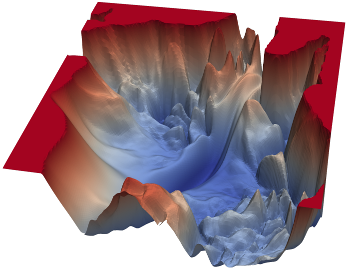
Introduction to Machine Learning

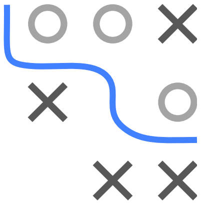
Advanced Risk Minimization
Properties of Loss Functions



Learning goals

- Statistical properties
- Robustness
- Optimization properties
- Some fundamental terminology

THE ROLE OF LOSS FUNCTIONS

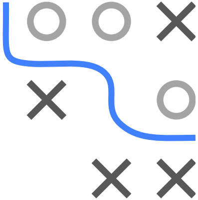


- Should be designed to measure errors appropriately
- **Statistical** properties: choice of loss implies statistical assumptions about the distribution of $y \mid \mathbf{x} = \tilde{\mathbf{x}}$ (see *maximum likelihood vs. empirical risk minimization*)
- **Robustness** properties:
  some losses more robust towards outliers than others
- **Optimization** properties: computational complexity of

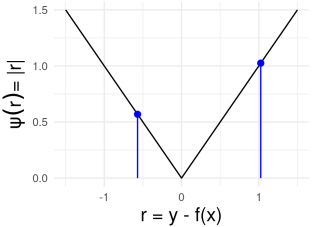
  $$\arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

  is influenced by choice of the loss

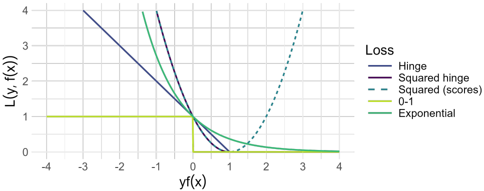
LOSSES WITH ONE ARGUMENT



- Regr. losses often only depend on **residuals** $r(\mathbf{x}) := y - f(\mathbf{x})$
- Classif. losses usually in terms of **margin**: $\nu(\mathbf{x}) := y \cdot f(\mathbf{x})$



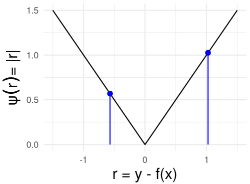
Distance-based: $L1$



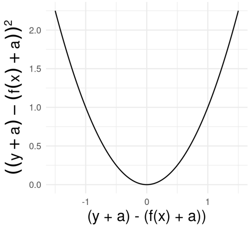
SOME BASIC PROPERTIES

A loss is

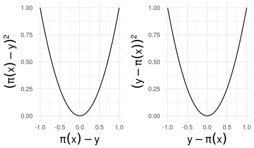
- **symmetric** if $L(y, f(\mathbf{x})) = L(f(\mathbf{x}), y)$
- **translation-invariant** if $L(y + a, f(\mathbf{x}) + a) = L(y, f(\mathbf{x}))$, $a \in \mathbb{R}$
- **distance-based** if it can be written in terms of residual
  $L(y, f(\mathbf{x})) = \psi(r)$ for some $\psi : \mathbb{R} \to \mathbb{R}$, and $\psi(r) = 0 \Leftrightarrow r = 0$



Distance-based: $L1$



Transl.-invariant.: $L2$



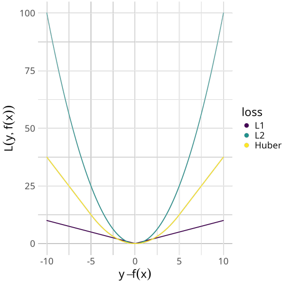
Symmetric: Brier score

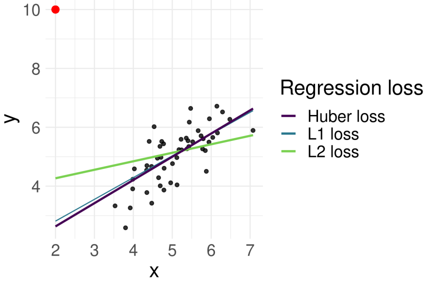
ROBUSTNESS

Outliers (in y) have large residuals $r(\mathbf{x}) = y - f(\mathbf{x})$. Some losses are more affected by large residuals than others. If loss goes up superlinearly (e.g. L2) it is not robust, linear (L1) or even sublinear losses are more robust.

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

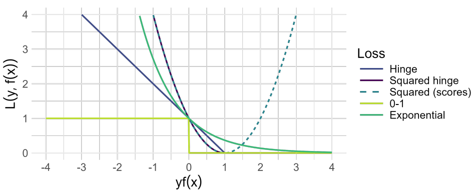
As a consequence, a model is less influenced by outliers than by “inliers” if the loss is **robust**.
Outliers e.g. strongly influence $L2$.





OPTIMIZATION PROPERTIES: SMOOTHNESS

- Measured by number of continuous derivatives. $f \in \mathcal{C}^k$ for k-times cont. differentiable. f and smooth for $f \in \mathcal{C}^\infty$
- Usually want to have at least gradients in optimization of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$
- If loss is not differentiable, 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, exponential and squared hinge losses are continuously differentiable. Hinge loss is continuous but not differentiable. 0-1 loss is not even continuous.

OPTIMIZATION PROPERTIES: CONVEXITY

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

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

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

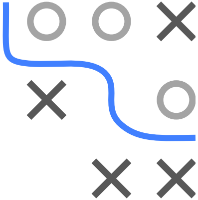
- In optimization, convex problems have several convenient properties, e.g. all local minima are global

- 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

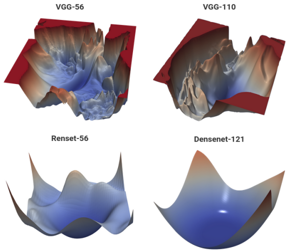
- Above holds for arbitrary functions, not only risks

OPTIMIZATION PROPERTIES: CONVEXITY



- Convexity of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ depends both on convexity of $L(\cdot)$ (given in most cases) and $f(\mathbf{x} \mid \boldsymbol{\theta})$ (often problematic)
- If we model our data using an exponential family distribution, we always get convex losses ▸ Wedderburn 1976
- For $f(\mathbf{x} \mid \boldsymbol{\theta})$ linear in $\boldsymbol{\theta}$, linear/logistic/softmax/poisson/... regression are convex problems (all GLMs)!

The problem on the bottom right is convex, the others are not (note that very high-dimensional surfaces are coerced into 3D here).



▸ Click for source