ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

inf | Informatik
Computer Science

Introduction to Machine Learning

Generalization and Model Validation

Prof. Andreas Krause
Learning and Adaptive Systems (las.ethz.ch)

Least-squares linear regression optimization

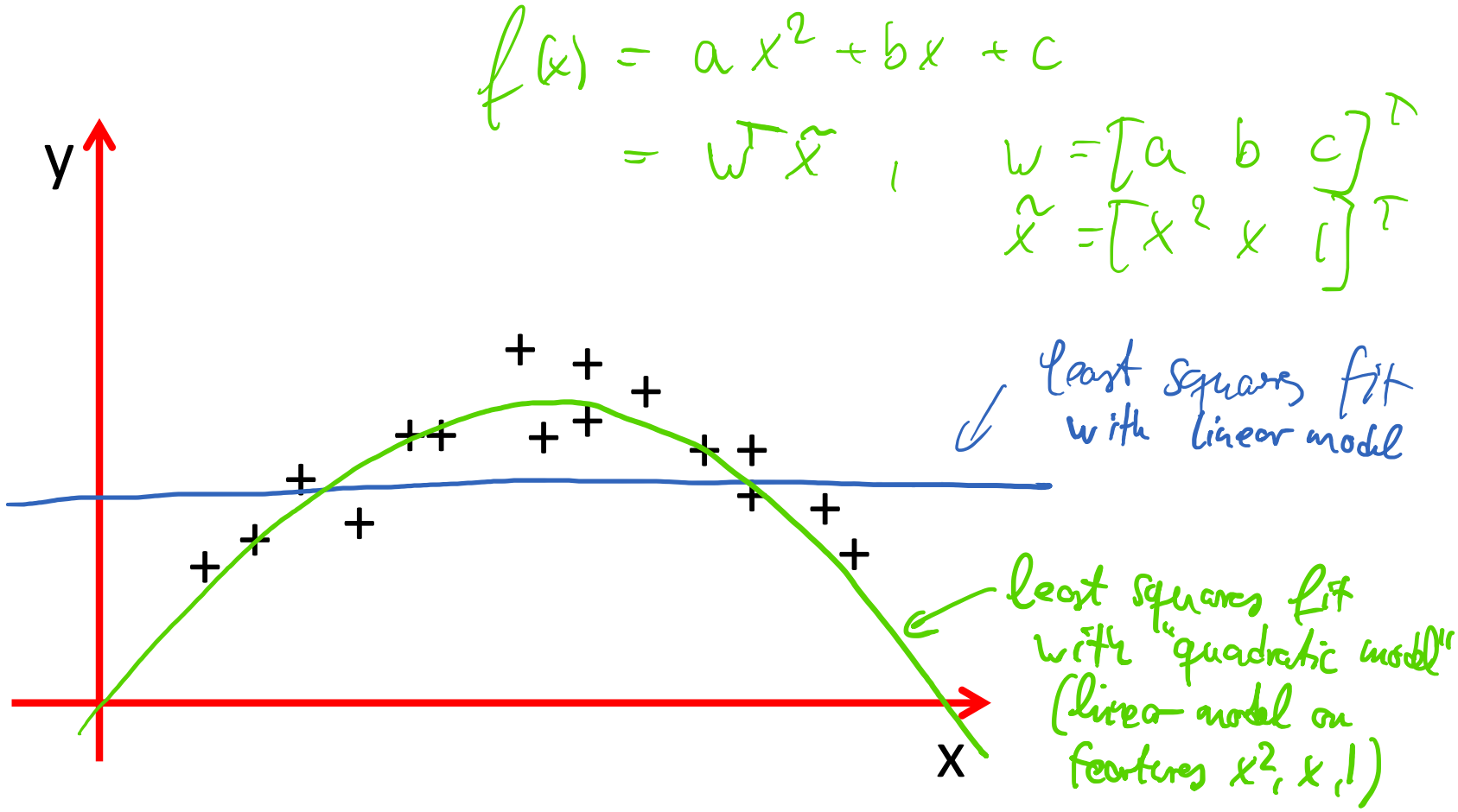
[Legendre 1805, Gauss 1809]

- Given data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Fitting nonlinear functions

- How about functions like this:



Linear regression for polynomials

We can fit non-linear functions via linear regression, using nonlinear features of our data (basis functions)

$$f(\mathbf{x}) = \sum_{i=1}^{D} w_i \phi_i(\mathbf{x})$$

$x \mapsto \tilde{x} = \phi(x)$

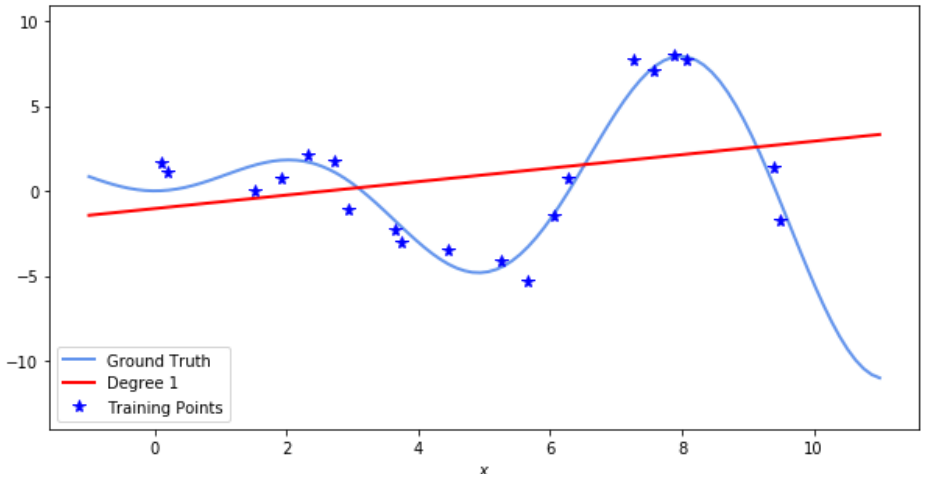
$D = |\phi(x)|$

In 1-d: $\phi(x) = [1, x, x^2, \ldots x^k]$

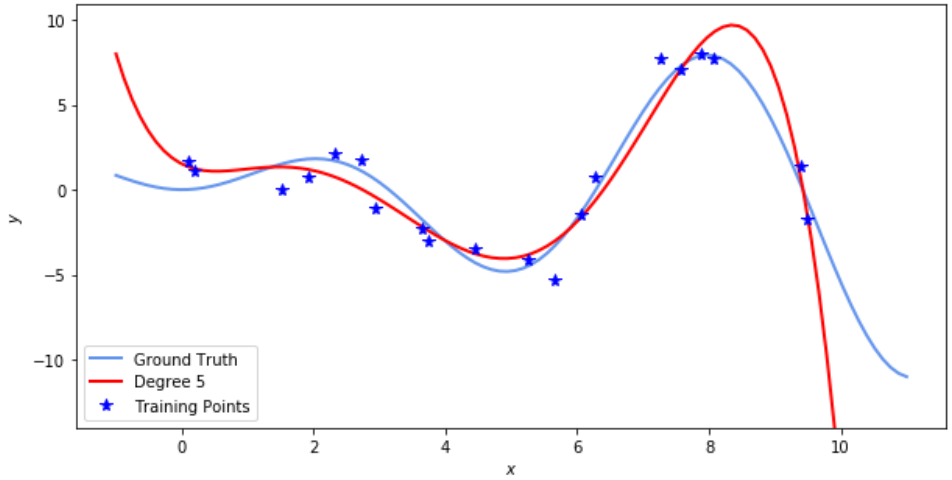
In 2-d: $\phi(x) = [1, x_1, x_2, x_1^2, x_2^2, x_1 \cdot x_2, \ldots]$

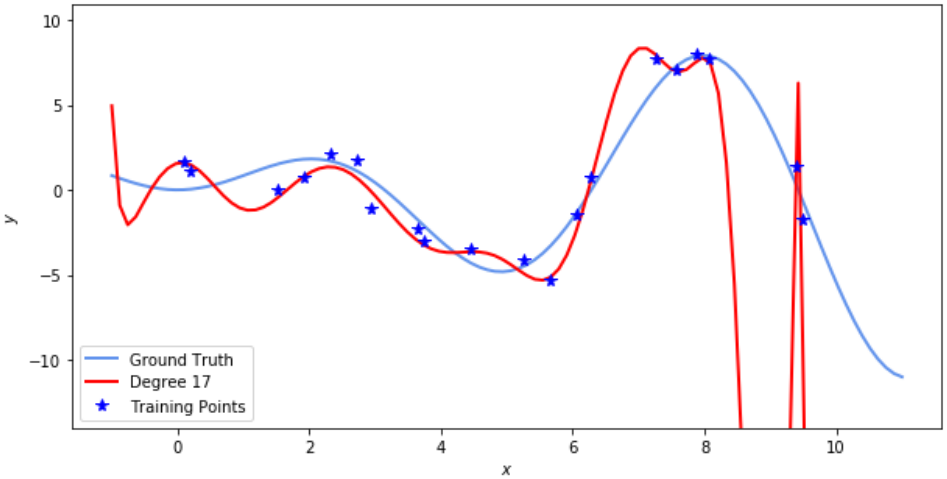
In d-d: $\phi(x)$ = vector of all monomials of degree up to k in $x_1 \ldots x_d$

Demo: Linear regression on polynomials



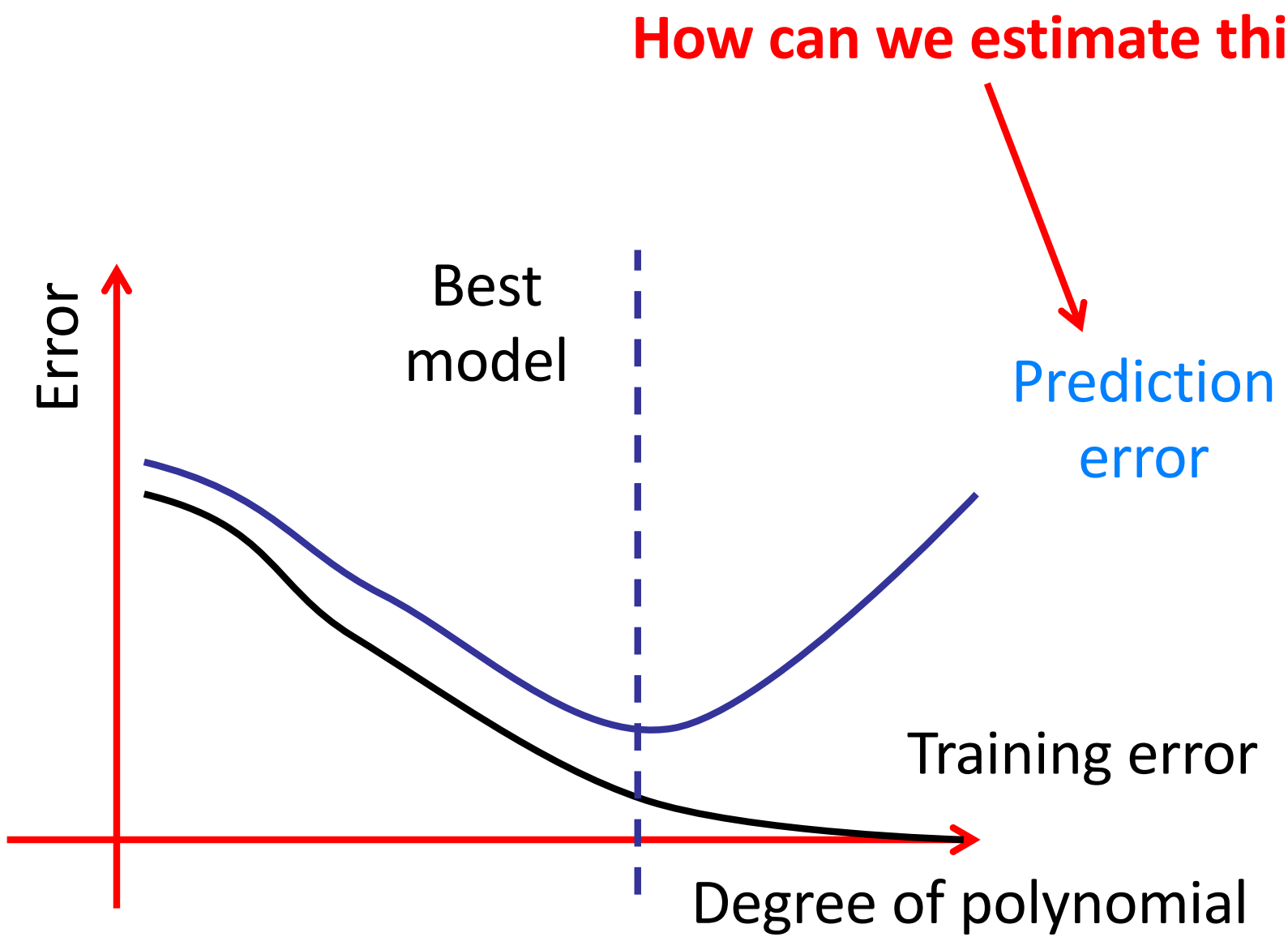
Underfitting





Overfitting

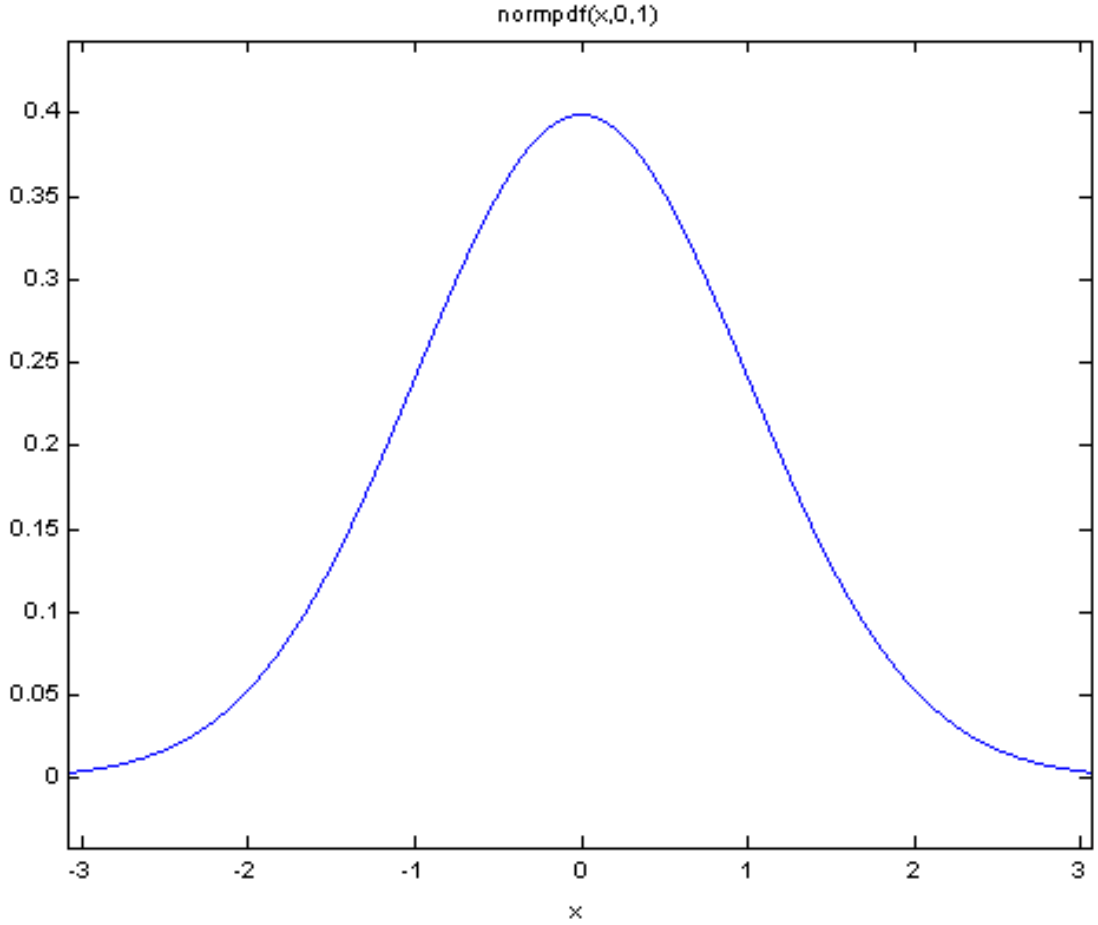
Model selection for linear regression with polynomials



Interlude: A note on probability

- You'll need to know about basic concepts in probability:
  - Random variables
  - Expectations (Mean, Variance etc.)
  - Independence (i.i.d. samples from a distribution, ...)
  - ...

Example: Gaussian distribution

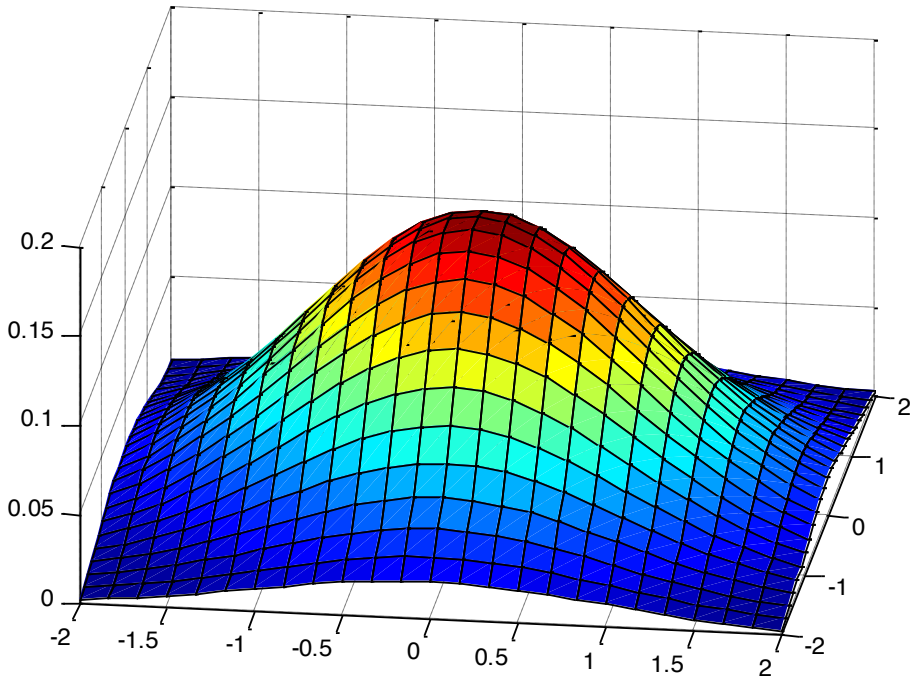


- σ = Standard deviation
- μ = mean

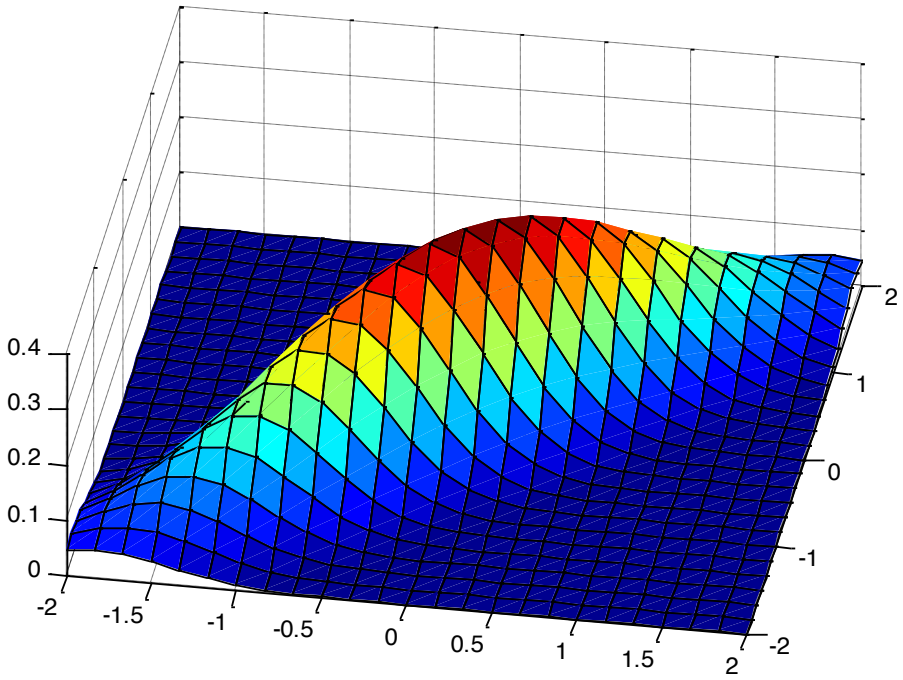
$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Example: Multivariate Gaussian

$$\frac{1}{2\pi\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \qquad \Sigma = \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix} \qquad \mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$$



$$\Sigma = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$



$$\Sigma = \begin{pmatrix}1 & 0.9\\ 0.9 & 1\end{pmatrix}$$

Interlude: Expectations

- Expected value of random variable X

$$\mathbb{E}[X] = \begin{cases} \sum_x p(x) \cdot x & \text{if } X \text{ discrete (e.g. Bernoulli)} \\ \int p(x) \cdot x \, dx & \text{if } X \text{ continuous (e.g. Gaussian)} \end{cases}$$

- Expected value of some function of X

$$\mathbb{E}[f(X)] = \begin{cases} \sum_x p(x) \cdot f(x) \\ \int p(x) f(x) \, dx \end{cases}$$

$$\text{E.g. } \mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- Linearity of expectation X, Y RVs (can be dependent!), $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$$

Achieving generalization

- Fundamental assumption: Our data set is generated independently and identically distributed (iid) from some unknown distribution *P*

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

- Our goal is to minimize the *expected error (true risk)* under *P*

$$R(\mathbf{w}) = \int P(\mathbf{x}, y)(y - \mathbf{w}^T\mathbf{x})^2 d\mathbf{x}dy$$

$$= \mathbb{E}_{\mathbf{x},y}[(y - \mathbf{w}^T\mathbf{x})^2]$$

Estimating the generalization error

- Estimate the true risk by the empirical risk on a sample data set D

$$\hat{R}_D(\mathbf{w}) = \frac{1}{|D|} \sum_{(\mathbf{x},y)\in D} (y - \mathbf{w}^T\mathbf{x})^2$$

- Why might this work?

Law of large numbers $\hat{R}_D(\mathbf{w}) \to R(\mathbf{w})$
for any fixed **w** almost surely as $|D| \to \infty$

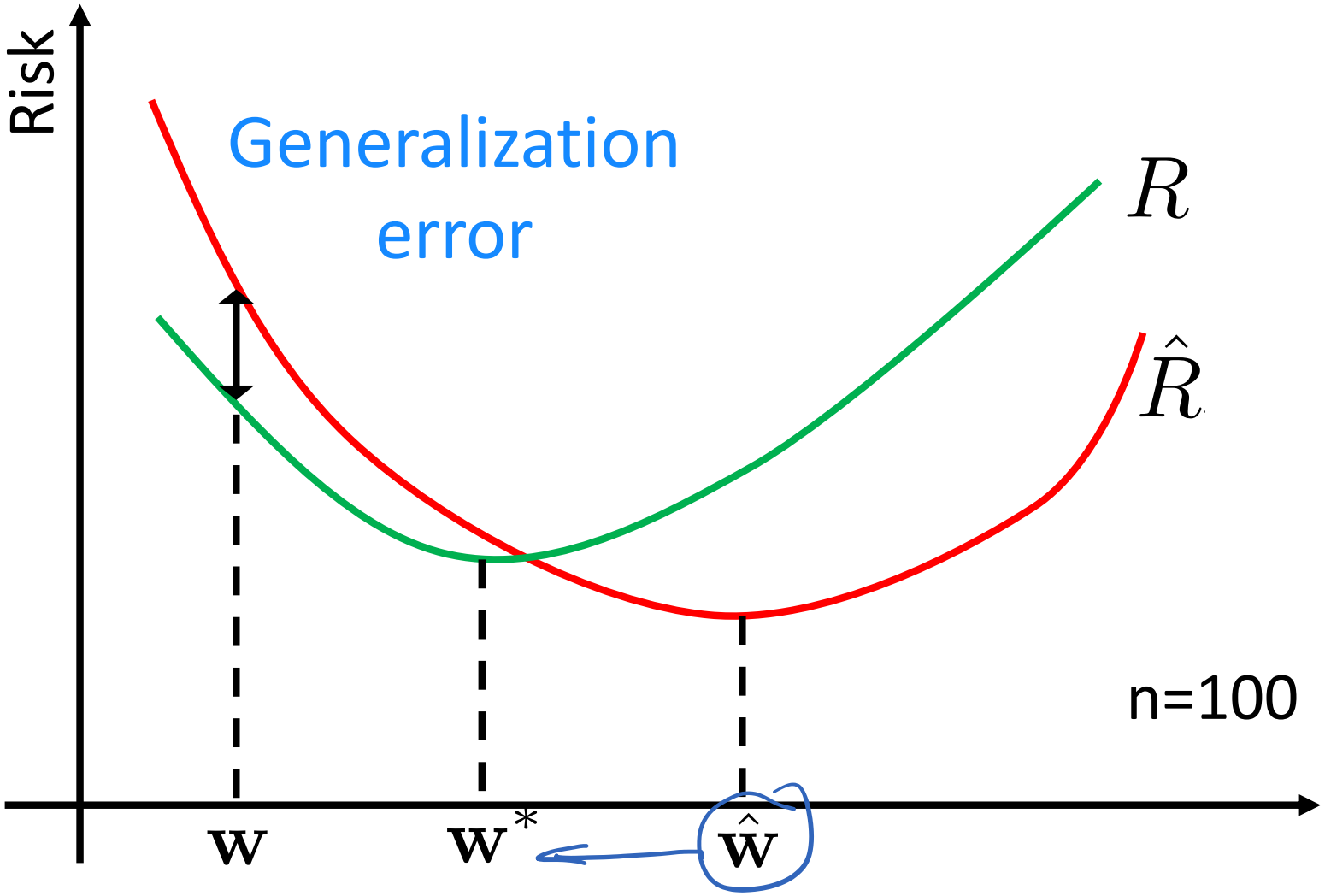
What happens if we optimize on training data?

- Suppose we are given training data D

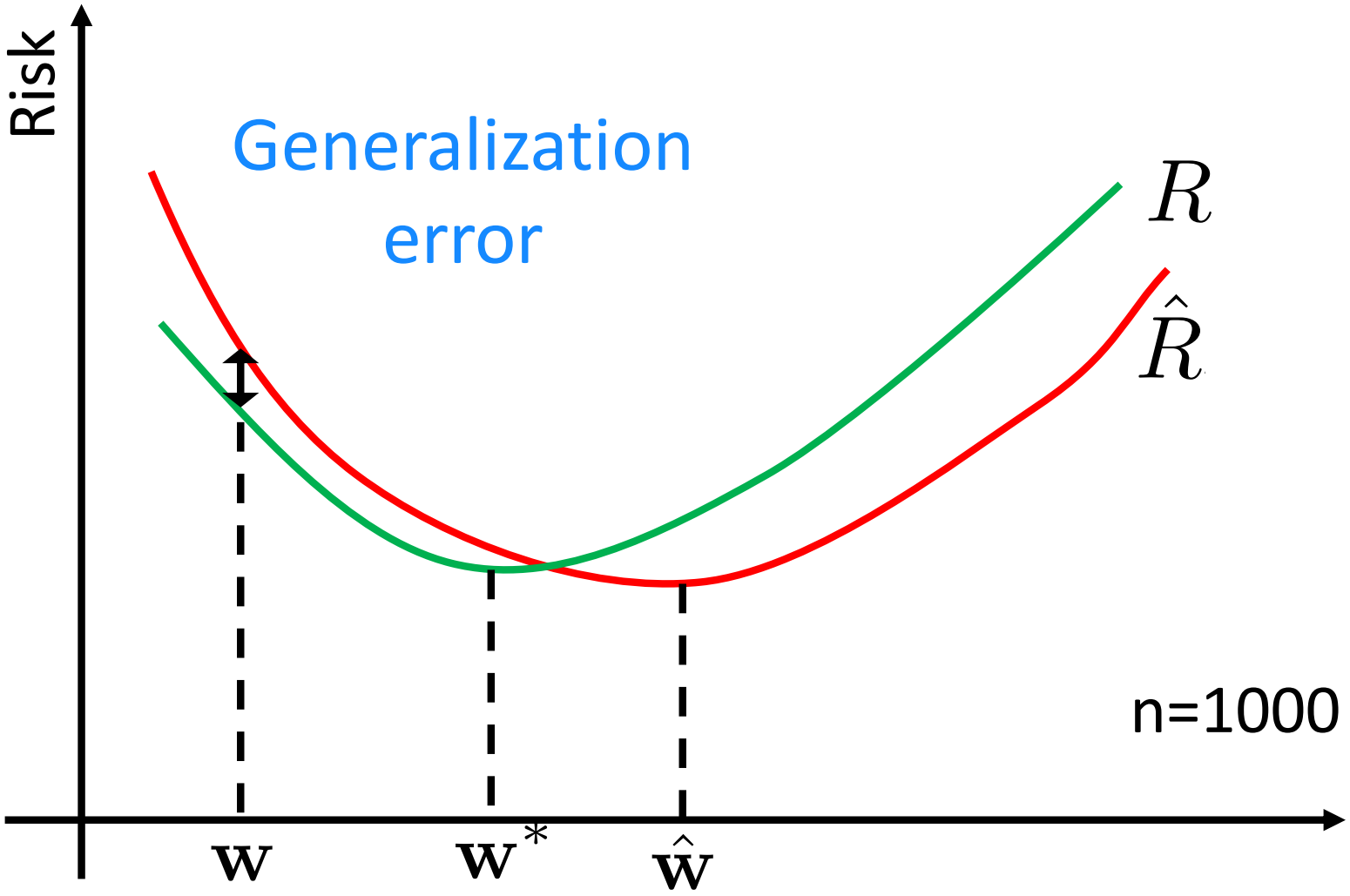
  Empirical Risk Minimization: $\hat{\mathbf{w}}_D = \operatorname{argmin}_{\mathbf{w}} \hat{R}_D(\mathbf{w})$

- Ideally, we wish to solve $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} R(\mathbf{w})$

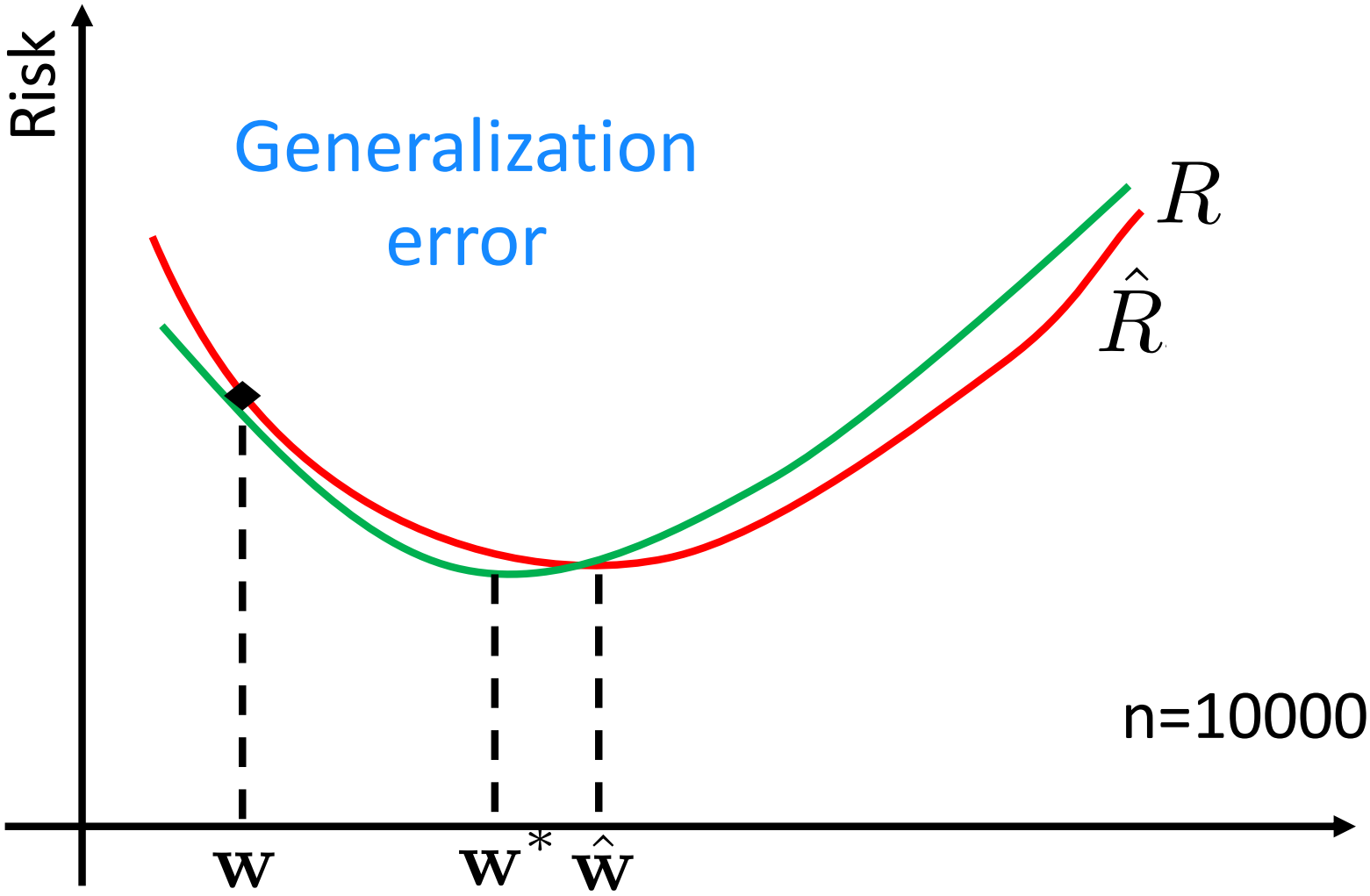
Empirical vs true risk



Empirical vs true risk



Empirical vs true risk



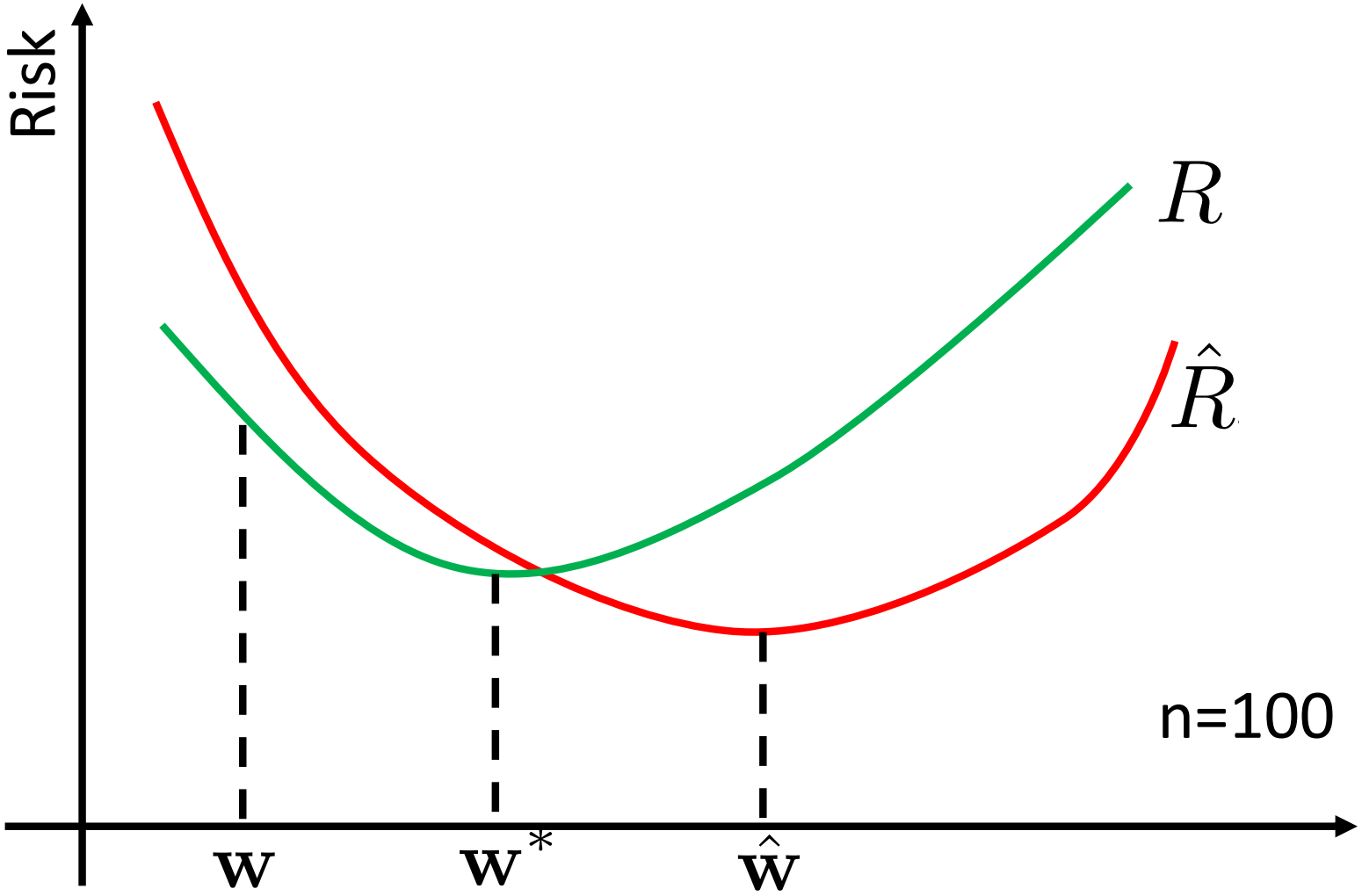
Outlook: Requirements for learning

- For learning via empirical risk minimization to be successful, need uniform convergence:

$$\sup_{\mathbf{w}} |R(\mathbf{w}) - \hat{R}_D(\mathbf{w})| \to 0 \text{ as } |D| \to \infty$$

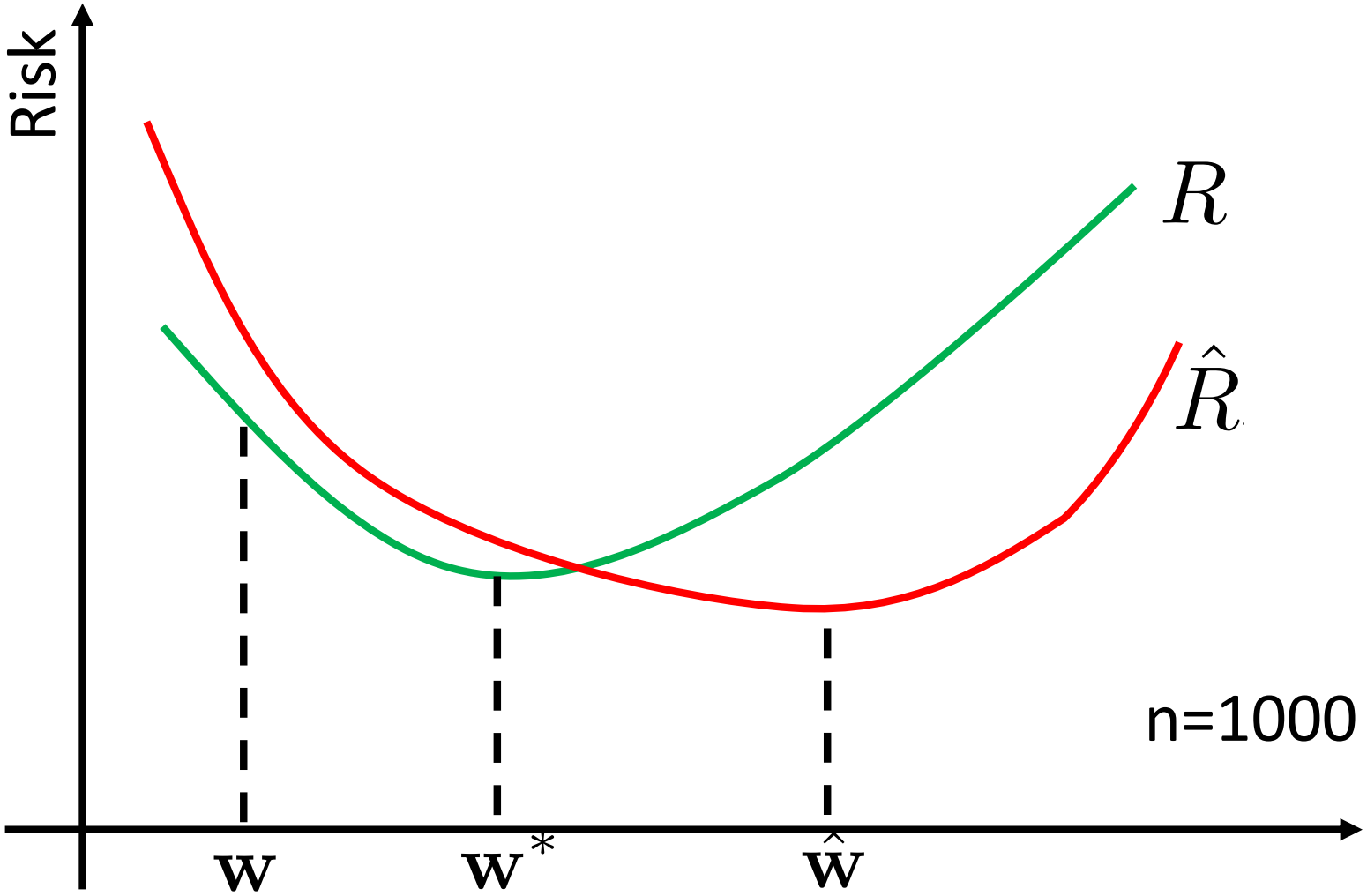
- This is not implied by law of large numbers alone, but depends on model class (holds, e.g., for squared loss on data distributions with bounded support)
  ➔ Statistical learning theory

What can go wrong in ERM

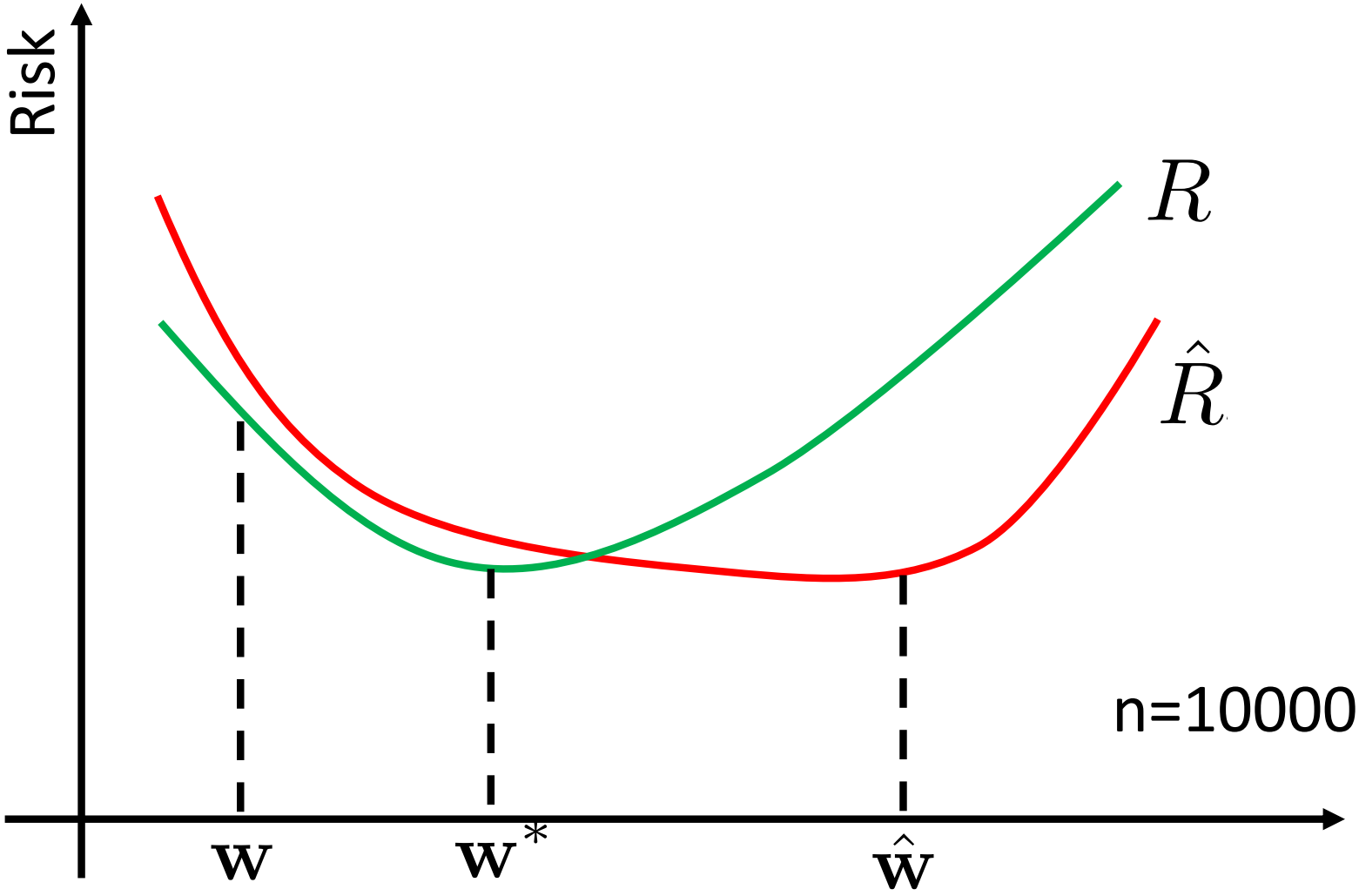


$n = |D|$

What can go wrong in ERM



What can go wrong in ERM



Learning from finite data

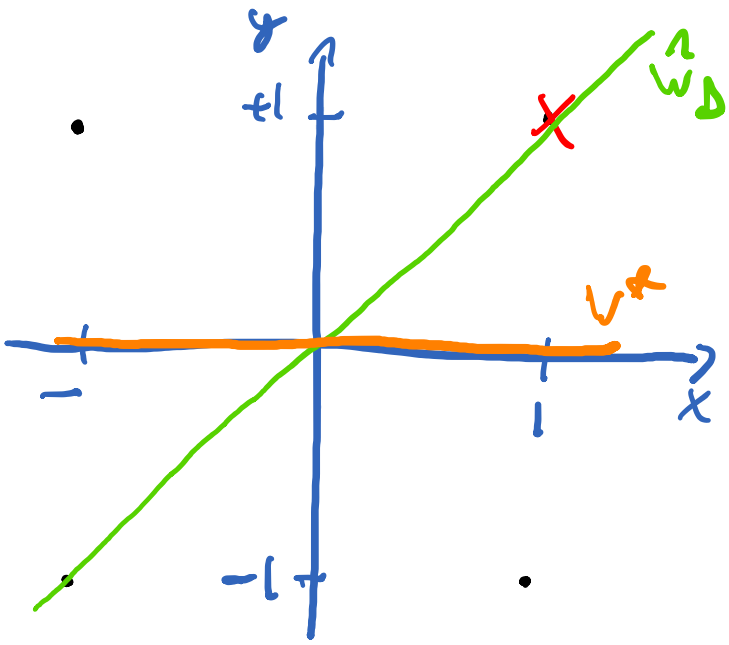
- Law of large numbers / uniform convergence are asymptotic statements (with $n \to \infty$)
- In practice one has finite amount of data.

- What can go wrong?

Simple example

$$\hat{\mathbf{w}}_D = \operatorname*{argmin}_{\mathbf{w}} \hat{R}_D(\mathbf{w})$$

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} R(\mathbf{w})$$



$\mathbb{E}_{\substack{D\sim P \\ |D|=1}}[R(\hat{w}_D)] = 2$

Fit $f(x) = w \cdot x$

$P = \text{Uniform}(\{(-1,-1),(-1,+1),(+1,-1),(+1,+1)\})$

$D = \{(x_1, y_1)\}$ $(x_1, y_1) \sim P$, e.g. $x_1 = y_1 = 1$

$\hookrightarrow \hat{w}_D = \operatorname{argmin}_w (y_1 - w \cdot x_1)^2 = 1$

$R(\hat{w}_D) = \mathbb{E}_{(x,y)\sim P}(y - \hat{w}_D \cdot x)^2$

$= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 2^2 = 2$

$w^* = 0$

$R(w^*) = 1 \cdot 1^2 = 1$

$\hat{R}_D(w^*) = 1$

$\hat{R}(\hat{w}_D) = 0$

$\mathbb{E}_D[\hat{R}(\hat{w}_D)] = 0$

What if we evaluate performance on training data?

$$\hat{\mathbf{w}}_D = \operatorname*{argmin}_{\mathbf{w}} \hat{R}_D(\mathbf{w}) \qquad \mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} R(\mathbf{w})$$

- In general, it holds that $\mathbb{E}_D\left[\hat{R}_D(\hat{\mathbf{w}}_D)\right] \leqslant \mathbb{E}_D\left[R(\hat{\mathbf{w}}_D)\right]$

$$\mathbb{E}_D\left[\hat{R}_D(\hat{w}_D)\right] \overset{\text{ERM}}{=} \mathbb{E}_D\left[\min_w \hat{R}_D(w)\right] \overset{\text{Jensen's}}{\leq} \min_w \mathbb{E}_D\left[\hat{R}_D(w)\right]$$

$$\overset{\text{Def.}}{=} \min_w \mathbb{E}_D\left[\frac{1}{|D|}\sum_{i=1}^{|D|}(y_i - w^T x_i)^2\right] \overset{\text{lin. exp.}}{=} \min_w \frac{1}{|D|}\sum_{i=1}^{|D|} \underbrace{\mathbb{E}_{(x_i,y_i)\sim p}(y_i - w^T x_i)^2}_{R(w)}$$

$$= \min_w R(w) \leq \mathbb{E}_D\left[R(\hat{w}_D)\right] \qquad \square$$

- **Thus, we obtain an overly optimistic estimate!**

More realistic evaluation?

- Want to avoid underestimating the prediction error
- **Idea**: Use separate test set from the same distribution P
- Obtain training and test data D_{train} and D_{test}
- Optimize ***w*** on training set

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \hat{R}_{train}(\mathbf{w})$$

- Evaluate on test set

$$\hat{R}_{test}(\hat{\mathbf{w}}) = \frac{1}{|D_{test}|} \sum_{(\mathbf{x},y)\in D_{test}} (y - \hat{\mathbf{w}}^T\mathbf{x})^2$$

- Then: $\mathbb{E}\left[\hat{R}_{test}(\hat{\mathbf{w}})\right] = \mathbb{E}\left[R(\hat{\mathbf{w}})\right]$

Why?

$D = D_{train}$, $V = D_{test}$, $D, V \sim P$

$$\mathbb{E}_{D,V}\left[\hat{R}_V(\hat{w}_D)\right] \overset{\text{ind.}}{=} \mathbb{E}_D\left[\mathbb{E}_V\left[\hat{R}_V(\hat{w}_D)\right]\right] = \mathbb{E}_D\left[\mathbb{E}_V\left[\frac{1}{|V|}\sum_{i=1}^{|V|}(y_i - \hat{w}_D^T x_i)^2\right]\right]$$

$$= \mathbb{E}_D\left[\frac{1}{|V|}\sum_{i=1}^{|V|}\underbrace{\mathbb{E}\left[(y_i - \hat{w}_D^T x_i)^2\right]}_{R(\hat{w}_D),\ \text{since } (y_i, x_i) \perp D}\right] = \mathbb{E}_D\left[R(\hat{w}_D)\right]$$

□

First attempt: Evaluation for model selection

- Obtain training and test data D_{train} and D_{test}
- Fit each candidate model (e.g., degree *m* of polynomial)

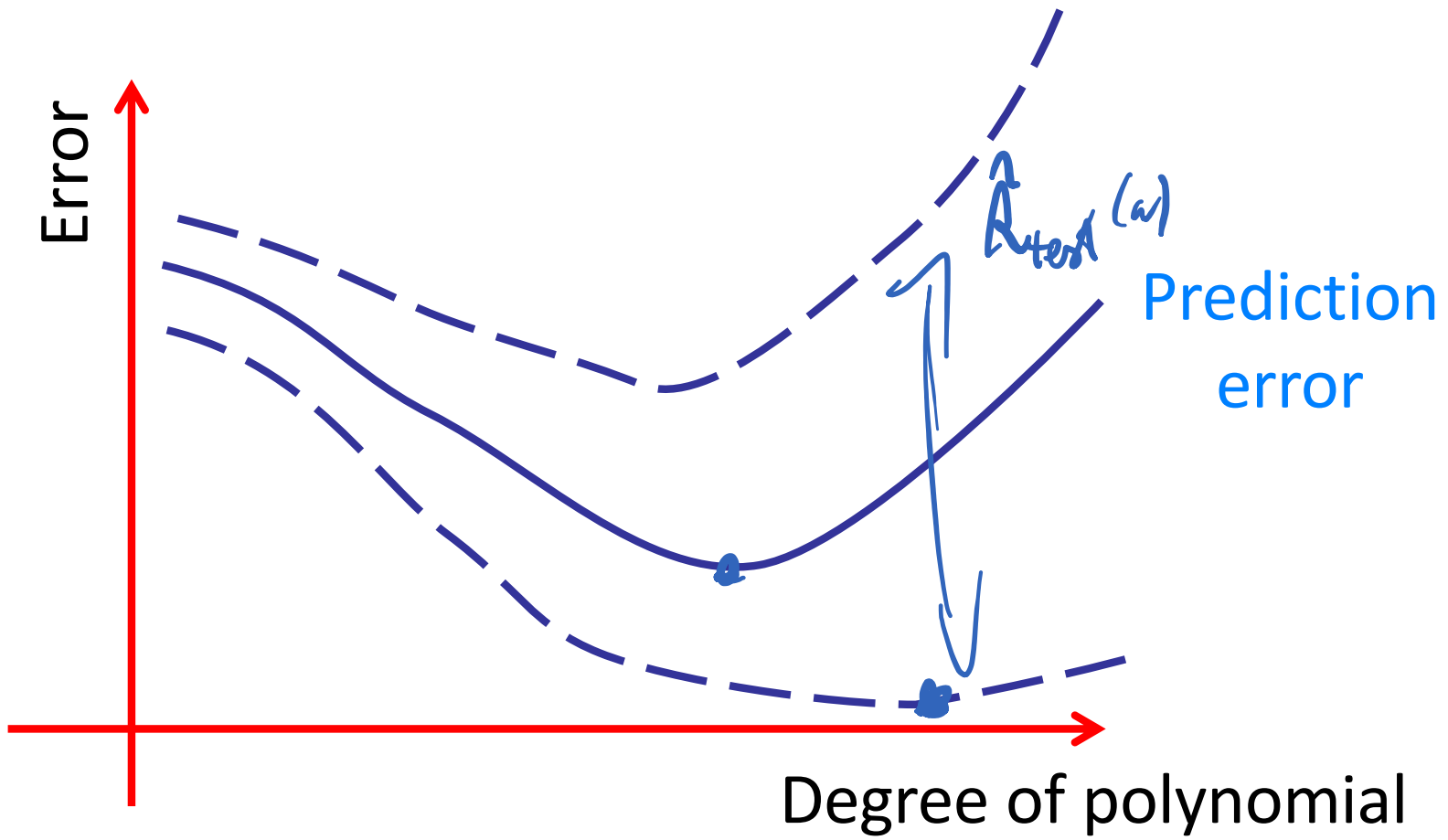
$$\hat{\mathbf{w}}_m = \underset{\mathbf{w}:\text{degree}(\mathbf{w})\leq m}{\text{argmin}} \hat{R}_{\text{train}}(\mathbf{w})$$

- Pick one that does best on test set:

$$\hat{m} = \underset{m}{\text{argmin}}\, \hat{R}_{\text{test}}(\hat{\mathbf{w}}_m)$$

- *Do you see a problem?*

Overfitting to *test* set



- Test error is itself random! Variance usually increases for more complex models
- Optimizing for *single* test set creates bias