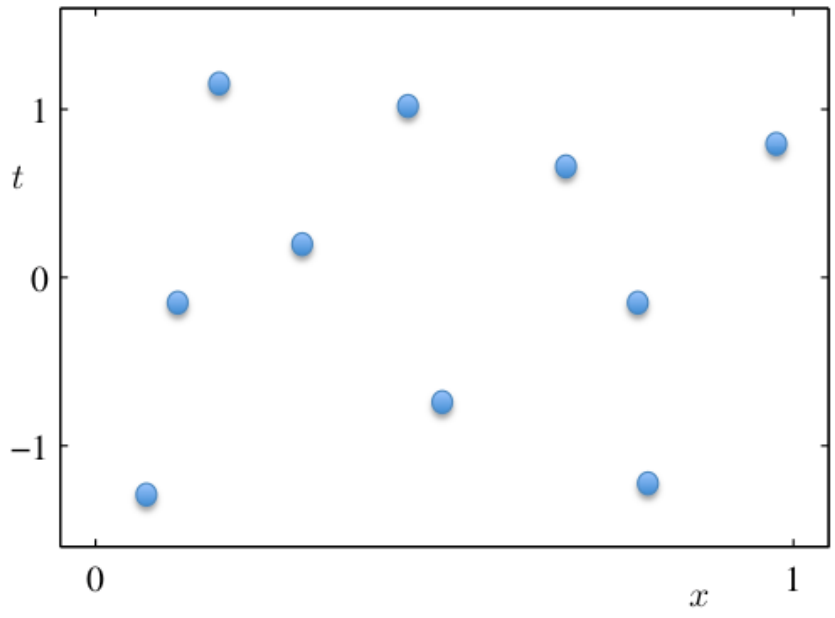
Rethinking generalization



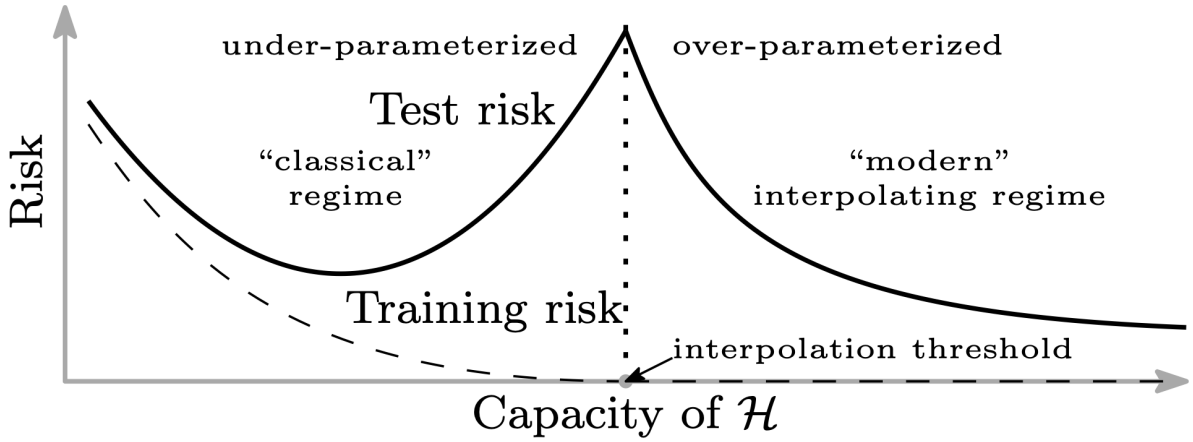
Rethinking generalization



Rethinking generalization

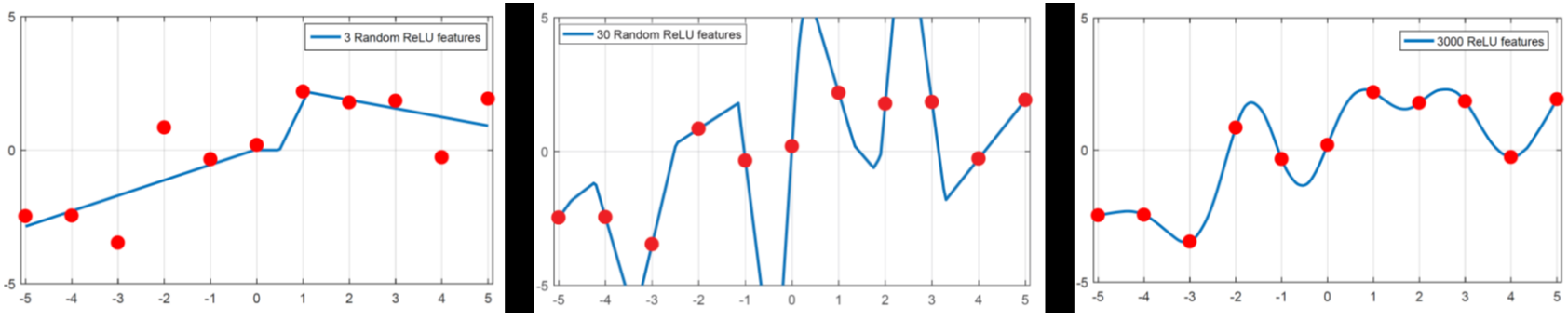


The double U-curve



(Figure reproduced from M. Belkin, *Fit without fear ...*, 2021.)

Lack of overfitting



(Figure reproduced from M. Belkin, *Fit without fear ...*, 2021.)