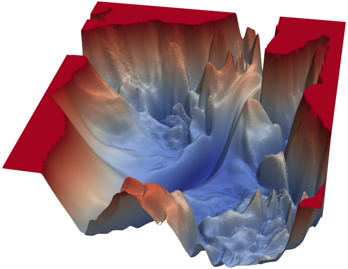
Introduction to Machine Learning

Properties of Loss Functions



Learning goals

- Know the concept of robustness
- Learn about analytical and computational properties of loss functions
- Understand that the loss function may influence convergence of the optimizer

THE ROLE OF LOSS FUNCTIONS

Why should we care about how to choose the loss function $L\left(y, f(\mathbf{x})\right)$?

- **Statistical** properties: Choice of loss implies statistical assumptions on the distribution of $y \mid \mathbf{x} = \mathbf{x}$ (see *Maximum Likelihood Estimation vs. Empirical Risk Minimization*).
- **Robustness** properties: Some loss functions are more robust towards outliers than others.
- **Analytical** properties: The computational / optimization complexity of the problem.

  $$\underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\, \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

  is influenced by the choice of the loss function.

BASIC TYPES OF REGRESSION LOSSES

- Regression losses usually only depend on the **residuals**

$$r \quad := \quad y - f(\mathbf{x})$$

- A loss is called **distance-based** if
  - it can be written in terms of the residual

$$L(y, f(\mathbf{x})) = \psi(r) \text{ for some } \psi : \mathbb{R} \to \mathbb{R}$$

  - $\psi(r) = 0 \Leftrightarrow r = 0$.
- A loss is **translation-invariant**, if $L(y + a, f(\mathbf{x}) + a) = L(y, f(\mathbf{x}))$.
- Losses are called **symmetric** if $L(y, f(\mathbf{x})) = L(f(\mathbf{x}), y)$.

@BB: This slide is a bit without context imo. We are also not saying anything about classification losses - should we?

ROBUSTNESS

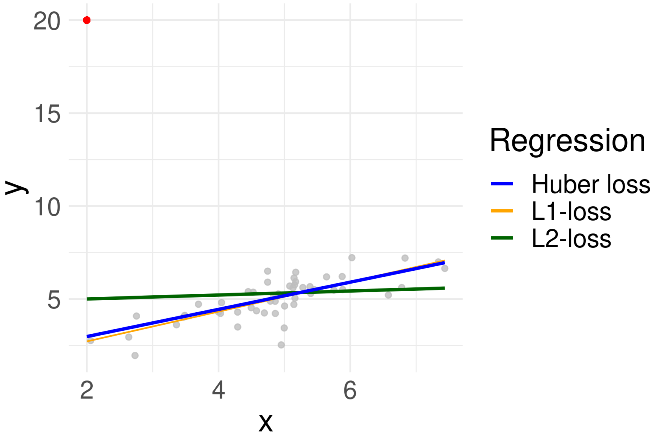
Outliers (in y) have large residuals $r = y - f(\mathbf{x})$. For some losses large residuals have a much bigger impact on the risk much more than for other losses.

y	$\hat{f}(\mathbf{x})$	L1	L2	Huber ($\epsilon = 5$)
1	0	1	1	0.5
5	0	5	25	12.5
10	0	10	100	37.5
50	0	50	2500	237.5

As a consequence, a model is less influenced by outliers than by inliers if the loss is robust.

ROBUSTNESS

The L2 loss is an example for a loss function that is not very robust towards outliers. It penalizes large residuals more than the L1 or the Huber loss. The L1 and the Huber loss are thus regarded robust.

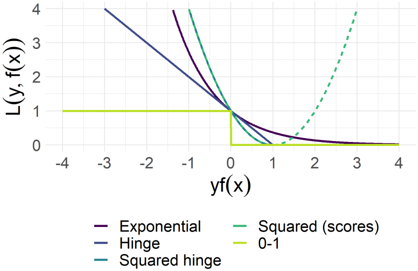


ANALYTICAL PROPERTIES: SMOOTHNESS

- **Smoothness** of a function is a property that is measured by the number of continuous derivatives it has.
- A function is said to be $\mathcal{C}^k$ if it is k times continuously differentiable. A function is $\mathcal{C}^\infty$ if it is continuously differently for all orders k.
- In contrast do derivative-free methods, derivative-based methods require a certain level of smoothness of the risk function $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$.
- Example: Gradient descent requires differentiability of the $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ (existence of $\nabla\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$), Newton-Raphson requires $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ to be twice differentiable (existence of Hessian $\nabla^2\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$).

ANALYTICAL PROPERTIES: SMOOTHNESS

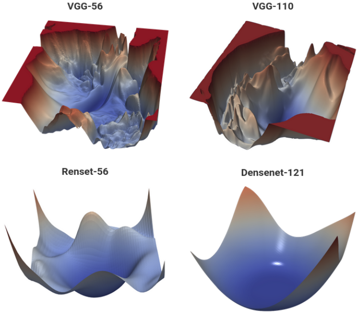
- If the loss function is not smooth, the risk minimization problem is in general not smooth either.
- Instead, derivative-free optimization need to be used (which might not be desirable)



Squared loss, exponential loss, and squared hinge loss are continuously differentiable. The hinge loss is continuous but not differentiable. The 0-1-loss is not even continuous.

ANALYTICAL PROPERTIES: CONVEXITY

- In optimization, convex optimization problems are desirable because they have a number of conventient properties. In particular, it holds for convex problems: A local minimum of a convex function is also a global minimum. A strictly convex function has at most **one** global minimum (uniqueness).



ANALYTICAL PROPERTIES: CONVEXITY

Li et al., 2018, Visualizing the Loss Landscape of Neural Nets. The problem on the bottom right is convex, the others are not.

- In practical terms complexity means that we do not need to worry to get stuck in a local minimum during risk minimization.
- Note that convexity of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ does not only depend on convexity of the loss function: Convexity of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is also determined by the choice of the hypothesis space!
- For example, if $L\left(y, f(\mathbf{x})\right)$ is convex in its second argument, and $f(\mathbf{x} \mid \boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, then $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is convex. If L is not convex, $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ might have multiple local minima (bad!).

ANALYTICAL PROPERTIES: CONVERGENCE

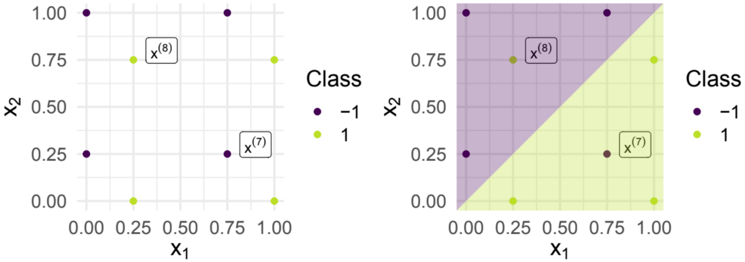
@BB: Do we want to cover this case here in such a detailed manner?

The choice of the loss function may also imply convergence behavior of the optimization problem.

Example: Gradient descent will not converge if we minimize the Bernoulli loss for linearly separable data.

First, we take a look at logistic regression for an almost linearly separable dataset consisting of the observations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(8)}$.

ANALYTICAL PROPERTIES: CONVERGENCE



Note: WLOG we estimate the model without intercept, s.t. we can visualize the regression coefficient $\boldsymbol{\theta}$ in 2D. Also, the symmetry of the data does not influence the generality of our conclusions.

Because of the symmetry of the data, the direction[1] of $\boldsymbol{\theta}$ is $\tilde{\boldsymbol{\theta}} := (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^\top$.

ANALYTICAL PROPERTIES: CONVERGENCE

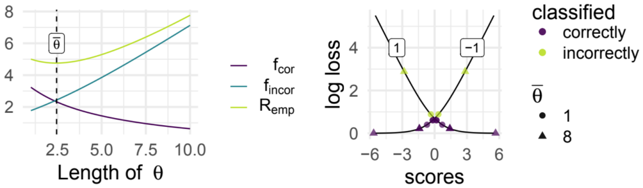
To find $\overline{\theta} := ||\boldsymbol{\theta}||_2$, we consider the empirical risk $\mathcal{R}_{\text{emp}}$ along $\tilde{\boldsymbol{\theta}}$:

$$\begin{aligned}\mathcal{R}_{\text{emp}} &= \sum_{i=1}^{8} \log\left[1 + \exp\left(-y^{(i)}\boldsymbol{\theta}^\top \mathbf{x}^{(i)}\right)\right] \\ &= \underbrace{\sum_{i=1}^{6} \log\left[1 + \exp\left(-\overline{\theta}\left|\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}\right|\right)\right]}_{=:\ f_{\text{cor}}(\overline{\theta})\ \text{(correctly classified)}} + \underbrace{\sum_{i=7}^{8} \log\left[1 + \exp\left(\overline{\theta}\left|\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}\right|\right)\right]}_{=:\ f_{\text{incor}}(\overline{\theta})\ \text{(incorrectly classified)}}.\end{aligned}$$

[1] $\boldsymbol{\theta}$ is perpendicular to the decision boundary and points to the "1"-space.

ANALYTICAL PROPERTIES: CONVERGENCE

Clearly, f_{cor} / f_{incor} are monotonically decreasing/increasing with rising length of $\boldsymbol{\theta}$:



- By removing obs. 7 and 8, we get a linearly separable dataset.
- This also means that we lose our "counterweight", i.e., if a parameter vector $\boldsymbol{\theta}$ is able to classify the samples perfectly, the vector $2\boldsymbol{\theta}$ also classifies the samples perfectly, with decreased risk.
- Therefore, an iterative optimizer such as gradient descent will continually increase $\boldsymbol{\theta}$ and never halt (in theory).

ANALYTICAL PROPERTIES: CONVERGENCE

- In such cases, regularization can guarantee convergence (see chapter on regularization).