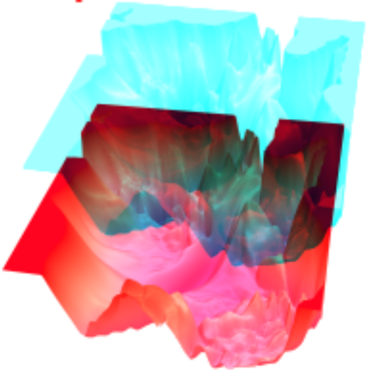
Introduction to Machine Learning

Advanced Risk Minimization

Properties of Loss Functions

Properties of Loss Functions



Learning goals

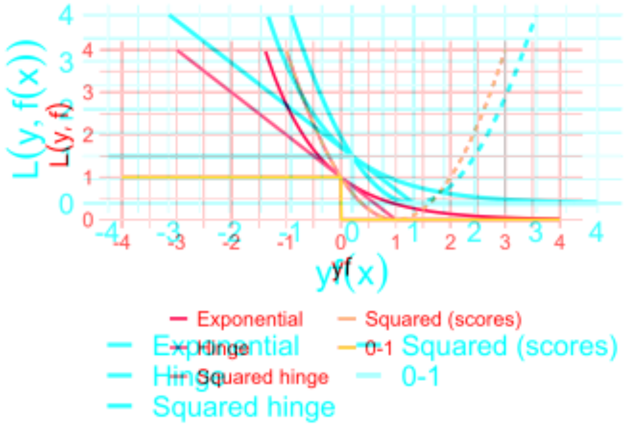
- Statistical properties
- Robustness
- Numerical properties
- Some fundamental terminology

Learning goals

- Statistical properties
- Robustness
- Numerical properties
- Some fundamental terminology

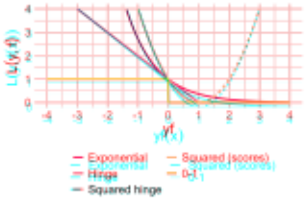
SOME BASIC TERMINOLOGY

Classification losses are usually expressed in terms of the **margin**:
$\nu := y \cdot f(\mathbf{x})$.



NUMERICAL PROPERTIES: SMOOTHNESS

- **Smoothness** of a function is a property measured by the number of continuous derivatives.
- Derivative-based optimization requires smoothness of the risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$
  - If loss is unsmooth, we might have to use derivative-free optimization (or worse, in case of 0-1)
  - Smoothness of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ not only depends on L, but also requires smoothness of $f(\mathbf{x})$!



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

NUMERICAL PROPERTIES: CONVEXITY

- A function $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is convex if

  $$\mathcal{R}_{\text{emp}}\left(t \cdot \boldsymbol{\theta} + (1-t) \cdot \tilde{\boldsymbol{\theta}}\right) \leq t \cdot \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + (1-t) \cdot \mathcal{R}_{\text{emp}}\left(\tilde{\boldsymbol{\theta}}\right)$$

  $\forall\, t \in [0, 1],\ \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$
  (strictly convex if the above holds with strict inequality).

- In optimization, convex problems have a number of convenient properties. E.g., all local minima are global.

  $\rightarrow$ strictly convex function has at most **one** global min (uniqueness).

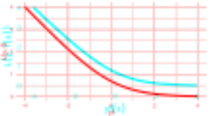
- For $\mathcal{R}_{\text{emp}} \in \mathcal{C}^2$, $\mathcal{R}_{\text{emp}}$ is convex iff Hessian $\nabla^2 \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is psd.

NUMERICAL PROPERTIES: CONVERGENCE

In case of **complete separation**, optimization might even fail entirely, e.g.:

- Margin-based loss that is strictly monotonicly decreasing in $y \cdot f$, e.g., **Bernoulli loss**:

$$L(y, f(\mathbf{x})) = \log(1 + \exp(-yf(\mathbf{x})))$$



- f linear in $\boldsymbol{\theta}$, e.g., **logistic regression** with $f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}$
- Data perfectly separable by our learner, so we can find $\boldsymbol{\theta}$:

$$y^{(i)} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = y^{(i)} \boldsymbol{\theta}^T \mathbf{x}^{(i)} > 0 \ \ \forall \mathbf{x}^{(i)}$$

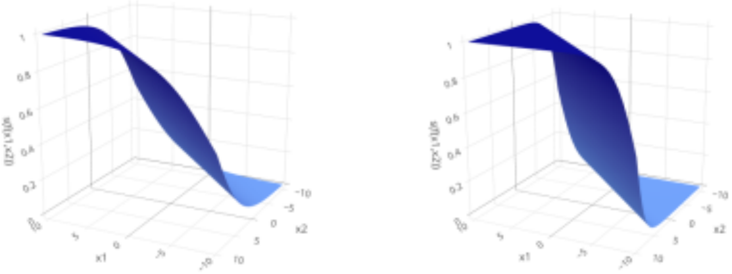
- Can now a construct a strictly better $\boldsymbol{\theta}$

$$\mathcal{R}_{\text{emp}}(2 \cdot \boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(2y^{(i)} \boldsymbol{\theta}^T \mathbf{x}^{(i)}\right) < \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

- As $||\boldsymbol{\theta}||$ increases, sum strictly decreases, as argument of L is strictly larger
- We can iterate that, so there is no local (or global) optimum, and no numerical procedure can converge

NUMERICAL PROPERTIES: CONVERGENCE / 2

- Geometrically, this translates to an ever steeper slope of the logistic/softmax function, i.e., increasingly sharp discrimination:



- In practice, data are seldomly linearly separable and misclassified examples act as counterweights to increasing parameter values.
- Besides, we can use **regularization** to encourage convergence to robust solutions.