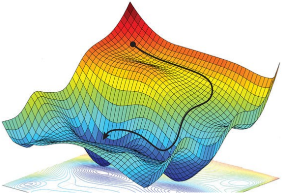
# Introduction to Machine Learning

# Pseudo-residuals and Gradient Descent



**Learning goals**

- Learn what pseudo-residuals are
- Understand the relationship between pseudo-residuals and gradient descent

# PSEUDO-RESIDUALS

- We further define **pseudo-residuals** as the negative first derivatives of loss functions w.r.t. $f(\mathbf{x})$

$$\tilde{r} \quad := \quad -\frac{\partial L(y, f(\mathbf{x}))}{\partial f(\mathbf{x})}.$$

- We will gain more intuition about the principle of pseudo-residuals in a later chapter.

# GD IN ML AND PSEUDO-RESIDUALS

By using the chain rule we see that

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \underbrace{\left.\frac{\partial L\left(y^{(i)}, f\right)}{\partial f}\right|_{f=f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)}}_{=-\tilde{r}^{(i)}} \cdot \nabla_{\boldsymbol{\theta}} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) \\
&= -\sum_{i=1}^{n} \tilde{r}^{(i)} \cdot \nabla_{\boldsymbol{\theta}} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)
\end{aligned}$$

For risk minimization, the update rule for the parameter $\boldsymbol{\theta}$ is

$$\begin{aligned}
\boldsymbol{\theta}^{[t+1]} &\leftarrow \boldsymbol{\theta}^{[t]} - \alpha^{[t]} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \left. L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}} \\
\boldsymbol{\theta}^{[t+1]} &\leftarrow \boldsymbol{\theta}^{[t]} + \alpha^{[t]} \sum_{i=1}^{n} \tilde{r}^{(i)} \cdot \left.\nabla_{\boldsymbol{\theta}} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}}
\end{aligned}$$

$\alpha^{[t]} \in [0, 1]$ is called “learning rate” in this context.