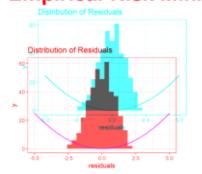
# Introduction to Machine Learning

Advanced Riski Minimization vs.

Maximum Likelihood Estimation vs.

Empirical Risk Minimization



#### Learning goals

- Learning goals
  - aUnderstand the connection between
  - maximum likelihood and risk
  - Learn the correspondence between a Gaussian error distribution and the L2 loss



#### MAXIMUM LIKELIHOOD

Let's consider regression from a maximum likelihood perspective. Assume:

$$y \mid \mathbf{x} \sim p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

Common case: true underlying relationship  $f_{true}$  with additive noise:

$$y = f_{\mathsf{true}}(\mathbf{x}) + \epsilon$$

ftrue(x) with Gaussian noise ~ N(0.1)

where  $f_{\text{true}}$  has params  $\theta$  and  $\epsilon$  a RV that follows some distribution  $\mathbb{P}_{\epsilon}$ , with  $\mathbb{E}[\epsilon] = 0$ . Also, assume  $\epsilon \perp \!\!\! \perp \mathbf{x}$ .

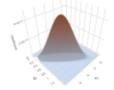


## MAXIMUM LIKELIHOOD /2

From a statistics / maximum-likelihood perspective, we assume (or we pretend) we know the underlying distribution  $p(y \mid \mathbf{x}, \theta)$ .

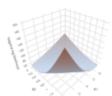
• Then, given i.i.d data  $\mathcal{D} = ((\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)}))$  from  $P_{xy}$  the maximum-likelihood principle is to maximize the **likelihood** 

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \rho\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$



or equivalently to minimize the negative log-likelihood

$$-\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$



### MAXIMUM LIKELIHOOD /3

From an ML perspective we assume our hypothesis space corresponds to the space of the (parameterized)  $f_{true}$ .

Simply define neg. log-likelihood as loss function

$$L(y, f(\mathbf{x} \mid \theta)) := -\log p(y \mid \mathbf{x}, \theta)$$

Then, maximum-likelihood = ERM

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

- NB: When we are only interested in the minimizer, we can ignore multiplicative or additive constants.



## **GAUSSIAN ERRORS - L2-LOSS**

Assume  $y = f_{\text{true}}(\mathbf{x}) + \epsilon$  with additive Gaussian errors, i.e.  $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ . Then

$$y \mid \mathbf{x} \sim N\left(f_{\mathsf{true}}(\mathbf{x}), \sigma^2\right)$$

The likelihood is then

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(y^{(i)} \mid f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma^{2}\right)$$

$$\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^{2}}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^{2}\right)$$



# GAUSSIAN ERRORS - L2-LOSS /2

Easy to see: minimizing neg. negative log-likelihood with Gaussian errors is the same as ERM with L2-loss:

$$-\ell(\boldsymbol{\theta}) = -\log\left(\mathcal{L}(\boldsymbol{\theta})\right)$$

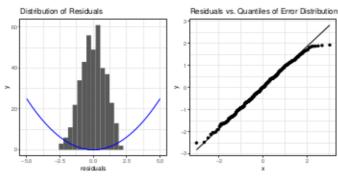
$$= -\log\left(\prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^{2}}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^{2}\right)\right)$$

$$\propto \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^{2}$$



# GAUSSIAN ERRORS - L2-LOSS /3

- We simulate data y | x ~ N (f<sub>true</sub>(x), 1) with f<sub>true</sub> = 0.2 ⋅ x
- Let's plot empirical errors as histogram, after fitting our model with L2-loss
- Q-Q-plot compares empirical residuals vs. theoretical quantiles of Gaussian





#### DISTRIBUTIONS AND LOSSES

ullet For every error distribution  $\mathbb{P}_{\epsilon}$  we can derive an equivalent loss function, which leads to the same point estimator for the parameter vector  $oldsymbol{ heta}$  as maximum-likelihood. Formally,

• 
$$\hat{\theta} \in \arg \max_{\theta} \mathcal{L}(\theta) \implies \hat{\theta} \in \arg \min_{\theta} - \log(\mathcal{L}(\theta))$$

 But: The other way around does not always work: We cannot derive a corresponding pdf or error distribution for every loss function – the Hinge loss is one prominent example, for which some probabilistic interpretation is still possible however, see



## DISTRIBUTIONS AND LOSSES /2

When does the reverse direction hold?

- If we can write the loss as  $L(y, f(\mathbf{x})) = L(y f(\mathbf{x})) = L(r)$  for  $r \in \mathbb{R}$ , then minimizing  $L(y f(\mathbf{x}))$  is equivalent to maximizing a conditional log-likelihood  $\log(p(y f(\mathbf{x}|\theta)))$  if
  - log(p(r)) is affine trafo of L (undoing the ∞):

$$\log(p(r)) = a - bL(r), \ a \in \mathbb{R}, b > 0$$

p is a pdf (non-negative and integrates to one)

Thus, a loss L corresponds to MLE under *some* distribution if there exist  $a \in \mathbb{R}, \ b > 0$  such that

$$\int_{\mathbb{D}} \exp(a - bL(r)) dr = 1$$

