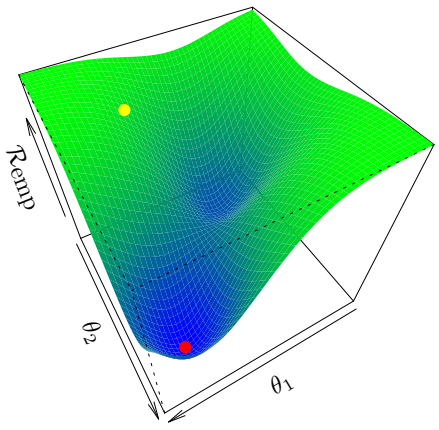
Introduction to Machine Learning

ML-Basics: Optimization

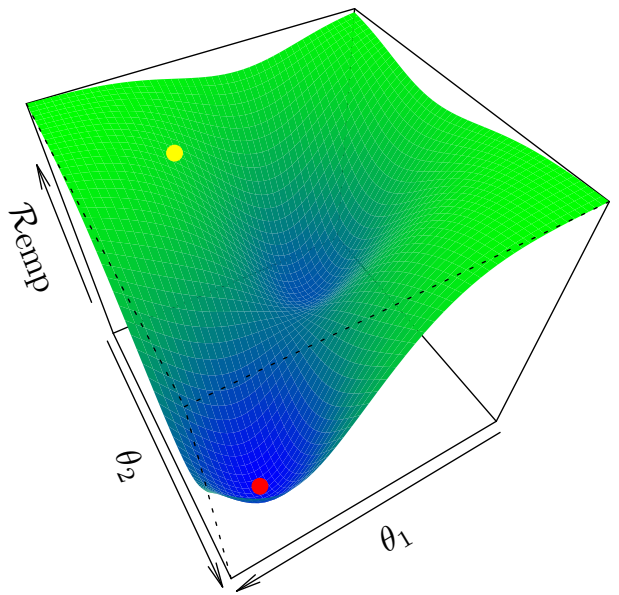


Learning goals

- Understand how the risk function is optimized to learn the optimal parameters of a model
- Understand the idea of gradient descent as a basic risk optimizer

INTRODUCTION

- We have seen, we can 1-on-1 identify models f with their parametrization $\boldsymbol{\theta} \in \Theta$.
- Hence we can express the associated empirical risk of the model as a function of these parameters: $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$.
- Therefore, when we try to find the best model, we actually traverse on the error surface from a starting point (yellow) with the goal of finding the point with the lowest empirical risk (red).



INTRODUCTION

Formally, this means that we find the best model $\hat{f}$ parametrized by parameters $\hat{\boldsymbol{\theta}} \in \Theta$ regarding an empirical risk $\mathcal{R}_{\text{emp}}$ by **minimizing** $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, i.e.,

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

For such a **(global) minimum** $\hat{\boldsymbol{\theta}}$ it obviously holds that

$$\forall \boldsymbol{\theta} \in \Theta : \quad \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) \leq \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}).$$

However, this does not imply that $\hat{\boldsymbol{\theta}}$ is unique by any means.

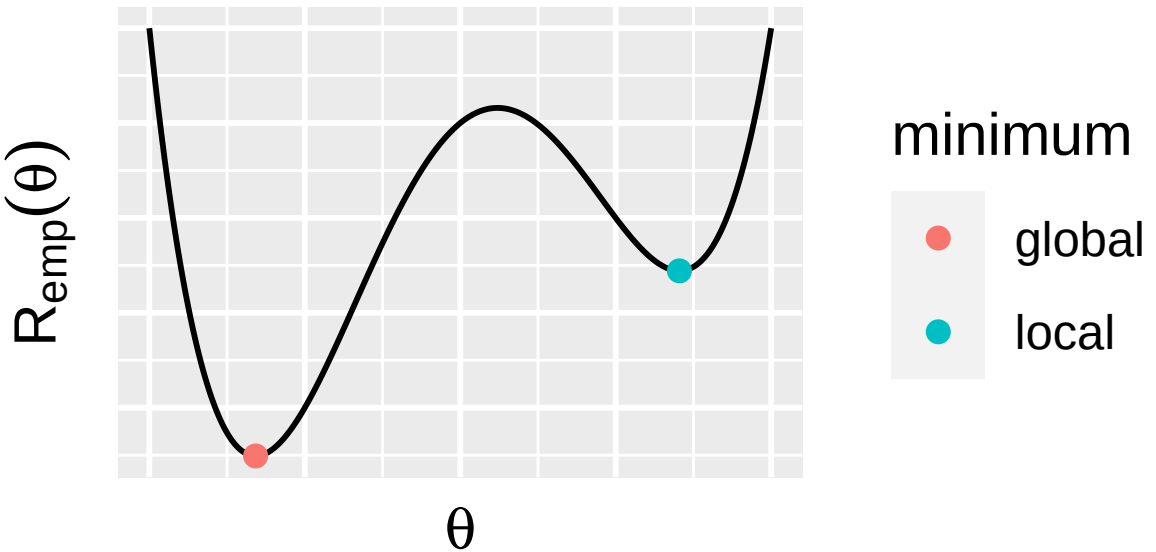
Which kind of technique we can use to solve the minimization problem strongly depends on the feature space. In this chapter we will focus on purely numeric features.

CONTINUOUS $\mathcal{R}_{\text{emp}}$

If the empirical risk $\mathcal{R}_{\text{emp}}$ is continuous in $\boldsymbol{\theta}$ we can define a **local minimum** $\hat{\boldsymbol{\theta}}$, such that

$$\exists \epsilon > 0 \; \forall \boldsymbol{\theta} \in \left\{ \bar{\boldsymbol{\theta}} \in \Theta \mid \left\| \hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}} \right\| < \epsilon \right\} : \quad \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) \leq \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}).$$

Clearly every global minimum is also a local minimum (if it exists). In general finding a local minimum is easier than finding a global minimum.

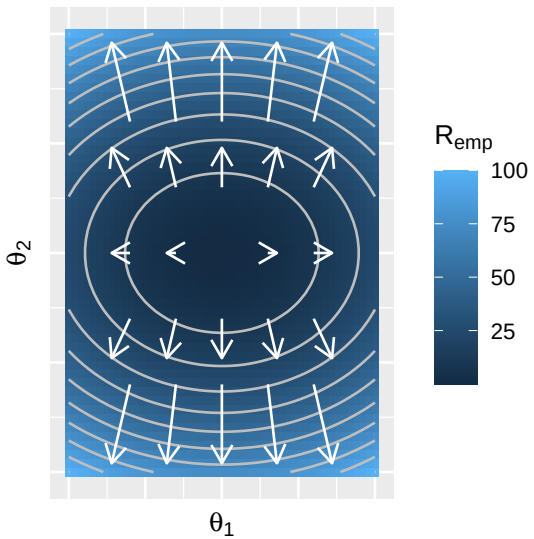


CONTINUOUSLY DIFFERENTIABLE $\mathcal{R}_{\text{emp}}$

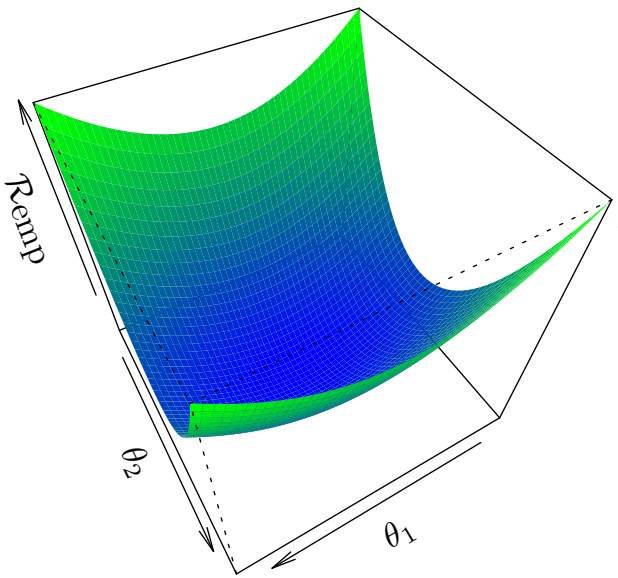
If the empirical risk $\mathcal{R}_{\text{emp}}$ is continuously differentiable in $\boldsymbol{\theta}$ then a **sufficient condition** for $\hat{\boldsymbol{\theta}}$ to be a local minimum is that the gradient

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) = 0$$

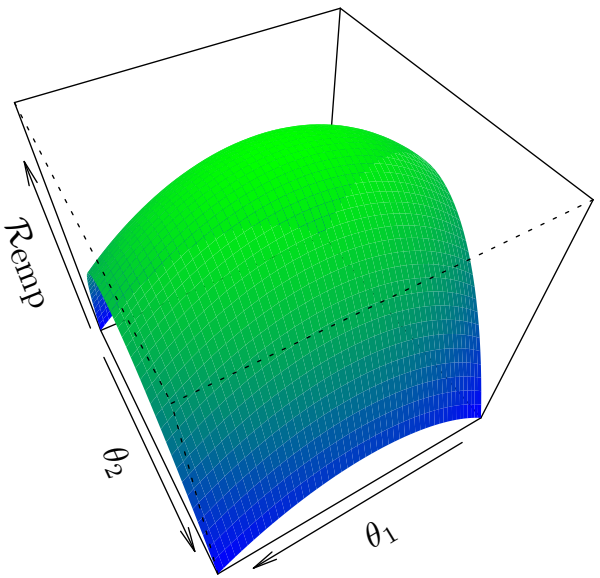
and the Hessian $\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}})$ is positive definite. Which makes sense, since, while the gradient can be thought of as the local direction and rate of fastest increase, the Hessian measures the local curvature of $\mathcal{R}_{\text{emp}}$.



$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$



const. pos. def. Hessian



const. neg. def. Hessian

LEAST SQUARES ESTIMATOR

Now, for given features $\mathbf{X} \in \mathbb{R}^{n \times p}$ and target $\mathbf{y} \in \mathbb{R}^n$, we want to find the best linear model regarding the squared error loss, i.e.,

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)}) \ .$$

With the sufficient condition for continously differentiable functions it can be shown that the **least squares estimator**

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

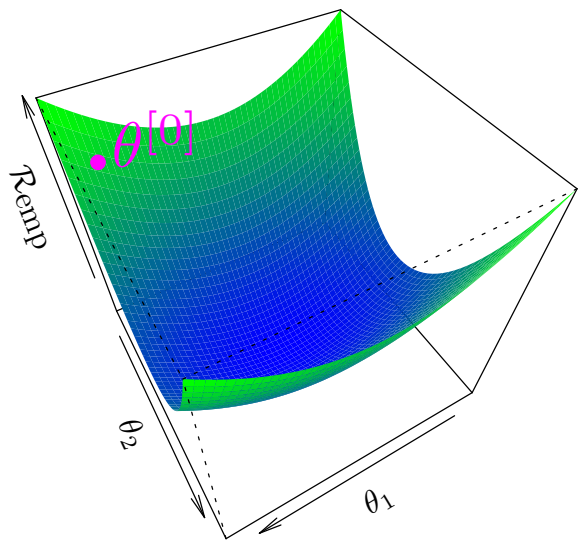
is a local minimum of $\mathcal{R}_{\text{emp}}$. Since, here, $\mathcal{R}_{\text{emp}}$ is a convex function it follows that there is only one minimum. Hence $\hat{\boldsymbol{\theta}}$ is the global minimum.

Note: Often such an analytical solution to our respective minimization problem does not exist. Therefore we need numerical methods which enable us to find an approximate solution.

GRADIENT DESCENT

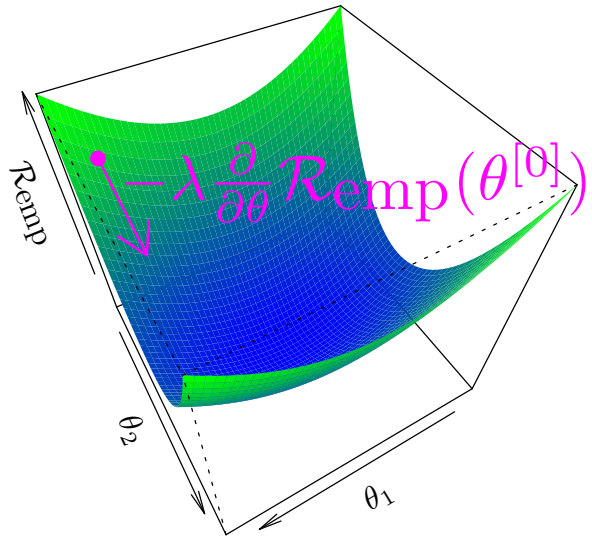
The simple idea of **gradient descent** (GD) is to follow iteratively from the t-th solution candidate $\boldsymbol{\theta}^{[t]}$ in the direction of the negative gradient, i.e., the direction of the steepest descent, with a learning rate λ to the $[t+1]$-th solution candidate $\boldsymbol{\theta}^{[t+1]}$, s.t.

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - \lambda \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[t]}).$$

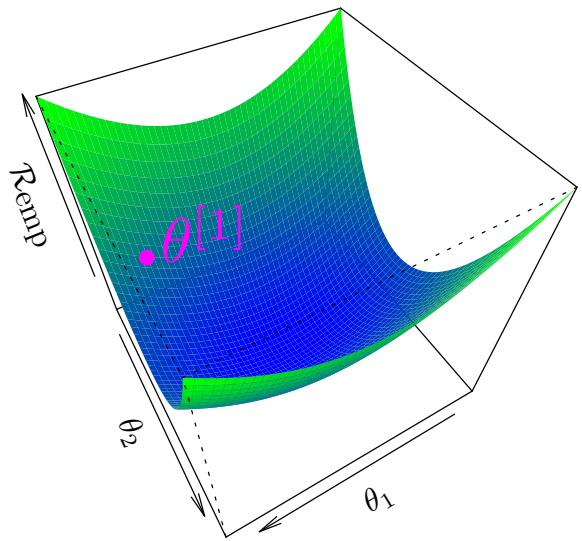


So we choose a random starting point $\boldsymbol{\theta}^{[0]}$ with associated empirical risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[0]}) = 76.25$.

GRADIENT DESCENT

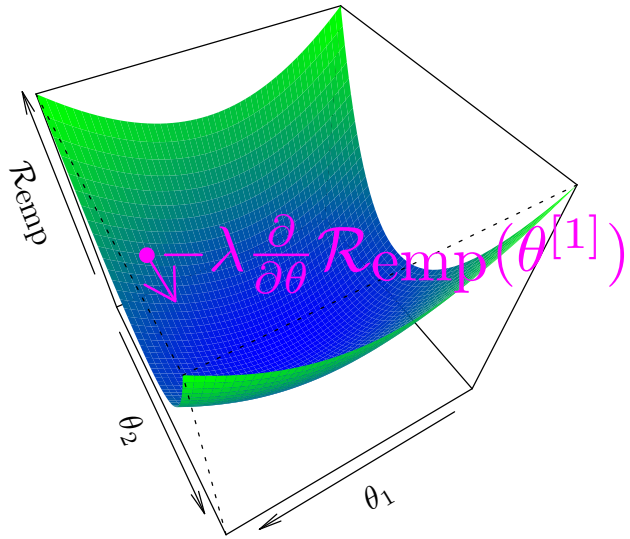


Now we follow in the direction of the negative gradient at $\boldsymbol{\theta}^{[0]}$.

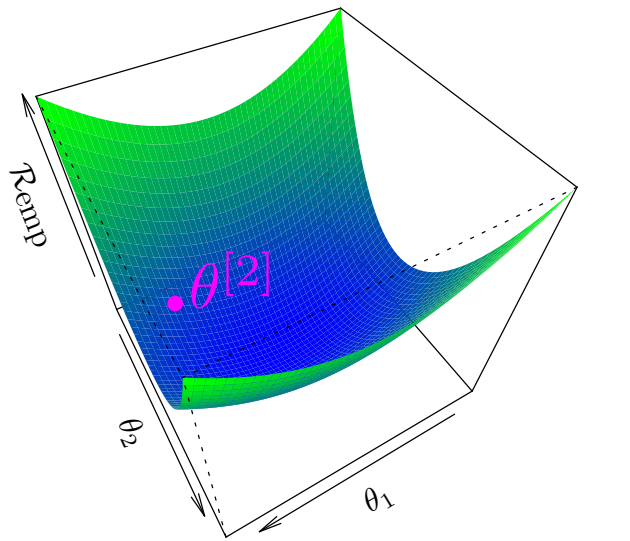


We arrive at $\boldsymbol{\theta}^{[1]}$ with associated empirical risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[1]}) \approx 42.73$.
Hence our estimation has improved since $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[1]}) < \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[0]})$.

GRADIENT DESCENT

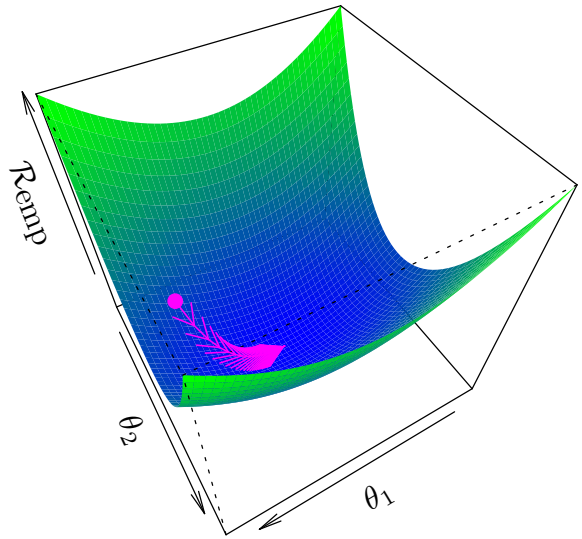


Again we follow in the direction of the negative gradient, but now at $\boldsymbol{\theta}^{[1]}$.

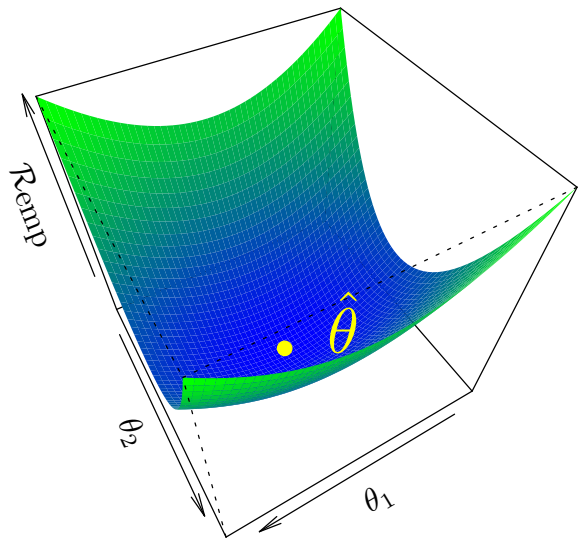


We arrive at $\boldsymbol{\theta}^{[2]}$ with associated empirical risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[2]}) \approx 25.08$. Thus clearly, we have improved our estimation again.

GRADIENT DESCENT



We iteratively improve our estimation by repeating this procedure until $\mathcal{R}_{\text{emp}}$ converges[1].



Eventually we arrive at a minimum $\hat{\theta}$ with $\mathcal{R}_{\text{emp}}(\hat{\theta}) = 0$.

[1] This is not the only stopping criterium; e.g., we could limit the number of iterations.

FURTHER TOPICS

- There exist many improvements of the GD method, e.g., we could also optimize the learning rate λ.
- GD is a so-called first-order method. Second-order methods use the Hessian (which must therefore exist) to refine the search direction.
- If the gradient of GD is not derived from the empirical risk of the whole data set, but instead from a randomly selected subset of it, we call the respective method **stochastic gradient descent** (SGD). For high-dimensional problems this can lead to a higher computational efficiency.
- Often it is desirable to not allow arbitrarily large $||\hat{\theta}||$, since this could result, among other things, in numerical instability of the method. This procedure is called **regularization**.

FURTHER TOPICS

learning vs optimuzation,.,.,.