

Deep Learning

Lecture 1: Fundamentals of machine learning

Prof. Gilles Louppe
g.louppe@uliege.be

Today

A recap on statistical learning:

- Supervised learning
- Empirical risk minimization
- Under-fitting and over-fitting
- Bias-variance dilemma

Statistical learning

Supervised learning

Consider an unknown joint probability distribution $p_{X,Y}$.

Assume training data

$$(\mathbf{x}_i, y_i) \sim p_{X,Y},$$

with $\mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \dots, N$.

- In most cases,
 - \mathbf{x}_i is a p -dimensional vector of features or descriptors,
 - y_i is a scalar (e.g., a category or a real value).
- The training data is generated i.i.d.
- The training data can be of any finite size N .
- In general, we do not have any prior information about $p_{X,Y}$.

Supervised learning is usually concerned with the two following inference problems:

- Classification: Given $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^p \times \Delta^C$, for $i = 1, \dots, N$, we want to estimate for any new \mathbf{x} ,

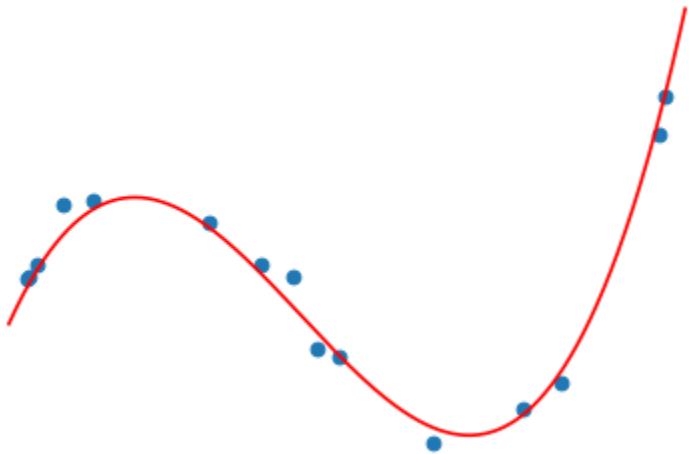
$$\arg \max_y p(Y = y | X = \mathbf{x}).$$

- Regression: Given $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^p \times \mathbb{R}$, for $i = 1, \dots, N$, we want to estimate for any new \mathbf{x} ,

$$\mathbb{E}[Y | X = \mathbf{x}].$$



Classification consists in identifying
a decision boundary between objects of distinct classes.



Regression aims at estimating relationships among (usually continuous) variables.

Probabilistic perspective

Supervised learning can be framed as probabilistic inference, where the goal is to estimate the conditional distribution

$$p(Y = y | X = \mathbf{x})$$

for any new (\mathbf{x}, y) .

This is the framing we will adopt in this course (starting from Lecture 2).

Empirical risk minimization

The traditional perspective on supervised learning is empirical risk minimization.

Consider a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ produced by some learning algorithm. The predictions of this function can be evaluated through a loss

$$\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R},$$

such that $\ell(y, f(\mathbf{x})) \geq 0$ measures how close the prediction $f(\mathbf{x})$ from y is.

Examples of loss functions

Classification:

$$\ell(y, f(\mathbf{x})) = \mathbf{1}_{y \neq f(\mathbf{x})}$$

Regression:

$$\ell(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$$

Let \mathcal{F} denote the hypothesis space, i.e. the set of all functions f than can be produced by the chosen learning algorithm.

We are looking for a function $f \in \mathcal{F}$ with a small **expected risk** (or generalization error)

$$R(f) = \mathbb{E}_{(\mathbf{x},y) \sim p_{X,Y}} [\ell(y, f(\mathbf{x}))].$$

This means that for a given data generating distribution $p_{X,Y}$ and for a given hypothesis space \mathcal{F} , the optimal model is

$$f_* = \arg \min_{f \in \mathcal{F}} R(f).$$

Since $p_{\mathbf{X}, \mathbf{Y}}$ is unknown, the expected risk cannot be evaluated and the optimal model cannot be determined.

However, if we have i.i.d. training data $\mathbf{d} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, N\}$, we can compute an estimate, the **empirical risk** (or training error)

$$\hat{R}(f, \mathbf{d}) = \frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathbf{d}} \ell(y_i, f(\mathbf{x}_i)).$$

This estimator is **unbiased** and can be used for finding a good enough approximation of f_* . This results into the **empirical risk minimization principle**:

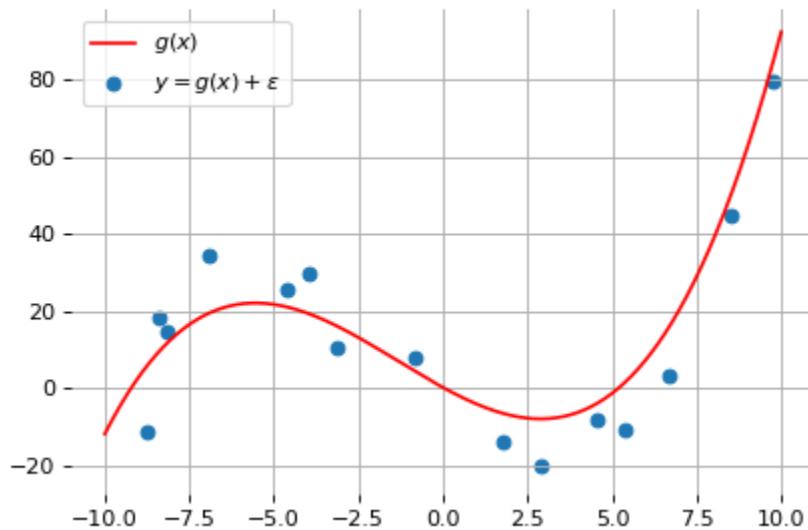
$$f_*^{\mathbf{d}} = \arg \min_{f \in \mathcal{F}} \hat{R}(f, \mathbf{d})$$

Most machine learning algorithms, including **neural networks**, implement empirical risk minimization.

Under regularity assumptions, empirical risk minimizers converge:

$$\lim_{N \rightarrow \infty} f_*^{\mathbf{d}} = f_*$$

Polynomial regression



Consider the joint probability distribution $p_{X,Y}$ induced by the data generating process

$$(x, y) \sim p_{X,Y} \Leftrightarrow x \sim U[-10; 10], \epsilon \sim \mathcal{N}(0, \sigma^2), y = g(x) + \epsilon$$

where $x \in \mathbb{R}$, $y \in \mathbb{R}$ and g is an unknown polynomial of degree 3.

Our goal is to find a function f that makes good predictions on average over $p_{X,Y}$.

Consider the hypothesis space $f \in \mathcal{F}$ of polynomials of degree 3 defined through their parameters $\mathbf{w} \in \mathbb{R}^4$ such that

$$\hat{y} \triangleq f(x; \mathbf{w}) = \sum_{d=0}^3 w_d x^d$$

For this regression problem, we use the squared error loss

$$\ell(y, f(x; \mathbf{w})) = (y - f(x; \mathbf{w}))^2$$

to measure how wrong the predictions are.

Therefore, our goal is to find the best value \mathbf{w}_* such that

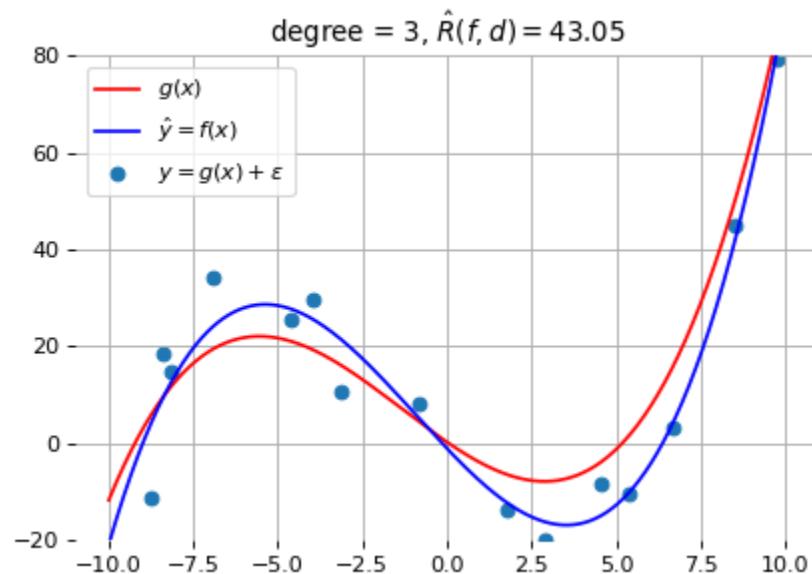
$$\begin{aligned}\mathbf{w}_* &= \arg \min_{\mathbf{w}} R(\mathbf{w}) \\ &= \arg \min_{\mathbf{w}} \mathbb{E}_{(x,y) \sim p_{X,Y}} [(y - f(x; \mathbf{w}))^2]\end{aligned}$$

Given a large enough training set $\mathbf{d} = \{(x_i, y_i) | i = 1, \dots, N\}$, the empirical risk minimization principle tells us that a good estimate $\mathbf{w}_*^{\mathbf{d}}$ of \mathbf{w}_* can be found by minimizing the empirical risk:

$$\begin{aligned}
 \mathbf{w}_*^{\mathbf{d}} &= \arg \min_{\mathbf{w}} \hat{R}(\mathbf{w}, \mathbf{d}) \\
 &= \arg \min_{\mathbf{w}} \frac{1}{N} \sum_{(x_i, y_i) \in \mathbf{d}} (y_i - f(x_i; \mathbf{w}))^2 \\
 &= \arg \min_{\mathbf{w}} \frac{1}{N} \sum_{(x_i, y_i) \in \mathbf{d}} (y_i - \sum_{d=0}^3 w_d x_i^d)^2 \\
 &= \arg \min_{\mathbf{w}} \frac{1}{N} \left\| \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}}_{\mathbf{y}} - \underbrace{\begin{pmatrix} x_1^0 \dots x_1^3 \\ x_2^0 \dots x_2^3 \\ \dots \\ x_N^0 \dots x_N^3 \end{pmatrix}}_{\mathbf{X}} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} \right\|^2
 \end{aligned}$$

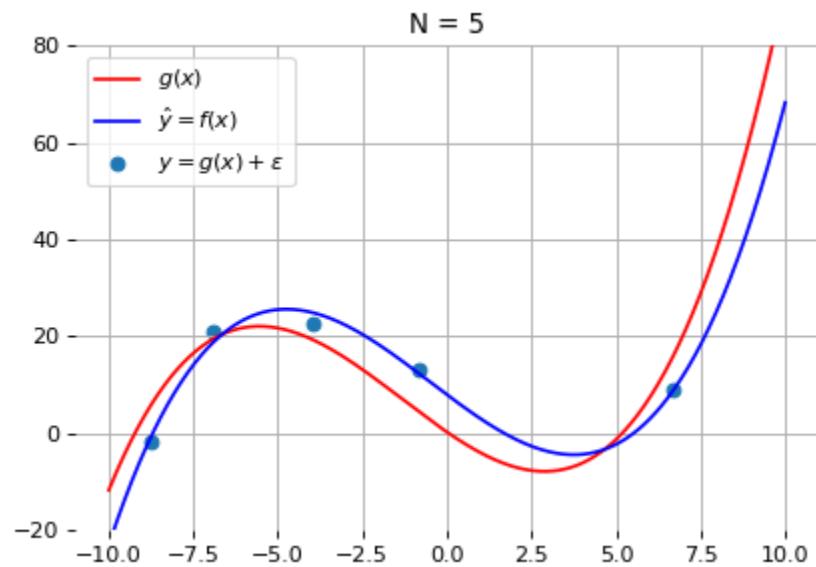
This is **ordinary least squares** regression, for which the solution is derived as

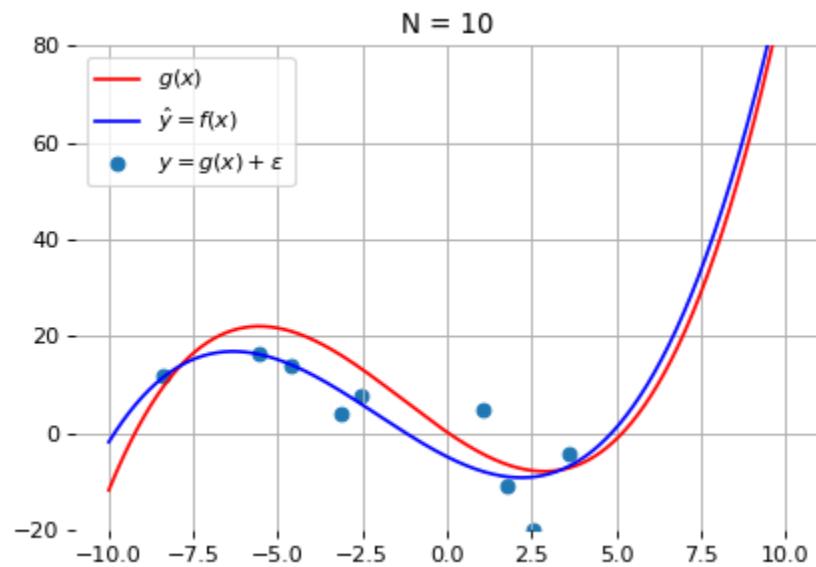
$$\mathbf{w}_*^d = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

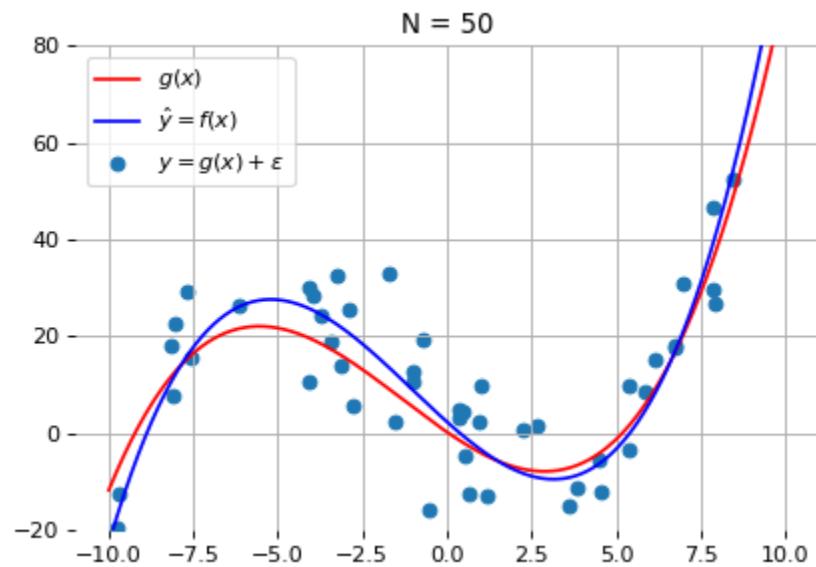


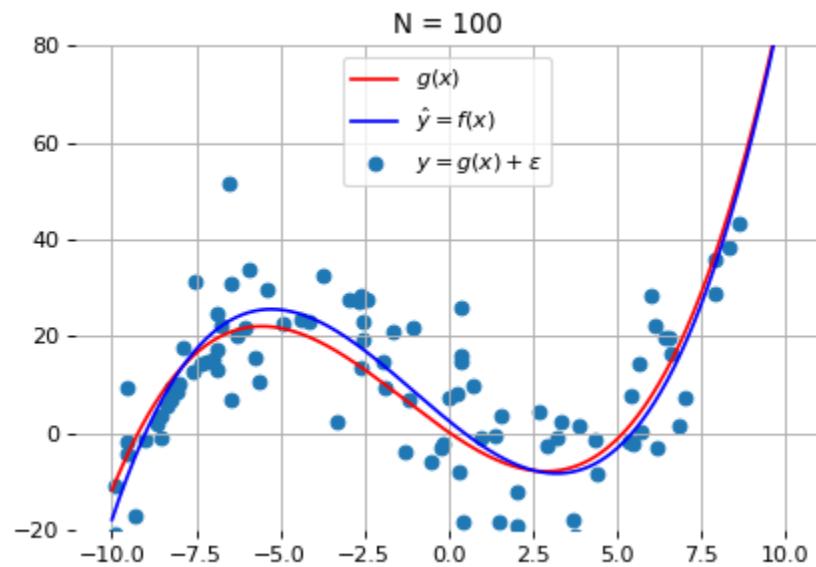
The expected risk minimizer \mathbf{w}_* within our hypothesis space is \mathbf{g} itself.

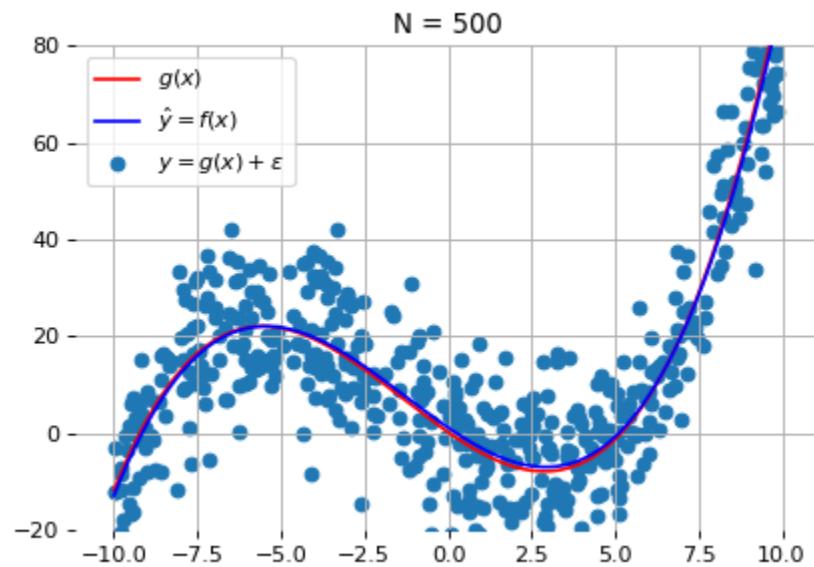
Therefore, on this toy problem, we can verify that $f(x; \mathbf{w}_*) \rightarrow f(x; \mathbf{w}_*) = g(x)$ as $N \rightarrow \infty$.





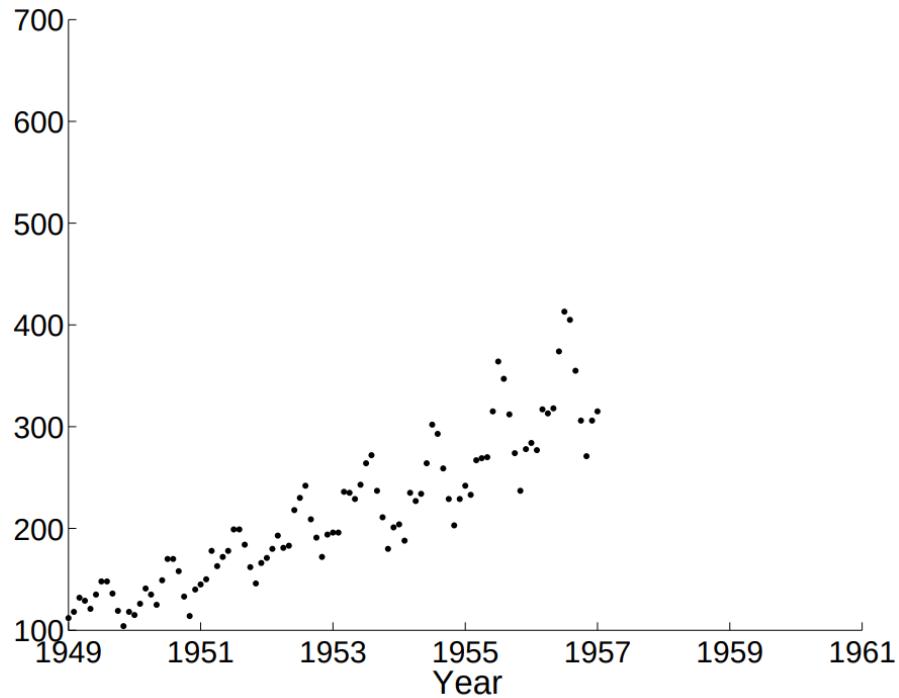






Under-fitting and over-fitting

What if we consider a hypothesis space \mathcal{F} in which candidate functions f are either too "simple" or too "complex" with respect to the true data generating process?

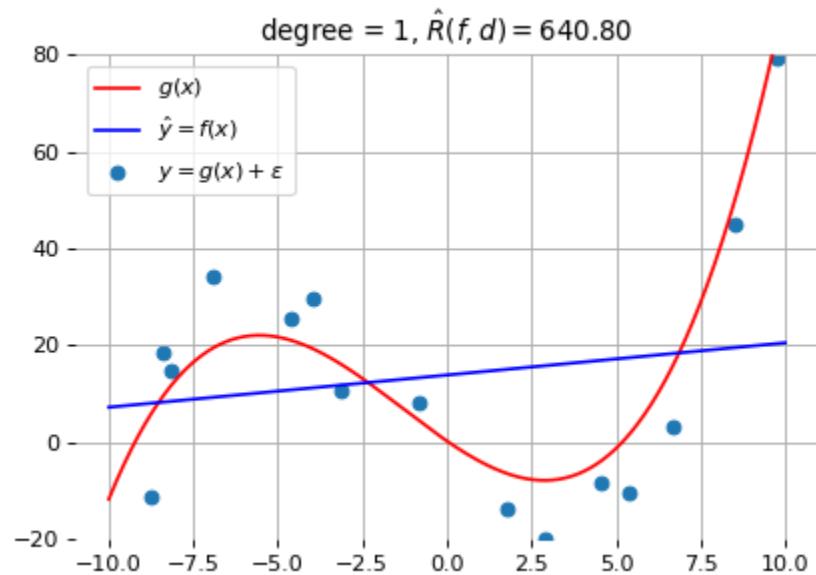


Which model would you choose?

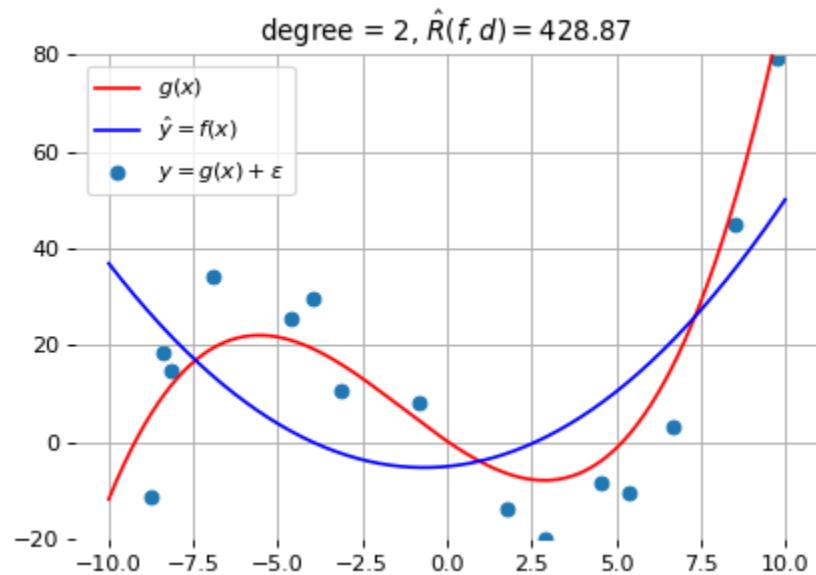
$$f_1(x) = w_0 + w_1 x$$

$$f_2(x) = \sum_{j=0}^3 w_j x^j$$

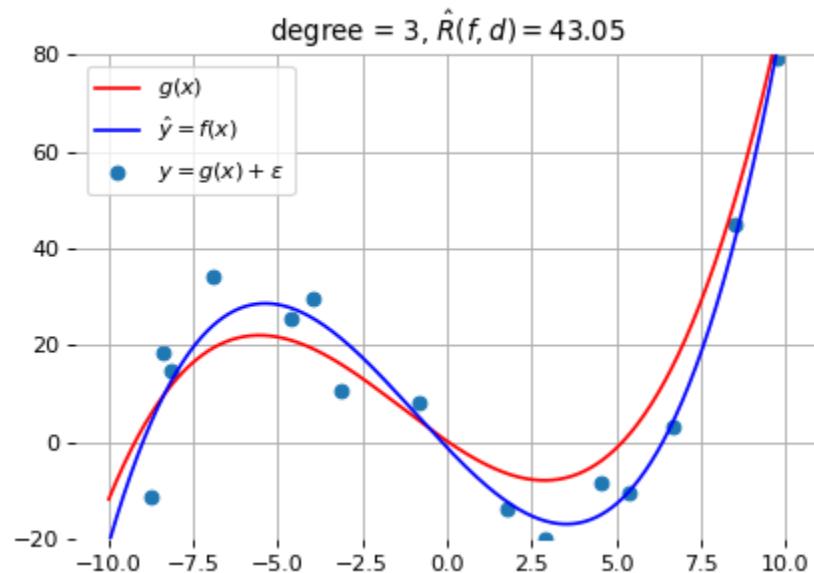
$$f_3(x) = \sum_{j=0}^{10^4} w_j x^j$$



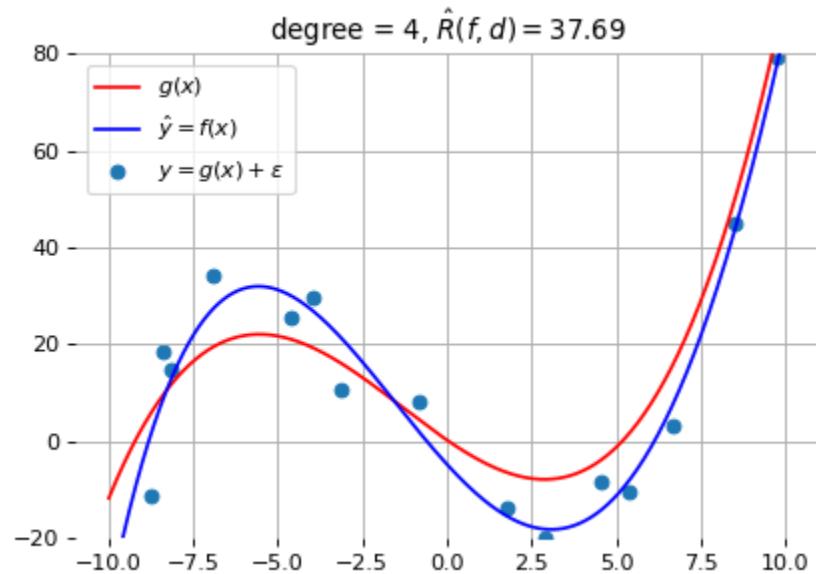
\mathcal{F} = polynomials of degree 1



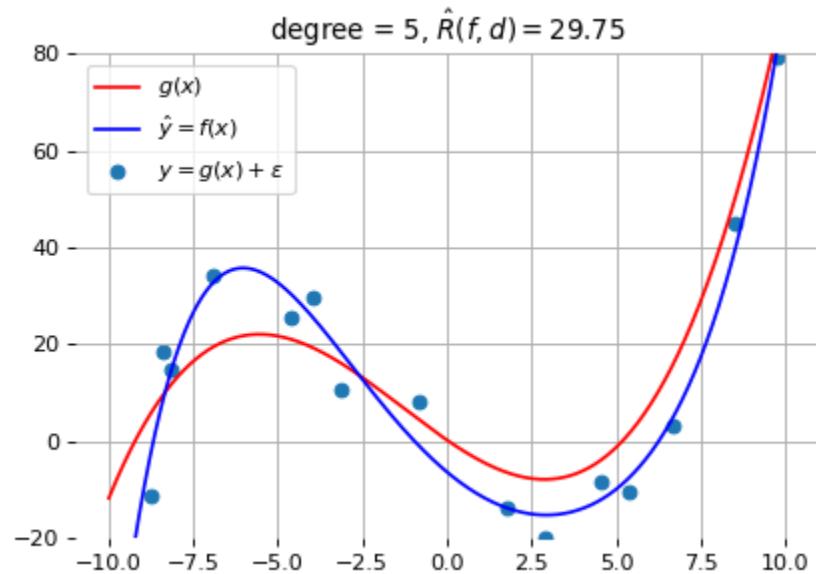
\mathcal{F} = polynomials of degree 2



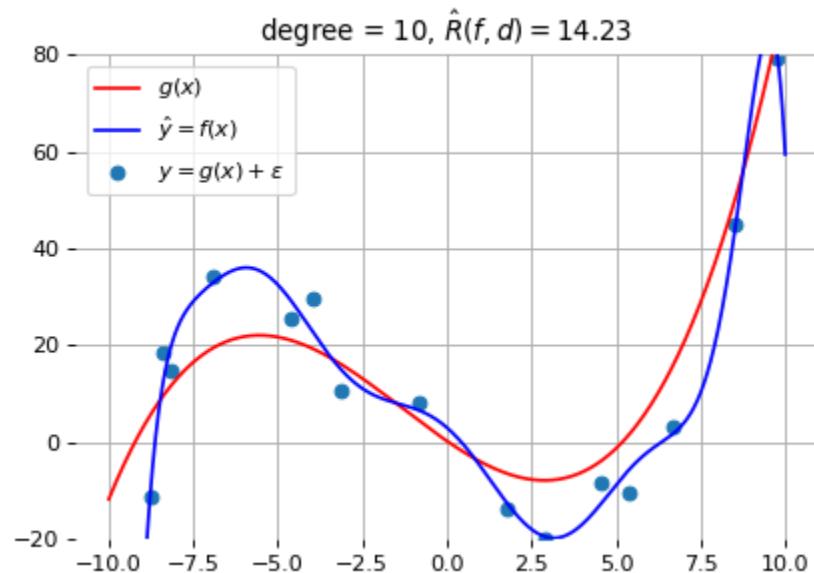
\mathcal{F} = polynomials of degree 3



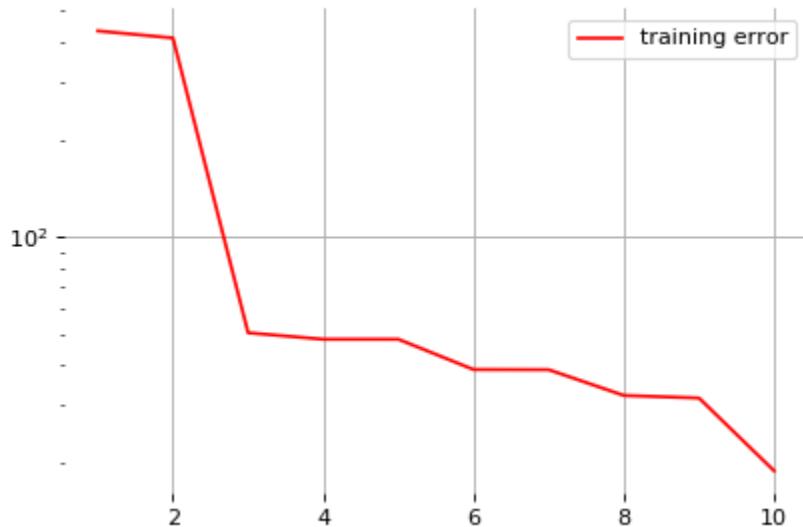
\mathcal{F} = polynomials of degree 4



\mathcal{F} = polynomials of degree 5



\mathcal{F} = polynomials of degree 10



Degree d of the polynomial VS. error.

Let $\mathcal{Y}^{\mathcal{X}}$ be the set of all functions $f : \mathcal{X} \rightarrow \mathcal{Y}$.

We define the **Bayes risk** as the minimal expected risk over all possible functions,

$$R_B = \min_{f \in \mathcal{Y}^{\mathcal{X}}} R(f),$$

and call the **Bayes optimal model** the model f_B that achieves this minimum.

No model f can perform better than f_B .

The **capacity** of an hypothesis space induced by a learning algorithm intuitively represents the ability to find a good model $f \in \mathcal{F}$ for any function, regardless of its complexity.

In practice, capacity can be controlled through hyper-parameters of the learning algorithm. For example:

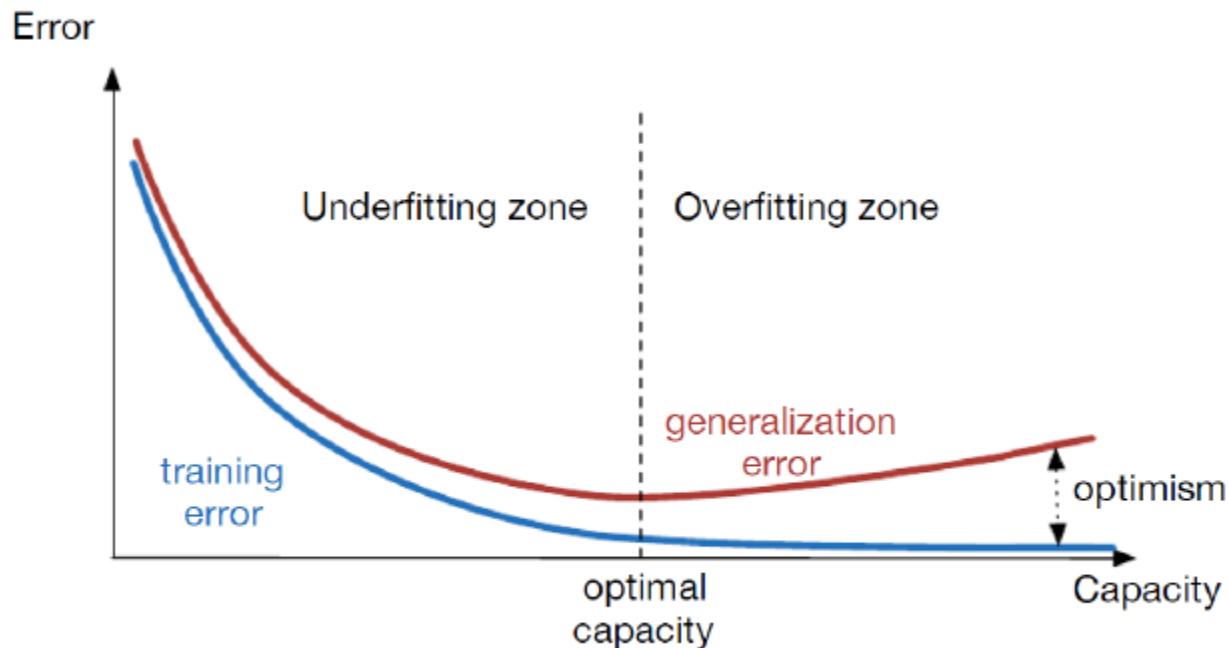
- The degree of the family of polynomials;
- The number of layers in a neural network;
- The number of training iterations;
- Regularization terms.

- If the capacity of \mathcal{F} is too low, then $f_B \notin \mathcal{F}$ and $R(f) - R_B$ is large for any $f \in \mathcal{F}$, including f_* and $f_*^{\mathbf{d}}$. Such models f are said to **underfit** the data.
- If the capacity of \mathcal{F} is too high, then $f_B \in \mathcal{F}$ or $R(f_*) - R_B$ is small. However, because of the high capacity of the hypothesis space, the empirical risk minimizer $f_*^{\mathbf{d}}$ could fit the training data arbitrarily well such that

$$R(f_*^{\mathbf{d}}) \geq R_B \geq \hat{R}(f_*^{\mathbf{d}}, \mathbf{d}) \geq 0.$$

In this situation, $f_*^{\mathbf{d}}$ becomes too specialized with respect to the true data generating process and a large reduction of the empirical risk (often) comes at the price of an increase of the expected risk of the empirical risk minimizer $R(f_*^{\mathbf{d}})$. In this situation, $f_*^{\mathbf{d}}$ is said to **overfit** the data.

Therefore, our goal is to adjust the capacity of the hypothesis space such that the expected risk of the empirical risk minimizer gets as low as possible.



When overfitting,

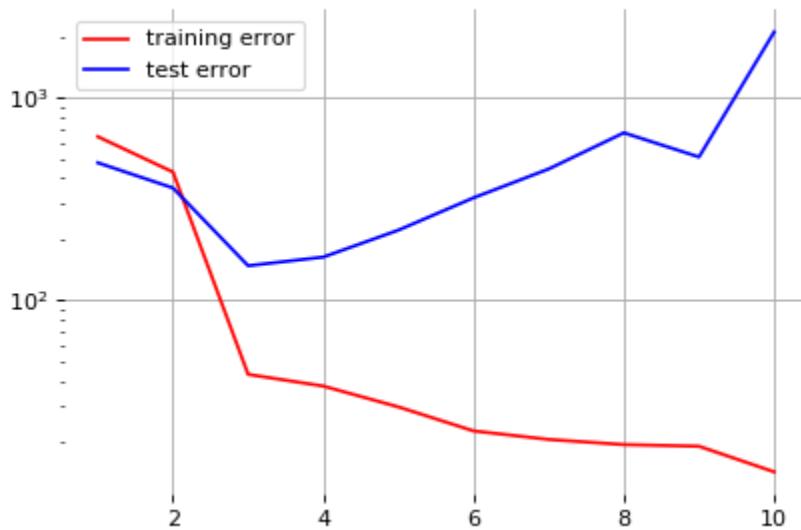
$$R(f_*^{\mathbf{d}}) \geq R_B \geq \hat{R}(f_*^{\mathbf{d}}, \mathbf{d}) \geq 0.$$

This indicates that the empirical risk $\hat{R}(f_*^{\mathbf{d}}, \mathbf{d})$ is a poor estimator of the expected risk $R(f_*^{\mathbf{d}})$.

Nevertheless, an unbiased estimate of the expected risk can be obtained by evaluating $f_*^{\mathbf{d}}$ on data \mathbf{d}_{test} independent from the training samples \mathbf{d} :

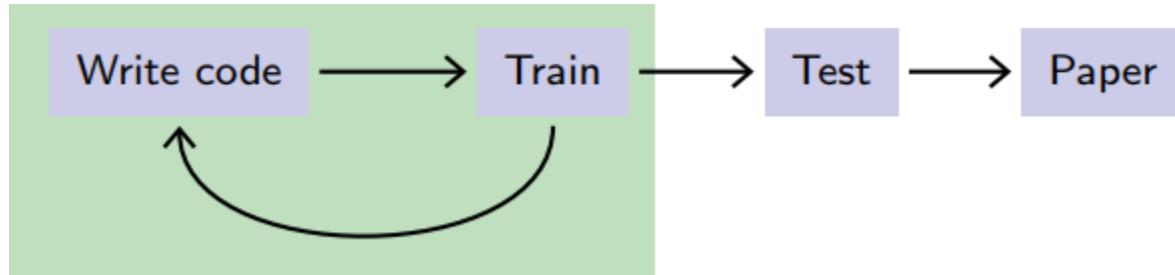
$$\hat{R}(f_*^{\mathbf{d}}, \mathbf{d}_{\text{test}}) = \frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathbf{d}_{\text{test}}} \ell(y_i, f_*^{\mathbf{d}}(\mathbf{x}_i))$$

This **test error** estimate can be used to evaluate the actual performance of the model. However, it should not be used, at the same time, for model selection.

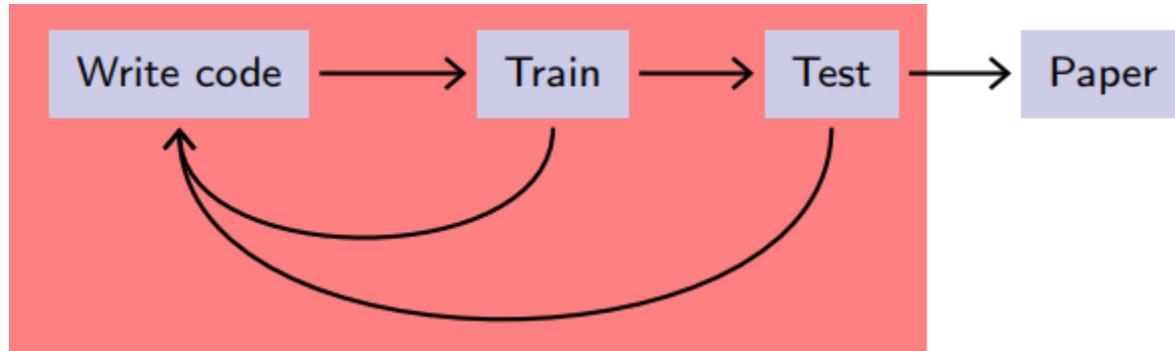


Degree d of the polynomial VS. error.

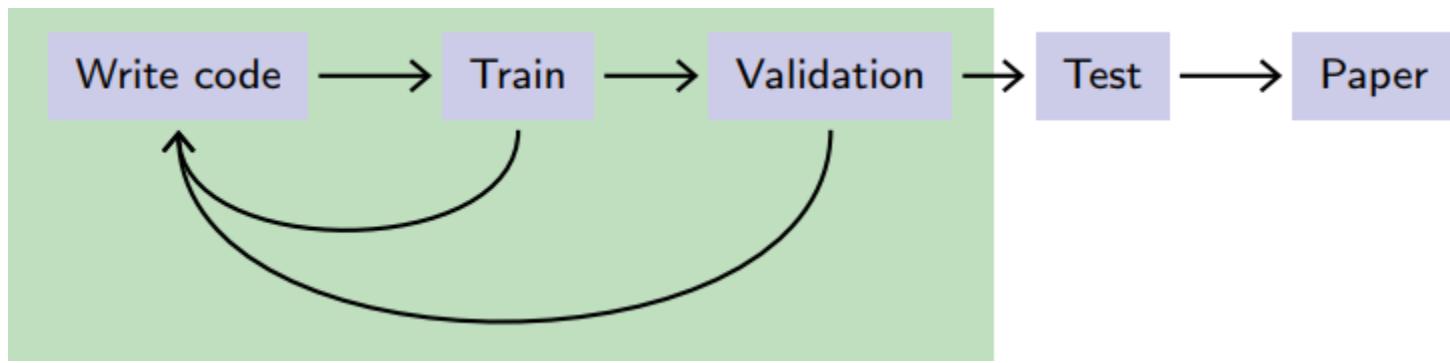
(Proper) evaluation protocol



There may be over-fitting, but it does not bias the final performance evaluation.



This should be **avoided** at all costs!



Instead, keep a separate validation set for tuning the hyper-parameters.

Bias-variance decomposition

Consider a fixed point \mathbf{x} and the prediction $\hat{Y} = f_*^{\mathbf{d}}(\mathbf{x})$ of the empirical risk minimizer at \mathbf{x} .

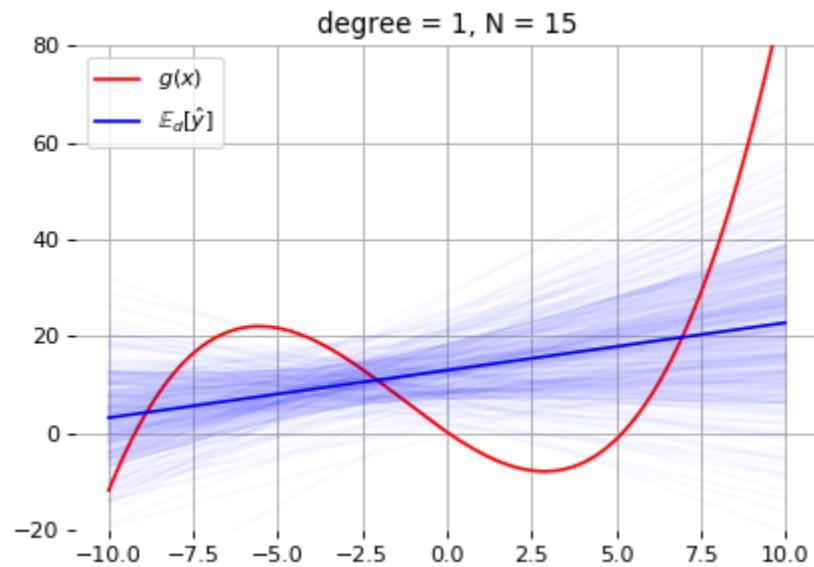
Then the local expected risk of $f_*^{\mathbf{d}}$ is

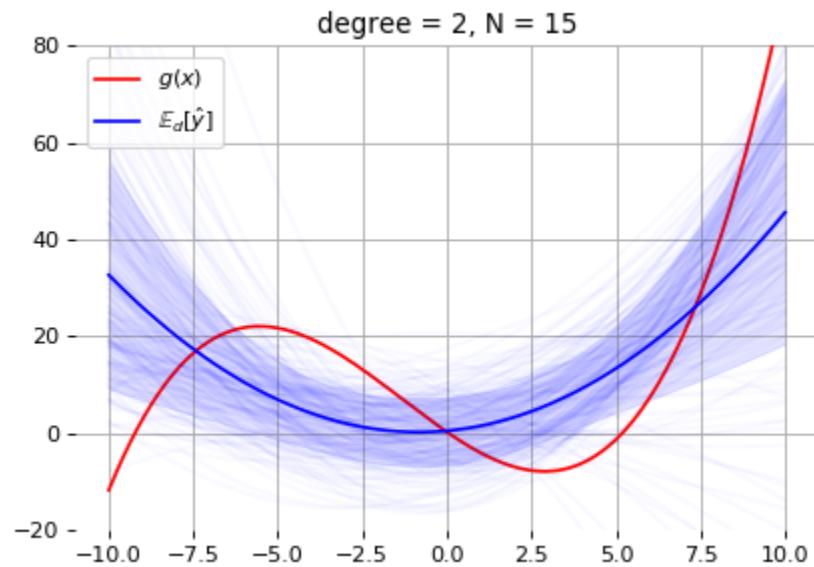
$$\begin{aligned} R(f_*^{\mathbf{d}}|\mathbf{x}) &= \mathbb{E}_{y \sim p_{Y|x}} [(y - f_*^{\mathbf{d}}(\mathbf{x}))^2] \\ &= \mathbb{E}_{y \sim p_{Y|x}} [(y - f_B(\mathbf{x}) + f_B(\mathbf{x}) - f_*^{\mathbf{d}}(\mathbf{x}))^2] \\ &= \mathbb{E}_{y \sim p_{Y|x}} [(y - f_B(\mathbf{x}))^2] + \mathbb{E}_{y \sim p_{Y|x}} [(f_B(\mathbf{x}) - f_*^{\mathbf{d}}(\mathbf{x}))^2] \\ &= R(f_B|\mathbf{x}) + (f_B(\mathbf{x}) - f_*^{\mathbf{d}}(\mathbf{x}))^2 \end{aligned}$$

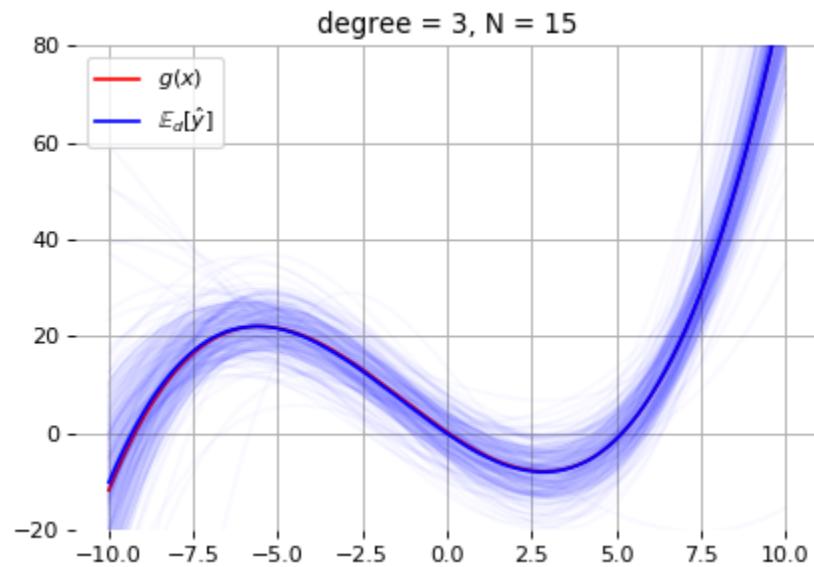
where

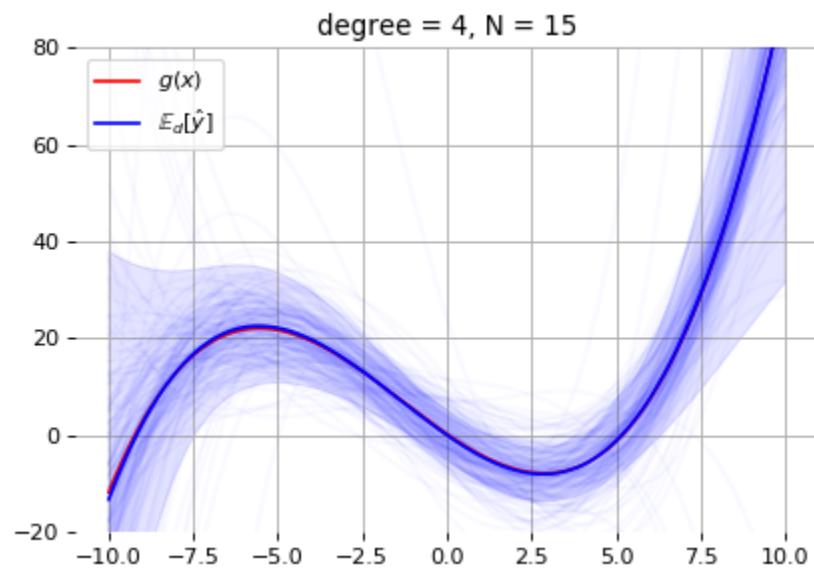
- $R(f_B|\mathbf{x})$ is the local expected risk of the Bayes model. This term cannot be reduced.
- $(f_B(\mathbf{x}) - f_*^{\mathbf{d}}(\mathbf{x}))^2$ represents the discrepancy between f_B and $f_*^{\mathbf{d}}$.

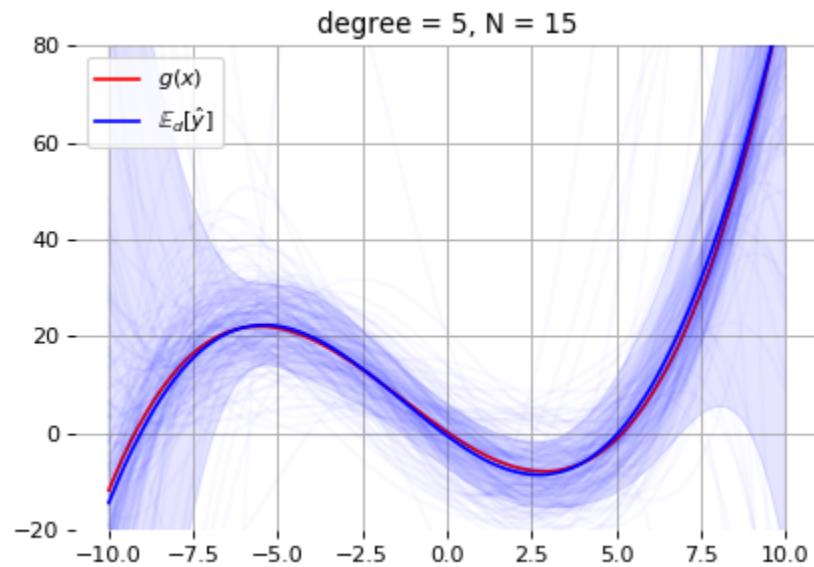
If $\mathbf{d} \sim p_{X,Y}$ is itself considered as a random variable, then $f_*^{\mathbf{d}}$ is also a random variable, along with its predictions \hat{Y} .











Formally, the expected local expected risk yields to:

$$\begin{aligned}\mathbb{E}_{\mathbf{d}} [R(f_*^{\mathbf{d}}|x)] &= \mathbb{E}_{\mathbf{d}} [R(f_B|x) + (f_B(x) - f_*^{\mathbf{d}}(x))^2] \\ &= R(f_B|x) + \mathbb{E}_{\mathbf{d}} [(f_B(x) - f_*^{\mathbf{d}}(x))^2] \\ &= \underbrace{R(f_B|x)}_{\text{noise}(x)} + \underbrace{(f_B(x) - \mathbb{E}_{\mathbf{d}} [f_*^{\mathbf{d}}(x)])^2}_{\text{bias}^2(x)} + \underbrace{\mathbb{E}_{\mathbf{d}} [(\mathbb{E}_{\mathbf{d}} [f_*^{\mathbf{d}}(x)] - f_*^{\mathbf{d}}(x))^2]}_{\text{var}(x)}\end{aligned}$$

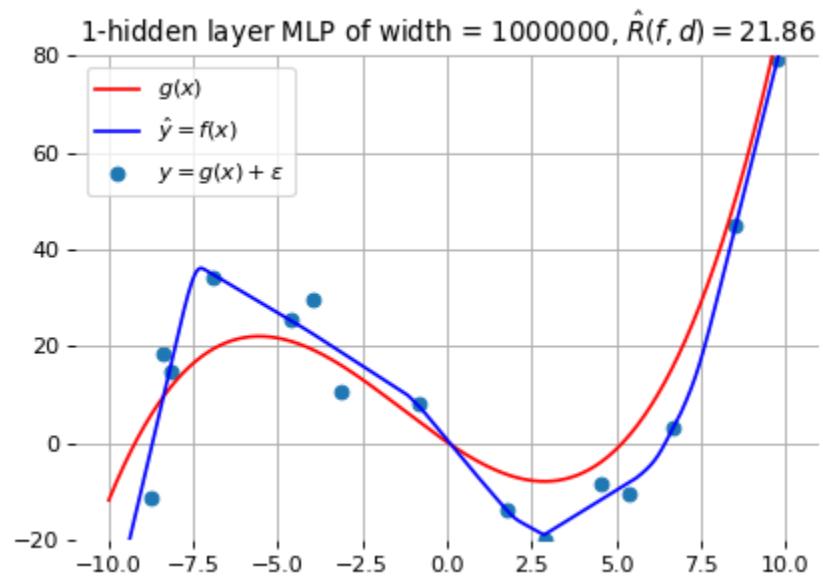
This decomposition is known as the **bias-variance** decomposition.

- The noise term quantifies the irreducible part of the expected risk.
- The bias term measures the discrepancy between the average model and the Bayes model.
- The variance term quantifies the variability of the predictions.

Bias-variance trade-off

- Reducing the capacity makes f_*^d fit the data less on average, which increases the bias term.
- Increasing the capacity makes f_*^d vary a lot with the training data, which increases the variance term.

What about a neural network with **millions** of parameters?



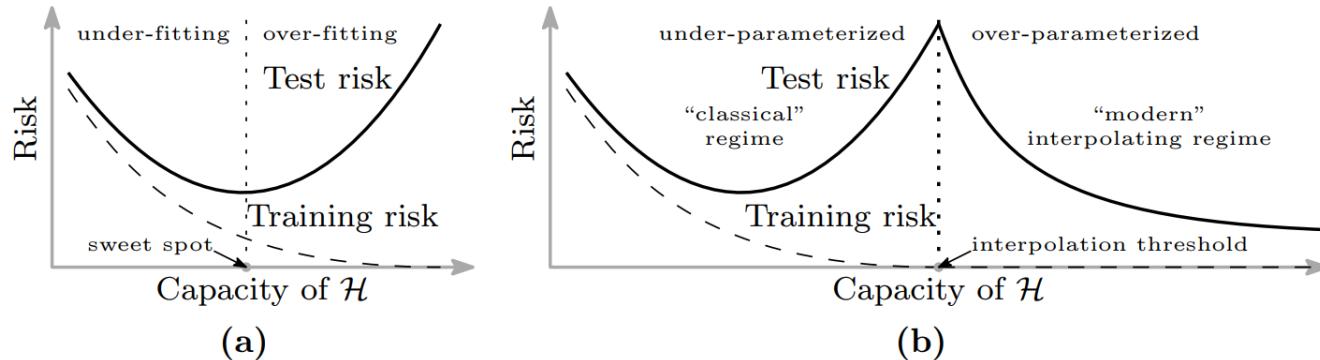


Figure 1: **Curves for training risk (dashed line) and test risk (solid line).** (a) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (b) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the “classical” regime) together with the observed behavior from using high capacity function classes (i.e., the “modern” interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

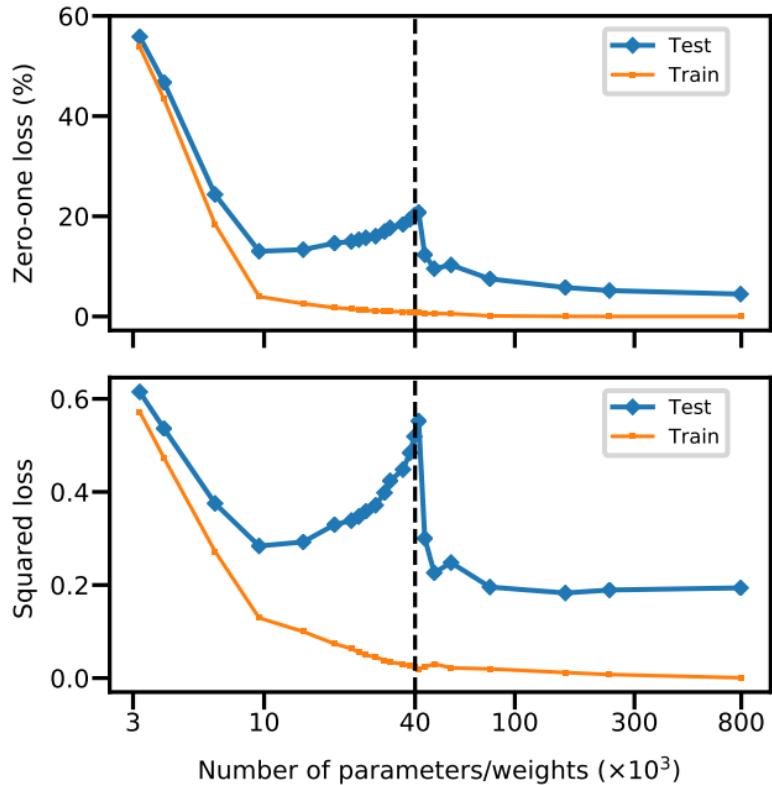


Figure 4: Double descent risk curve for fully connected neural network on MNIST. Training and test risks of network with a single layer of H hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d+1) \cdot H + (H+1) \cdot K$. The interpolation threshold (black dotted line) is observed at $n \cdot K$.

