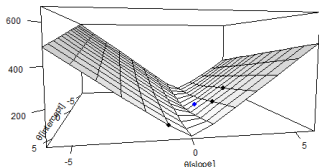


# Introduction to Machine Learning

## Introduction: Losses & Risk Minimization



### Learning goals

- Know the concept of loss
- Understand the relationship between loss and risk
- Understand the relationship between risk minimization and finding the best model

# HOW TO EVALUATE MODELS

- In the training, we want to optimize  $\theta$ . To score  $\theta$ , we have to compare the actual output with the predicted output:

Features $x$		Target $y$	?
--------------	--	------------	---

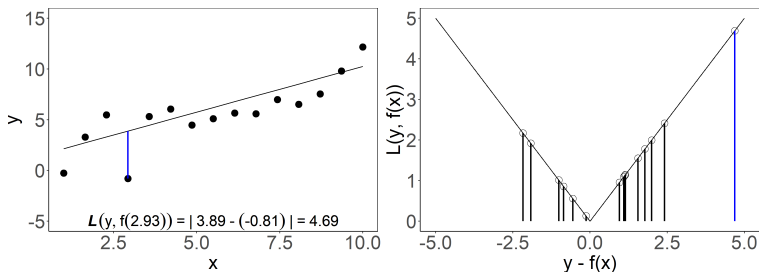
- We need to define a suitable criterion, e.g.:
  - Absolute error  $|2588 - 2220| = 368$
  - Squared error:  $(2588 - 2220)^2 = 135,424$
- The choice of this metric has a major influence on the final model, as it determines what a *good* model is.
- It will determine the ranking of the different models  $f \in \mathcal{H}$ .
- The metric we use is called the **loss function**.

# LOSS

The **loss function**  $L(y, f(\mathbf{x}))$  quantifies the "quality" of the prediction  $f(\mathbf{x})$  of a single observation  $\mathbf{x}$ :

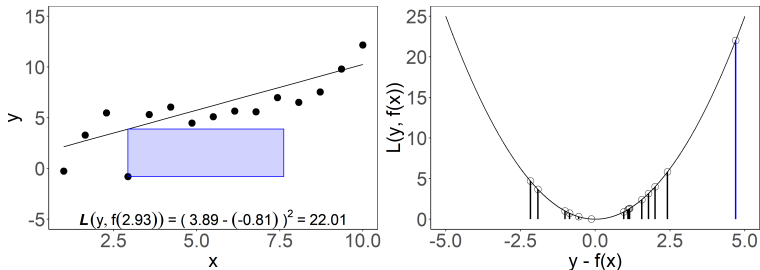
$$L : \mathcal{Y} \times \mathbb{R}^g \rightarrow \mathbb{R}.$$

How "close"  $f(\mathbf{x})$  is to  $y$  can be quantified e. g. by the absolute loss  $L(y, f(\mathbf{x})) = |f(\mathbf{x}) - y|$ .



# LOSS

Often, we use the L2-loss  $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ :



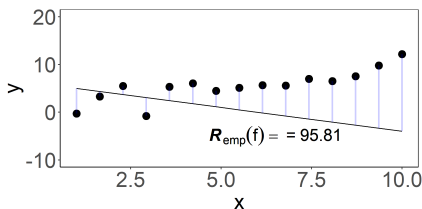
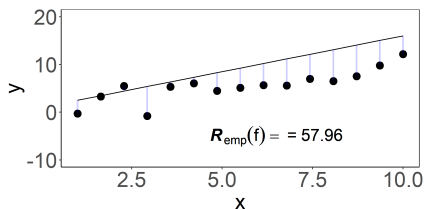
# RISK

The **risk function** quantifies the "quality" of the whole model.

The ability of a model  $f$  to reproduce the association between  $\mathbf{x}$  and  $y$  that is present in the data  $\mathcal{D}$  can be measured by the **summed loss**, also called "**empirical risk**":

$$\mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^n L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right).$$

$$\mathcal{R} : \mathcal{H} \rightarrow \mathbb{R}.$$



# RISK

## Notes:

- The risk is often denoted as empirical mean over  $L(y, f(\mathbf{x}))$

$$\bar{\mathcal{R}}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^n L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right).$$

The factor  $\frac{1}{n}$  does not make a difference in optimization, so we will consider  $\mathcal{R}_{\text{emp}}(f)$  most of the time.

- Since the model  $f$  is usually defined by **parameters**  $\theta$  in a parameter space  $\Theta$ , this becomes:

$$\mathcal{R} : \mathbb{R}^d \rightarrow \mathbb{R}.$$

$$\begin{aligned}\mathcal{R}_{\text{emp}}(\theta) &= \frac{1}{n} \sum_{i=1}^n L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \theta\right)\right) \\ \hat{\theta} &= \arg \min_{\theta \in \Theta} \mathcal{R}_{\text{emp}}(\theta)\end{aligned}$$

# RISK MINIMIZATION

The best model is the model with the smallest risk.

If we have a finite number of models  $f$ , we can compare the risk  $\mathcal{R}_{\text{emp}}(\theta)$  of all models:

Model	$\theta_{\text{intercept}}$	$\theta_{\text{slope}}$	$\mathcal{R}_{\text{emp}}(\theta)$
$f_1$	2	3	194.62
$f_2$	3	2	127.12
$f_3$	6	-1	95.81
$f_4$	1	1.5	57.96

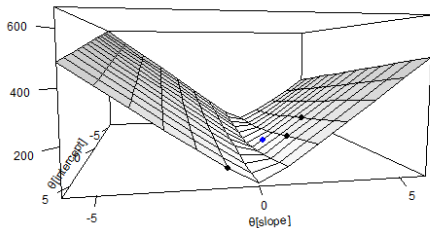
# RISK MINIMIZATION

But: Normally, the hypothesis space  $\mathcal{H}$  is infinitely large.

As the the mapping of the hypothesis space to its parameters is bijective, we can consider the error surface depending on the parameters  $\theta$ :

$$\mathcal{R} : \mathbb{R}^d \rightarrow \mathbb{R}.$$

Model	$\theta_{intercept}$	$\theta_{slope}$	$\mathcal{R}_{emp}(\theta)$
$f_1$	2	3	194.62
$f_2$	3	2	127.12
$f_3$	6	-1	95.81
$f_4$	1	1.5	57.96





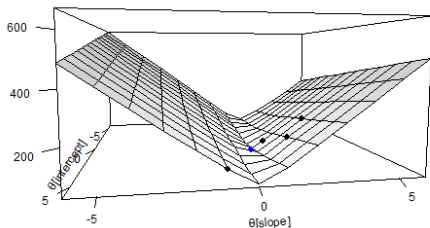
# RISK MINIMIZATION

The process of finding the best model is called **empirical risk minimization** (ERM).

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \mathcal{R}_{\text{emp}}(\theta).$$

$$\mathcal{R} : \mathbb{R}^d \rightarrow \mathbb{R}.$$

Model	$\theta_{\text{intercept}}$	$\theta_{\text{slope}}$	$\mathcal{R}_{\text{emp}}(\theta)$
$f_1$	2	3	194.62
$f_2$	3	2	127.12
$f_3$	6	-1	95.81
$f_4$	1	1.5	57.96
$f_5$	1.25	0.90	23.40



# RISK MINIMIZATION

Most learners in ML try to solve the above *optimization problem*, which implies a tight connection between ML and optimization.

# FURTHER REMARKS

- For regression tasks, the loss often only depends on the residual  $L(y, f(\mathbf{x})) = L(y - f(\mathbf{x})) = L(\epsilon)$ .
- The choice of loss implies which kinds of errors are important or not – requires *domain knowledge*!
- For learners that correspond to probabilistic models, the loss determines / is equivalent to distributional assumptions.
- Since learning can be re-phrased as minimizing the loss, the choice of loss strongly affects the computational difficulty of learning:
  - How smooth is  $\mathcal{R}_{\text{emp}}(\theta)$  in  $\theta$ ?
  - Is  $\mathcal{R}_{\text{emp}}(\theta)$  differentiable so that we can use gradient-based methods?
  - Does  $\mathcal{R}_{\text{emp}}(\theta)$  have multiple local minima or saddlepoints over  $\Theta$ ?