UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Lesson 3.

The learning problem

DATA SCIENCE AND AUTOMATION COURSE

MASTER DEGREE SMART TECHNOLOGY ENGINEERING

TEACHER
Mirko Mazzoleni

PLACE
University of Bergamo

Outline

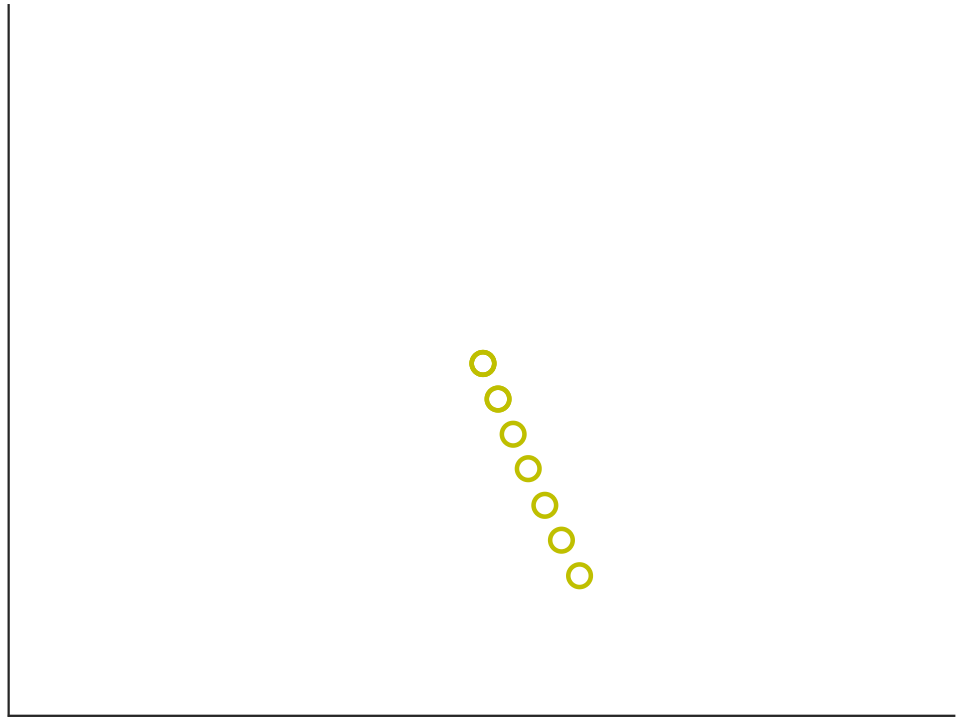
1. Feasibility of learning
2. Generalization bounds
3. Bias-Variance tradeoff
4. Learning curves

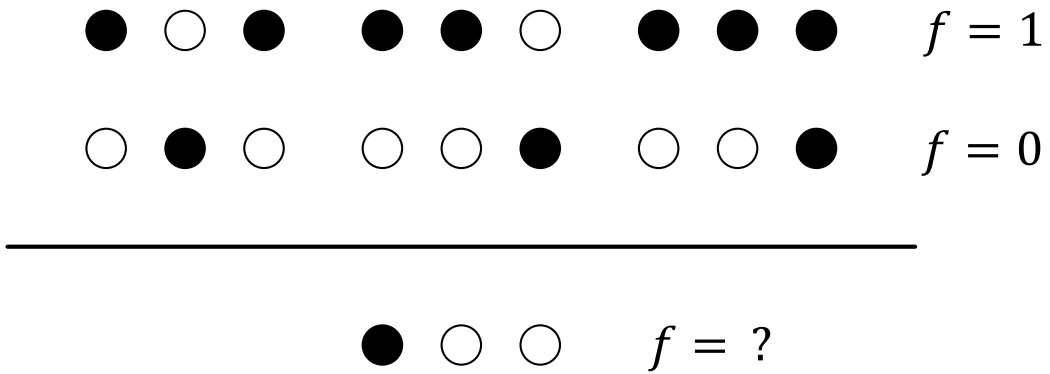
Outline

1. **Feasibility of learning**
2. Generalization bounds
3. Bias-Variance tradeoff
4. Learning curves

Puzzle

Focus on supervised learning: which are the plausible response values of the unknown function, on positions of the input space that we have not seen? [5]

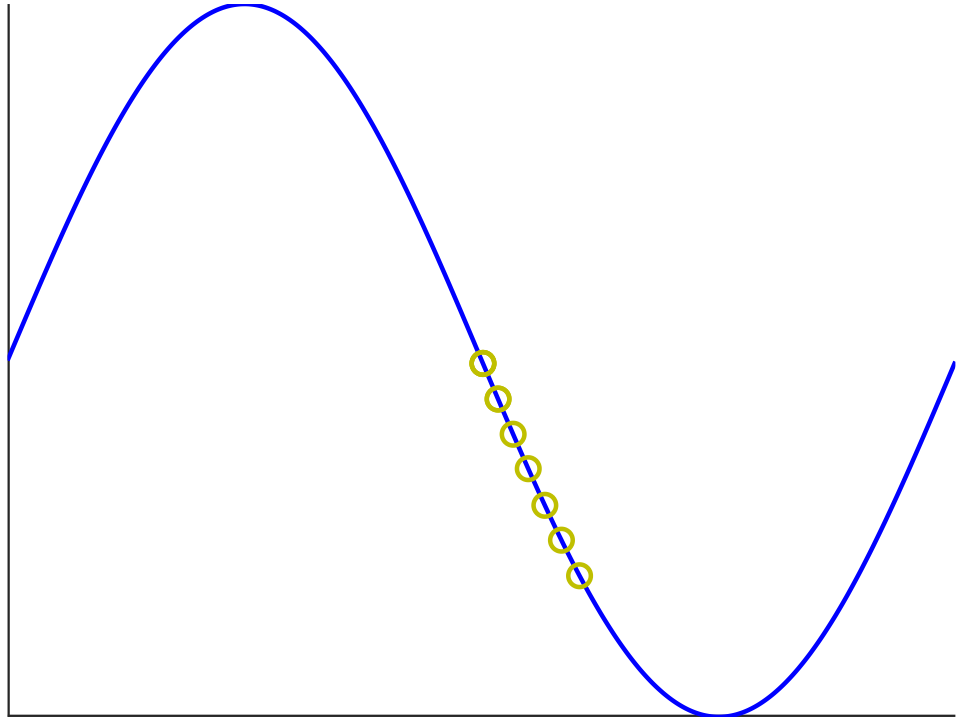


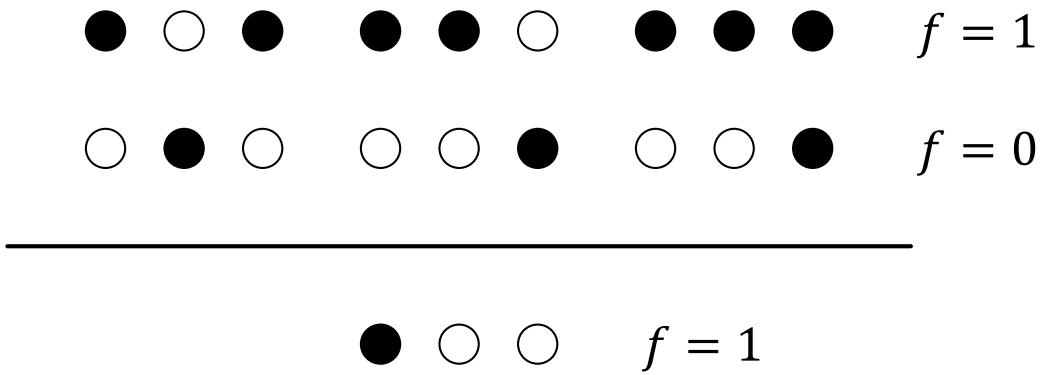


Puzzle

It is not possible to know how the function behaves outside the observed points

(Hume's induction problem) [7]





Feasibility of learning

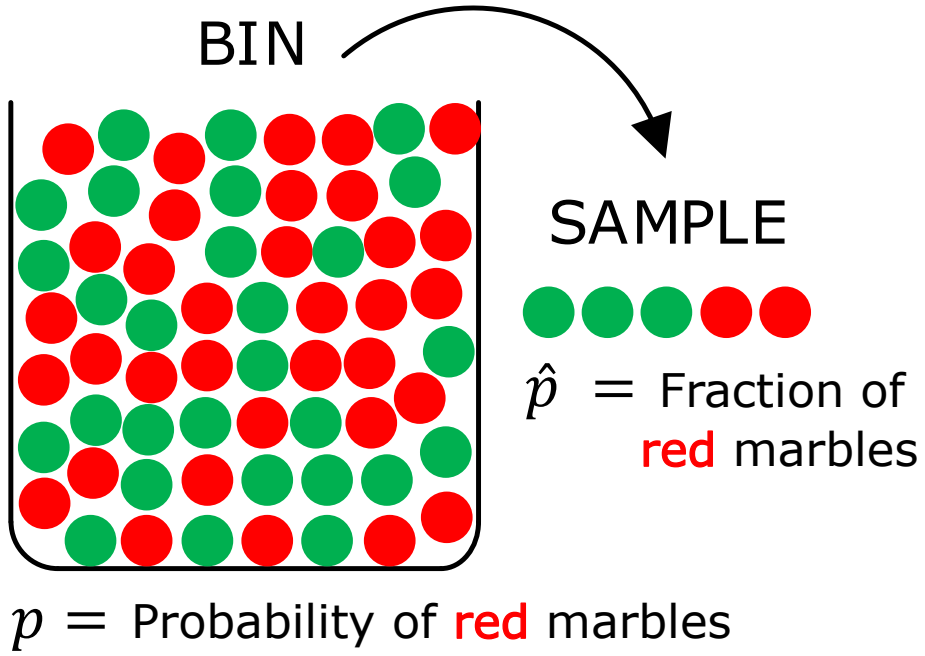
Focus on **supervised learning**, **dichotomous classification** case

Problem: learn an unknown function f

Solution: Impossible ☹. The function can assume any value outside the data we have

Experiment

- Consider a "bin" with red and green marbles
- $\mathbb{P}[$ picking a **red** marble $] = p$
- The value of p is unknown to us
- Pick N marbles independently
- Fraction of red marbles in the sample $= \hat{p}$



Does $\hat{p}$ say something about p?

No!

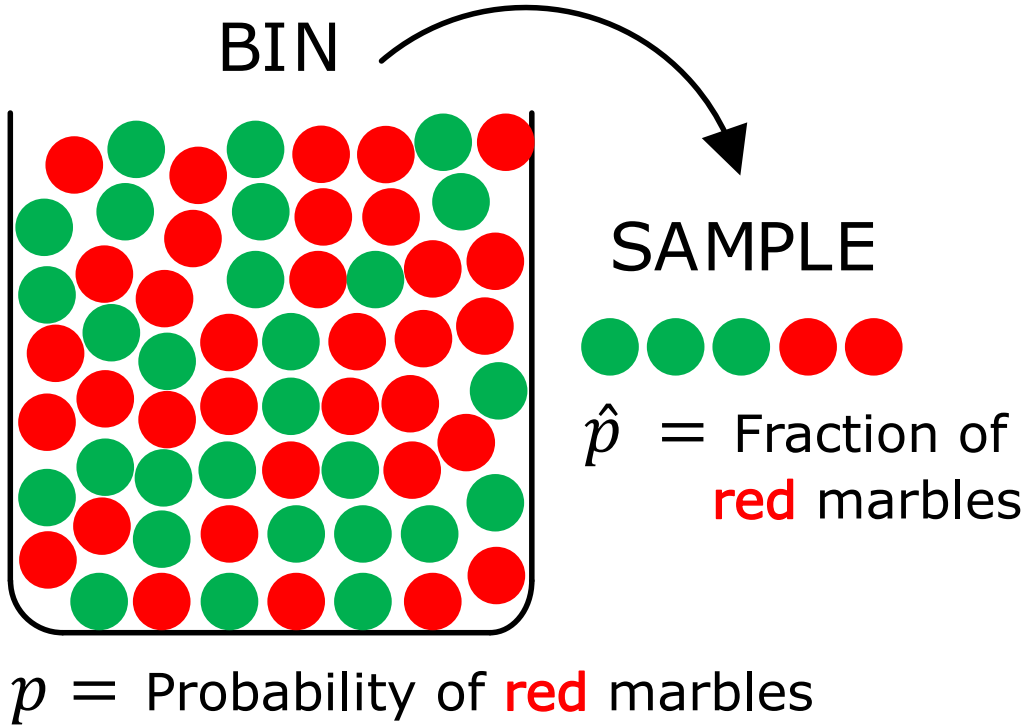
Sample can be mostly **green** while bin is mostly **red**

Possible

Yes!

Sample frequency $\hat{p}$ is likely close to bin frequency p *(if the sample is sufficiently large)*

Probable



Does $\hat{p}$ say something about p?

In a big sample (large N), $\hat{p}$ is probably close to p (within ϵ)

This is stated by the **Hoeffding's inequality:**

$$\mathbb{P}[|\hat{p} - p| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$

The statement $\hat{p} = p$ is P.A.C. (Probably Approximately Correct)

- The quantity $|\hat{p} - p| > \varepsilon$ is a **bad event**, we want its probability to be low
- The bound is valid for all N and ε is a **margin of error**
- The bound does not depend on p
- If we set for a lower margin ε, we have to increase the data N in order to have a small probability of $|\hat{p} - p| > \epsilon$ (bad event) happening

Connection to learning

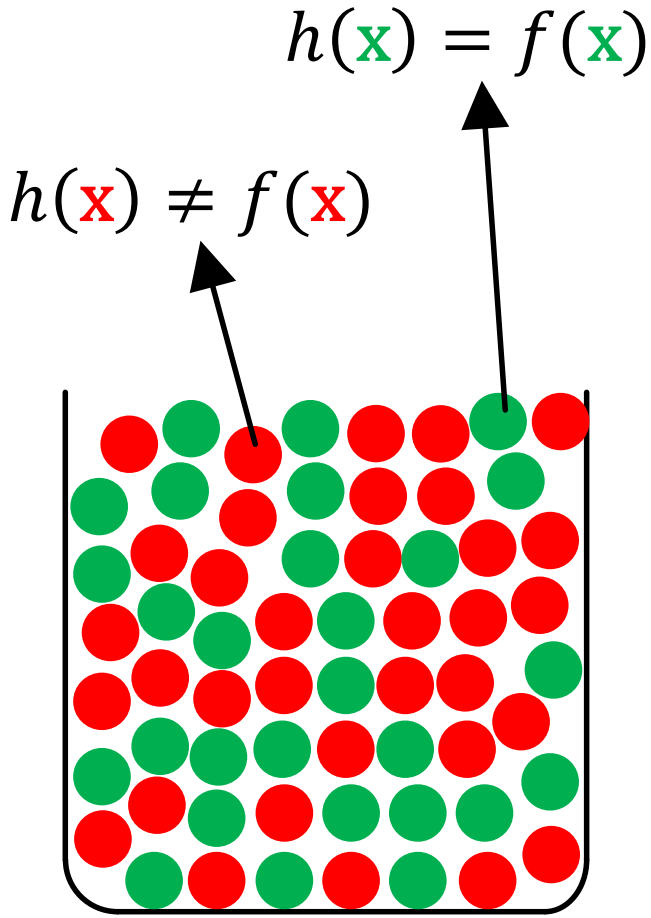
Bin: The unknown is a number p

Learning: The unknown is a function $f: \mathcal{X} \to \mathcal{Y}$

Each marble ● is a input point $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{d \times 1}$

For a specific hypothesis $h \in \mathcal{H}$:

- Hypothesis **got it right** $h(\boldsymbol{x}) = f(\boldsymbol{x})$
- Hypothesis **got it wrong** $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$



Both p and $\hat{p}$ depend on the particular hypothesis h

$\hat{p} \to$ in-sample error $E_{in}(h)$

$p \to$ out-of-sample error $E_{out}(h)$

The **Out of sample error** $E_{out}(h)$ is the quantity that really matters

Connection to REAL learning

In a learning scenario, the function h **is not fixed a priori**

- The *learning algorithm* is used to fathom the hypothesis space $\mathcal{H}$, to find the best hypothesis $h \in \mathcal{H}$ that **matches the sampled data** → call this hypothesis g
- With many hypotheses, there is more probability to find a good hypothesis g **only by chance** → the function can be perfect on sampled data but bad on unseen data

There is therefore an **approximation - generalization tradeoff** between:

- Perform well on the **given** (training) dataset
- Perform well on **unseen** data

Connection to REAL learning

Probability of a "bad event" is less than a huge number ➔ not useful bound

The Hoeffding's inequality becomes:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 2Me^{-2\varepsilon^2 N}$$

where M is the number of hypotheses in $\mathcal{H}$ → M can be infinity ☹

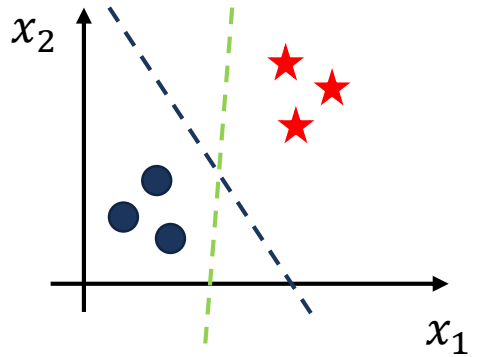
The quantity $E_{out}(g) - E_{in}(g)$ is called the **generalization error**

It turns out that the number of hypotheses M can be replaced by a quantity $m_{\mathcal{H}}(N)$ (called the **growth function**) which is eventually **bounded by a polynomial**

Connection to REAL learning

It turns out that the number of hypotheses M can be replaced by a quantity $m_{\mathcal{H}}(N)$ (called the **growth function**) which is eventually **bounded by a polynomial**

- This is due to the fact the M hypotheses will be very overlapping → They generate the same "classification dichotomy"



Vapnik-Chervonenkis Inequality

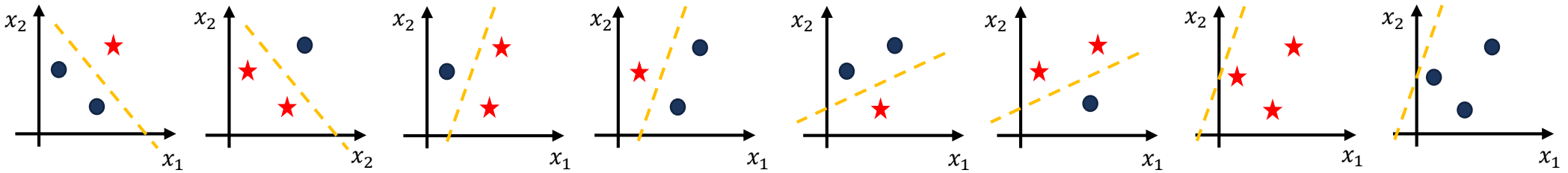
$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\varepsilon^2 N}$$

Generalization theory

The **VC-dimension** is a single parameter that characterizes the growth function

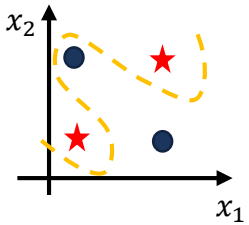
Definition

The Vapnik-Chervonenkis dimension of a hypothesis set $\mathcal{H}$ is the max number of points for which the hypothesis can generate all possible classification dichotomies



$N = 3$

Max n° dichotomies $2^N = 8$



$N = 4$

The linear model is not able to provide all 2^4 dichotomies, we would need a nonlnear one

It can be shown that:

- If the d_{VC} is **finite**, then $m_{\mathcal{H}} \leq N^{dVC} + 1$ → this is a polynomial that will be eventually dominated by e^{-N} → generalization guarantees
- For **linear models** $y = \sum_{j=1}^{d-1} \theta_j x_j + \theta_0$ we have that $d_{VC} = d$ → can be interpreted as the number of parameters of the model

Rearranging things

Start from the VC inequality:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\varepsilon^2 N}$$

δ

Get ε in terms of δ:

$$\delta = 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\varepsilon^2 N} \Rightarrow \varepsilon = \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Ω

Interpretation

- I want to be at most ε away from E_{out}, given that I have E_{in}
- I want this statement to be correct $\geq (1-\delta)\%$ of the times
- Given any two of N, δ, ε it is possible to compute the remaining element

Outline

1. Feasibility of learning

2. **Generalization bounds**

3. Bias-Variance tradeoff

4. Learning curves

Generalization bound

Following previous reasoning, it is possible to say that, with probability $\geq 1 - \delta$:

$$|E_{in}(g) - E_{out}(g)| \leq \Omega(N, \mathcal{H}, \delta) \quad \Rightarrow -\Omega(N, \mathcal{H}, \delta) \leq E_{in}(g) - E_{out}(g) \leq \Omega(N, \mathcal{H}, \delta)$$

Solving for inequalities leads to:

1. $E_{out}(g) \geq E_{in}(g) - \Omega(N, \mathcal{H}, \delta)$ → Not of much interest ☹
2. $\boldsymbol{E_{out}(g) \leq E_{in}(g) + \Omega(N, \mathcal{H}, \delta)}$ → Bound on the out of sample error! ☺

Observations

- $E_{in}(g)$ is known
- The penalty Ω can be computed if $d_{VC}(\mathcal{H})$ is known and δ is chosen

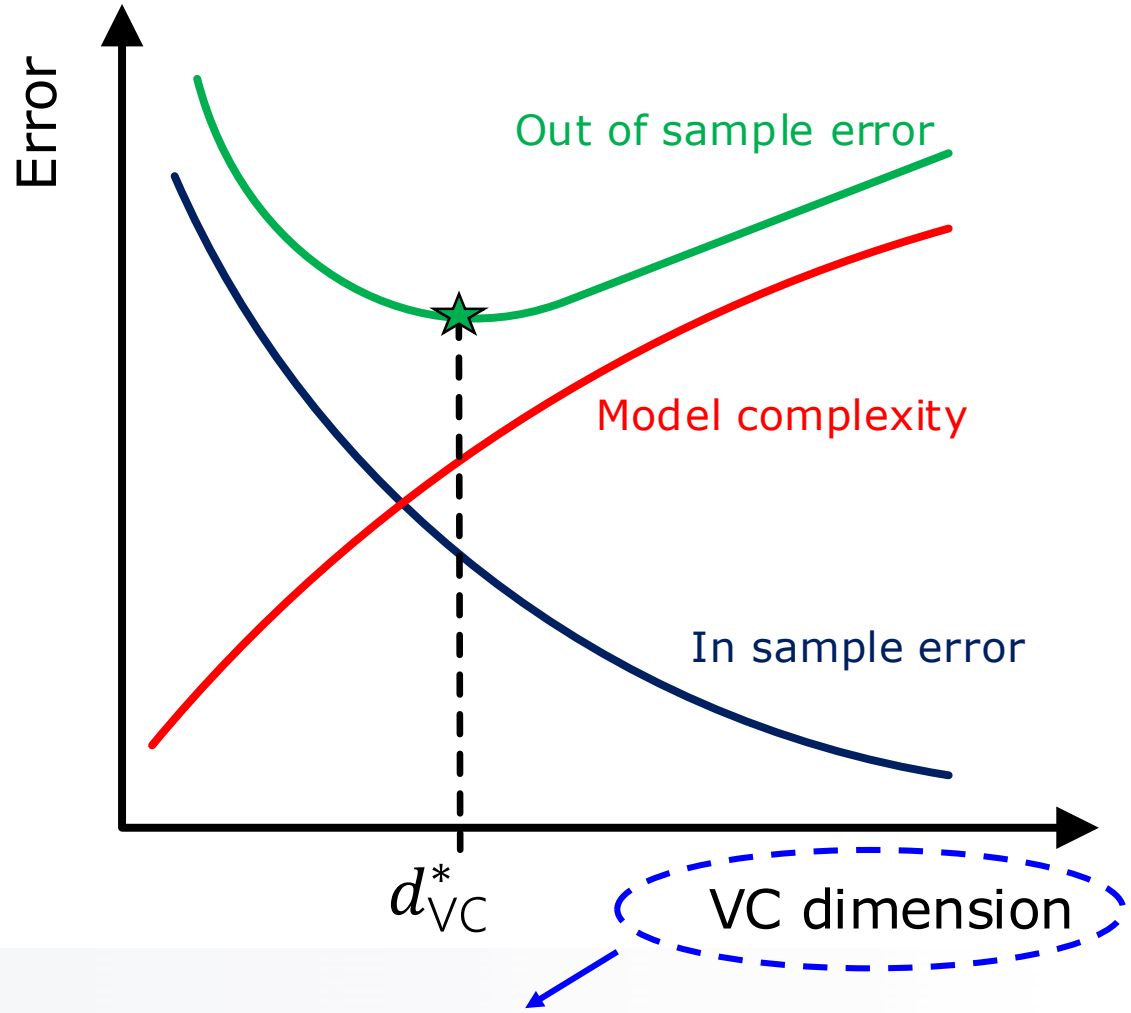
Generalization bound

$$\Omega = \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}$$

Analysis of the generalization bound $E_{out}(g) \leq E_{in}(g) + \Omega(N, \mathcal{H}, \delta)$

- $\Omega \uparrow$ if $d_{VC} \uparrow$ → More penalty for model complexity
- $\Omega \uparrow$ if $\delta \downarrow$ → More penalty for higher confidence
- $\Omega \downarrow$ if $N \uparrow$ → Less penalty with more examples
- $E_{in} \downarrow$ if $d_{VC} \uparrow$ → A more complex model can fit the data better

The optimal model is a **compromise** between E_{in} and Ω



Model complexity ≈ number of model parameters

Take home lessons

Generalization theory, based on the concept of VC-dimension, studies the cases in which it is possible to **generalize** out of sample what we find in sample

- The takeaway concept is that learning is **feasible** in a **probabilistic** way
- If we are able to deal with the approximation-generalization tradeoff, we can say with **high probability** that the **generalization error is small**

Rule of thumb

How many data points N are required to ensure a good generalization bound?

$$N \geq 10 \cdot d_{VC}$$

General principle

Match the **"model complexity"** to the **data resources**, not to the **target complexity**

Outline

1. Feasibility of learning

2. Generalization bounds

3. **Bias-Variance tradeoff**

4. Learning curves

Approximation vs. generalization

The ultimate goal is to have a small E_{out}: good approximation of f out of sample

- More complex $\mathcal{H} \Rightarrow$ better chances of **approximating** f in sample $\rightarrow$ if $\mathcal{H}$ is too simple, we fail to approximate f and we end up with a large E_{in}

- Less complex $\mathcal{H} \Rightarrow$ better chance of **generalizing** out of sample $\rightarrow$ if $\mathcal{H}$ is too complex, we fail to generalize well

Ideal: $\mathcal{H} = \{ f \}$ winning lottery ticket ☺

Approximation vs. generalization

The example shows:

- perfect fit on in sample (training) data

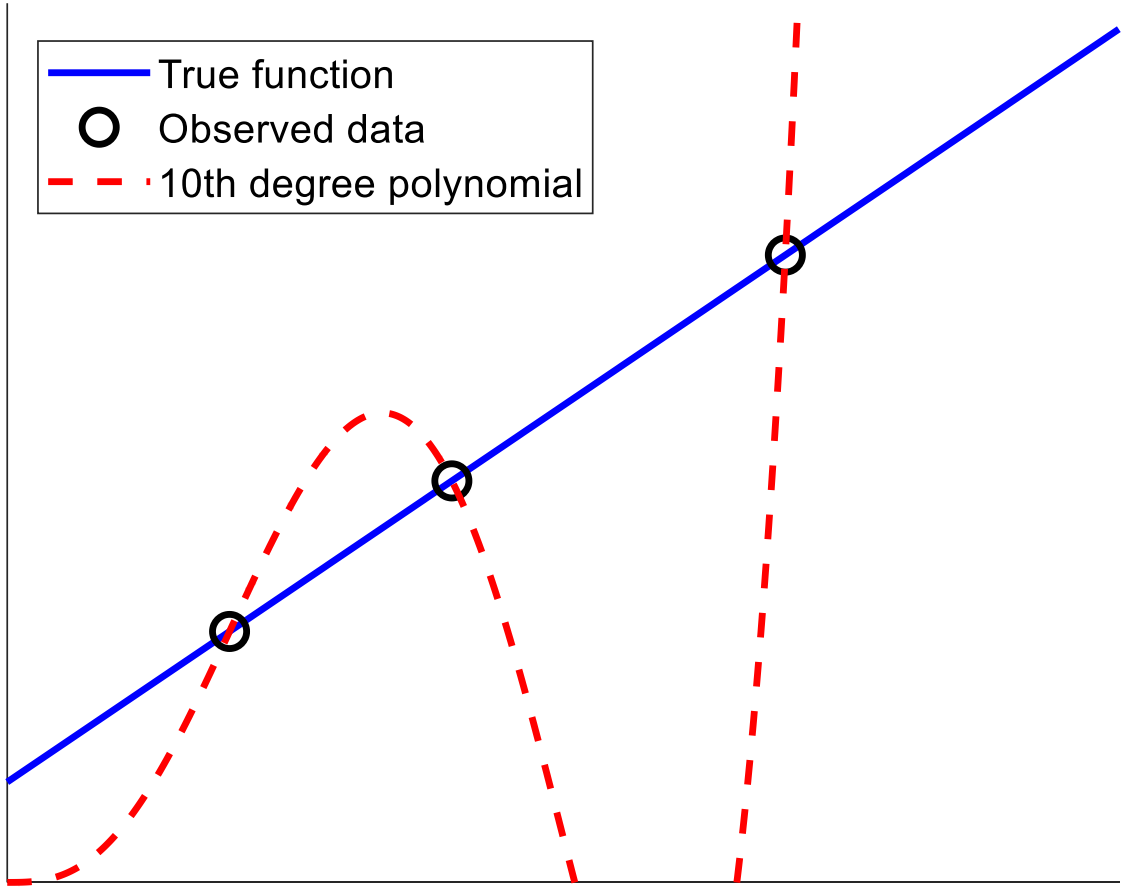
$$\downarrow$$

$$E_{in} = 0$$

- low fit on out of sample (test) data

$$\downarrow$$

E_{out} **huge**



Quantifying the tradeoff

VC analysis was one approach: $E_{out} \leq E_{in} + \Omega$

Bias-variance analysis is another: decomposing E_{out} into:

1. How well $\mathcal{H}$ can approximate f → Bias
2. How well we can zoom in on a good $h \in \mathcal{H}$, using the available data → Variance

It applies to **real valued targets** and uses **squared error**

The learning algorithm is not obliged to minimize squared error loss. However, we measure its produced hypothesis's bias and variance using squared error

Outline

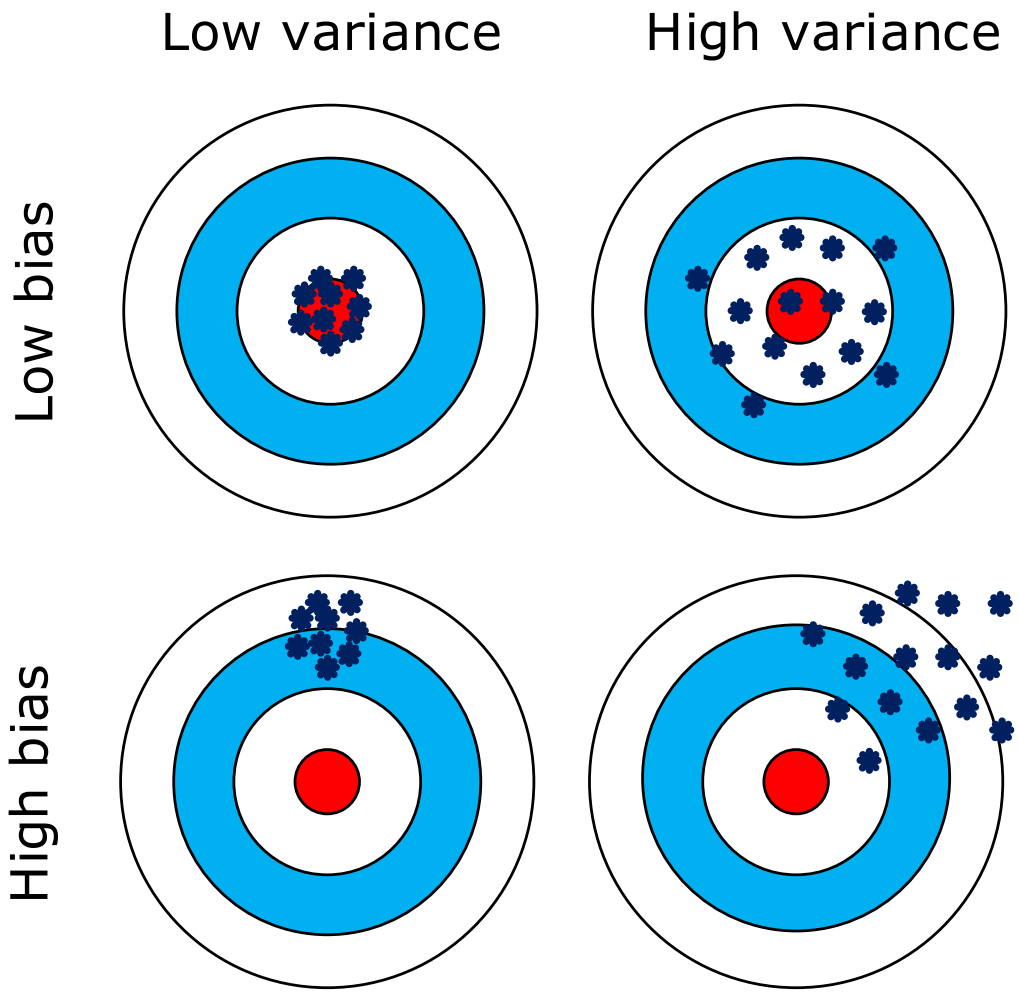
1. Feasibility of learning

2. Generalization bounds

3. **Bias-Variance tradeoff**

4. Learning curves

Bias and variance



Bias and variance

The out of sample error is (making explicit the dependence of g on $\mathcal{D}$)

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

The **expected** out of sample error of the learning model is independent of the particular realization of data set used to find $g^{(\mathcal{D})}$:

$$\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$$

$$= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$$

Bias and variance

Focus on $\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$ Define the "average" hypothesis $\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right]$

This average hypothesis can be derived by imagining many datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ and building it by $\bar{g}(\mathbf{x}) \approx \frac{1}{N}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(\mathbf{x})$ → this is a conceptual tool, and $\bar{g}$ does not need to belong to the hypothesis set

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}) + \bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}\Bigg[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2 + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2$$

$$+ 2 \cdot \left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)\Bigg]$$

Bias and variance

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\textbf{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\textbf{bias}(\mathbf{x})}$$

Therefore

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\textbf{bias}(\mathbf{x}) + \textbf{var}(\mathbf{x})\right] \\ &= \quad \textbf{bias} \quad + \quad \textbf{var}\end{aligned}$$

Bias and variance

Interpretation

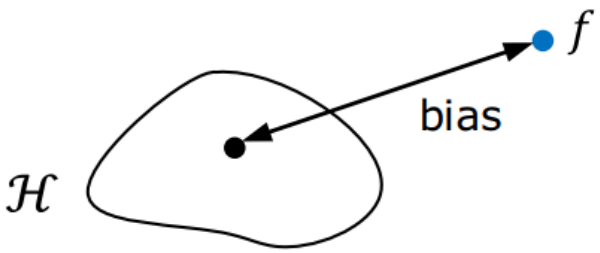
- The **bias** term $\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2$ measures how much our learning model is biased away from the target function

  In fact, $\bar{g}$ has the benefit of learning from an unlimited number of datasets, so it is only **limited in its ability** to approximate f by the limitations of the learning model itself

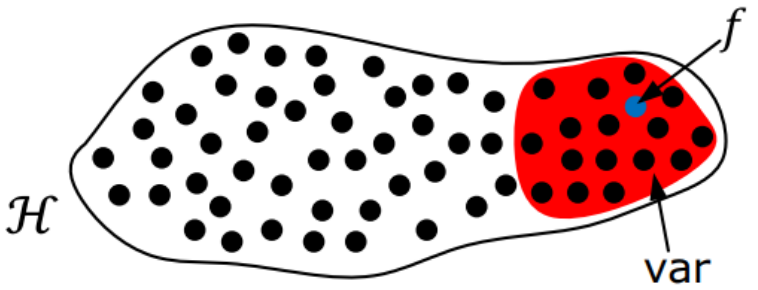
- The **variance** term $\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]$ measures the variance in the final hypothesis, depending on the data set, and can be thought as how much the final chosen hypothesis **differs from the "mean"** (best) hypothesis

Bias and variance

$$\text{bias} = \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2$$



$$\text{variance} = \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]$$



Very small model. Since there is only one hypothesis, both the average function $\bar{g}$ and the final hypothesis $g^{(\mathcal{D})}$ will be the same, for any dataset. Thus, `var = 0`. The bias will depend solely on how well this single hypothesis approximates the target f, and unless we are extremely lucky, we expect a large bias

Very large model. The target function is in $\mathcal{H}$. Different data sets will led to different hypotheses that agree with f on the data set, and are spread around f in the red region. Thus, `bias` $\approx$ `0` because $\bar{g}$ is likely to be close to f. The `var` is large (heuristically represented by the size of the red region)

Outline

1. Feasibility of learning

2. Generalization bounds

3. Bias-Variance tradeoff

4. **Learning curves**

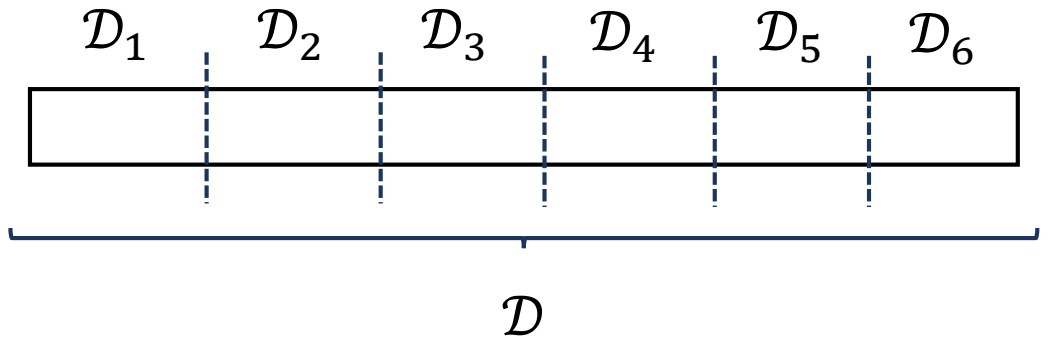
Learning curves

How it is possible to know if a model is suffering from bias or variance problems?

The learning curves provide a **graphical representation** for assessing this, by plotting:

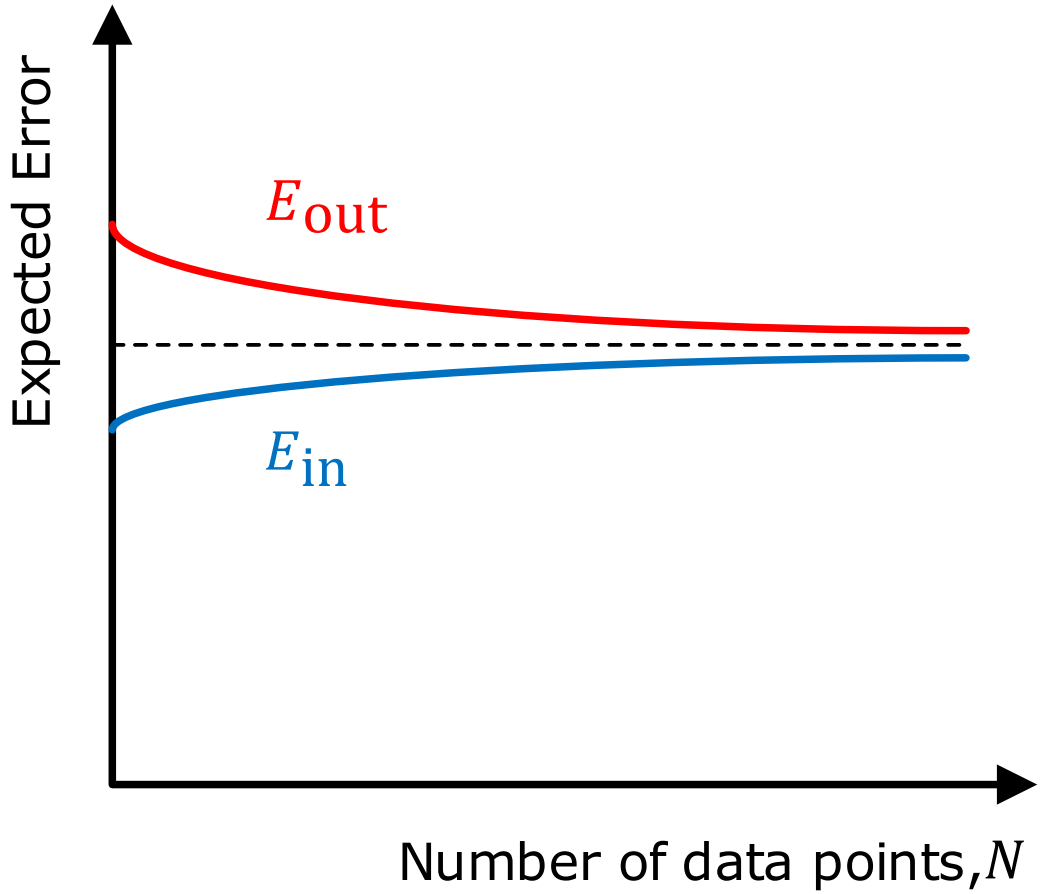
- the **expected** out of sample error $\mathbb{E}_{\mathcal{D}}\left[E_{out}\left(g^{\mathcal{D}}\right)\right]$
- the **expected** in sample error $\mathbb{E}_{\mathcal{D}}\left[E_{in}\left(g^{\mathcal{D}}\right)\right]$

with respect to the **number of data** N

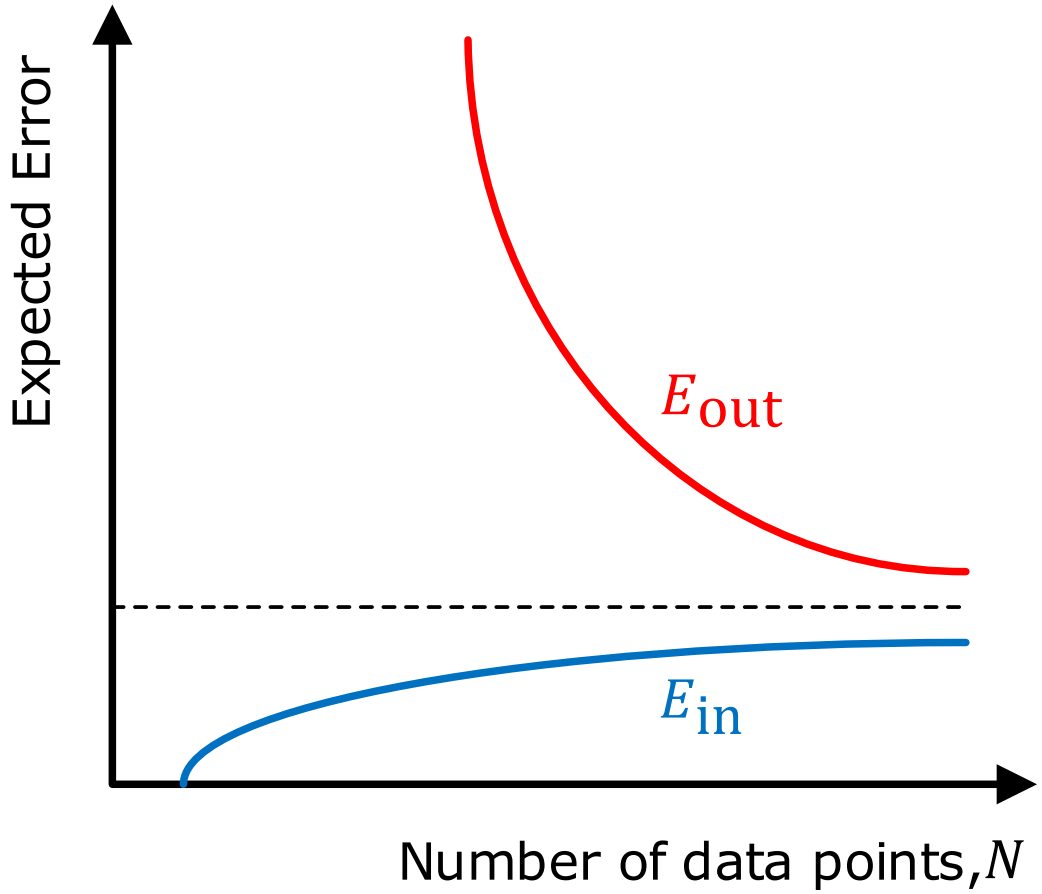


In practice, the curves are **computed from one dataset**, or by dividing it into more parts and taking the mean curve resulting from the various sub-datasets

Learning curves



Simple model



Complex model

Learning curves

Interpretation

- Bias can be present when the expected **error is quite** high and E_{in} is **similar** to E_{out}
- When bias is present, getting **more data is not likely** to help
- Variance can be present when **there is a gap** between E_{in} and E_{out}
- When variance is present, getting **more data is likely** to help

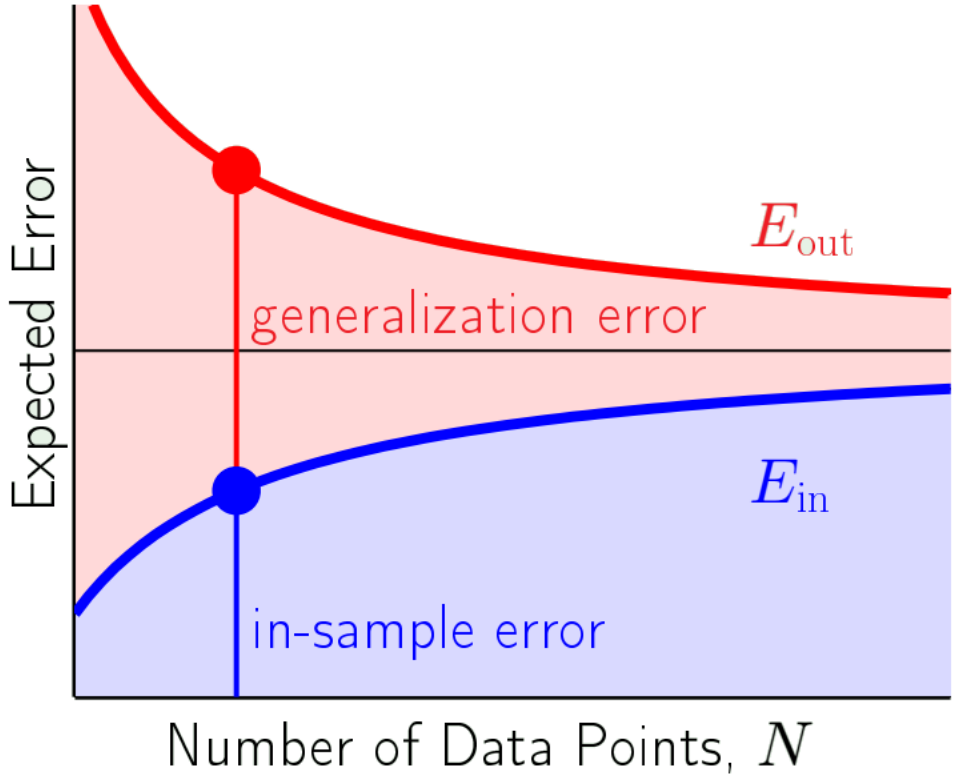
Fixing bias

- Try add more features
- Try polynomial features
- Try a more complex model
- Boosting

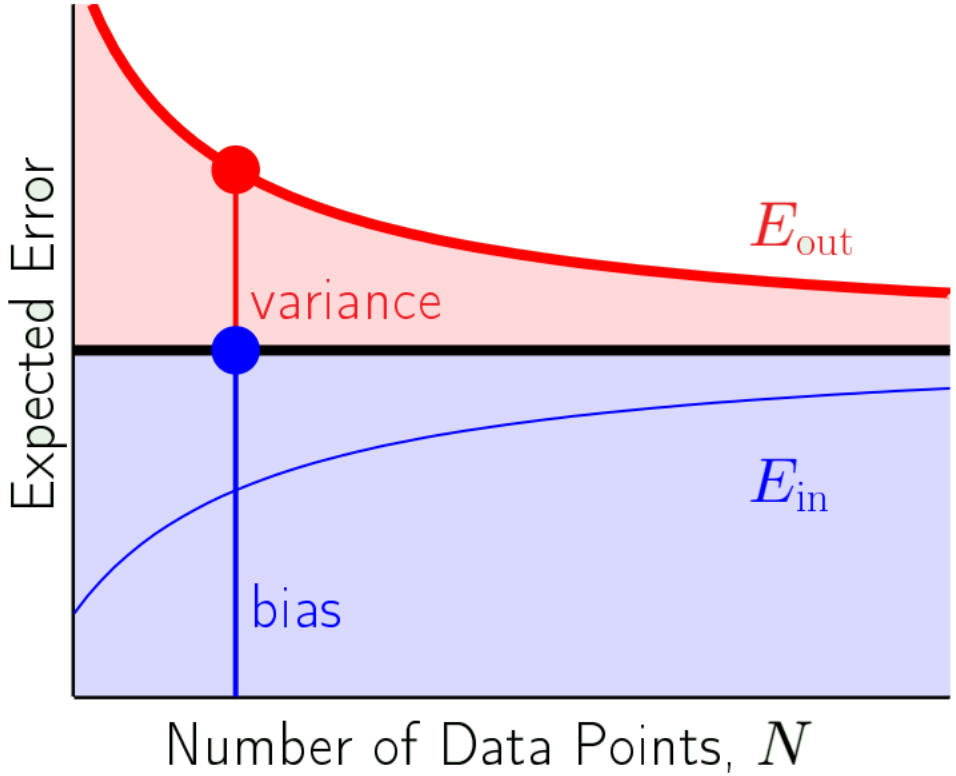
Fixing variance

- Try a smaller set of features
- Get more training examples
- *Regularization*
- Bagging

Learning curves: VC vs. bias-variance analysis



VC analysis



bias-variance

Pictures taken from [5]

References

1. Provost, Foster, and Tom Fawcett. "*Data Science for Business: What you need to know about data mining and data-analytic thinking*". O'Reilly Media, Inc., 2013.
2. Brynjolfsson, E., Hitt, L. M., and Kim, H. H. *"Strength in numbers: How does data driven decision making affect firm performance?"* Tech. rep., available at SSRN: http://ssrn.com/abstract=1819486, 2011.
3. Pyle, D. "*Data Preparation for Data Mining*". Morgan Kaufmann, 1999.
4. Kohavi, R., and Longbotham, R. *"Online experiments: Lessons learned".* Computer, 40 (9), 103–105, 2007.
5. Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. *"Learning from data"*. AMLBook, 2012.
6. Andrew Ng. *"Machine learning"*. Coursera MOOC. (https://www.coursera.org/learn/machine-learning)
7. Domingos, Pedro. "*The Master Algorithm".* Penguin Books, 2016.
8. Christopher M. Bishop, "*Pattern recognition and machine learning", Springer-Verlag New York, 2006.*