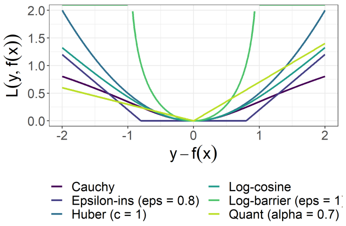
Introduction to Machine Learning

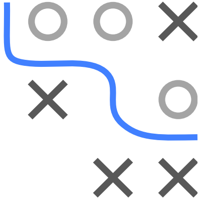
Advanced Risk Minimization
Advanced Regression Losses



Learning goals

- Know the Huber loss
- Know the log-cosh loss
- Know the Cauchy loss
- Know the log-barrier loss
- Know the ϵ-insensitive loss
- Know the quantile loss

ADVANCED LOSS FUNCTIONS



Special loss functions can be used to estimate non-standard posterior components, to measure errors customarily or which are designed to have special properties like robustness.

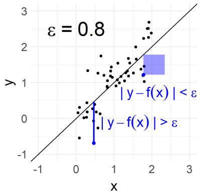
Examples:

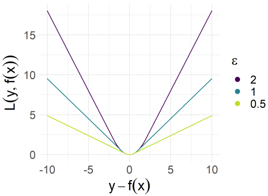
- Quantile loss: Overestimating a clinical parameter might not be as bad as underestimating it.
- Log-barrier loss: Extremely under- or overestimating demand in production would put company profit at risk.
- ϵ-insensitive loss: A certain amount of deviation in production does no harm, larger deviations do.

HUBER LOSS

$$L(y, f) = \begin{cases} \frac{1}{2}(y - f)^2 & \text{if } |y - f| \leq \epsilon \\ \epsilon|y - f| - \frac{1}{2}\epsilon^2 & \text{otherwise} \end{cases}, \quad \epsilon > 0$$

- Piece-wise combination of $L1/L2$ to have robustness/smoothness
- Analytic properties: convex, differentiable (once)



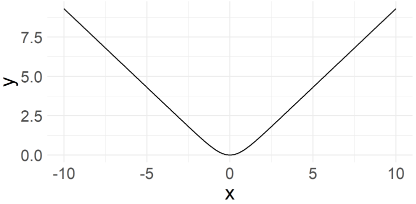


- Risk minimizer and optimal constant do not have a closed-form solution. To fit a model numerical optimization is necessary.
- Solution behaves like **trimmed mean**: a (conditional) mean of two (conditional) quantiles.

LOG-COSH LOSS ▸ R. A. Saleh and A. Saleh 2022

$$L(y, f) = \log(\cosh(|y - f|)) \quad \text{where} \cosh(x) := \frac{e^x + e^{-x}}{2}$$

- Logarithm of the hyperbolic cosine of the residual.
- Approximately equal to $0.5(|y - f|)^2$ for small residuals and to $|y - f| - \log 2$ for large residuals, meaning it works a smoothed out $L1$ loss using $L2$ around the origin.
- Has all the advantages of Huber loss and is, moreover, twice differentiable everywhere.

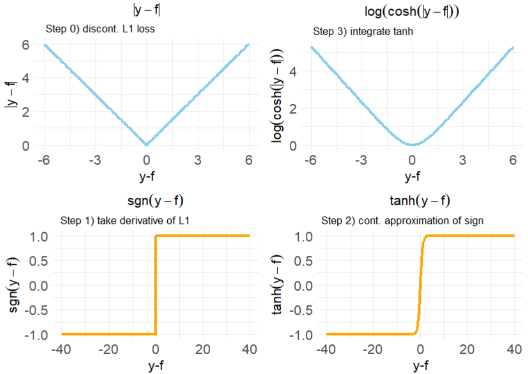


LOG-COSH LOSS ▸ R. A. Saleh and A. Saleh 2022 / 2

What is the idea behind the log-cosh loss?

Essentially, we

1. take derivative of $L1$ loss w.r.t. $y - f$, which is the $\text{sign}(y - f)$ function
2. eliminate discontinuity at 0 by approximating $\text{sign}(y - f)$ using the cont. differentiable $\tanh(y - f)$
3. finally integrate the smoothed sign function "up again" to obtain smoothed $L1$ loss $\log(\cosh(y - f)) = \log(\cosh(|y - f|))$



The log-cosh approach to obtain a differentiable approximation of the $L1$ loss can also be extended to differentiable quantile/pinball losses.

LOG-COSH LOSS ▸ R. A. Saleh and A. Saleh 2022 / 3

The cosh(θ, σ) distribution:

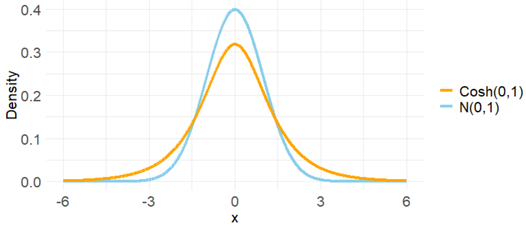
The (normalized) reciprocal $\cosh(x)$ defines a pdf by its positivity on $\mathbb{R}$ and since $\int_{-\infty}^{\infty} \frac{1}{\pi \cosh(x)} dx = 1$.

We can define a location-scale family of distributions (using θ and σ) resembling Gaussians with **heavier tails**.

It is easy to check that ERM using the log-cosh loss is equivalent to MLE of the $\cosh(\theta, 1)$ distribution.

- $p(x|\theta, \sigma) = \frac{1}{\pi\sigma \cosh\left(\frac{x-\theta}{\sigma}\right)}$
- $\mathbb{E}_{X\sim p}[X] = \theta$
- $\mathrm{Var}_{X\sim p}[X] = \frac{1}{4}(\pi^2\sigma^2)$
- $\hat{\theta}^{MLE} = \arg\max_\theta \prod_{i=1}^n \frac{1}{\pi \cosh(x_i - \theta)} \equiv$

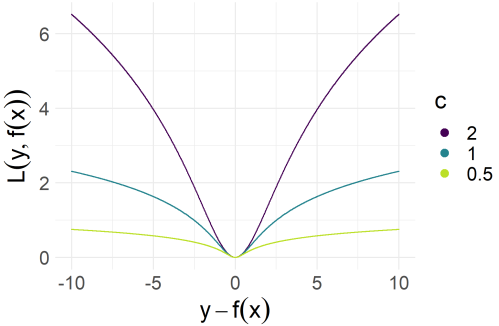
  $\hat{\theta} = \arg\min_\theta \sum_{i=1}^n \log(\cosh(x_i - \theta))$



CAUCHY LOSS

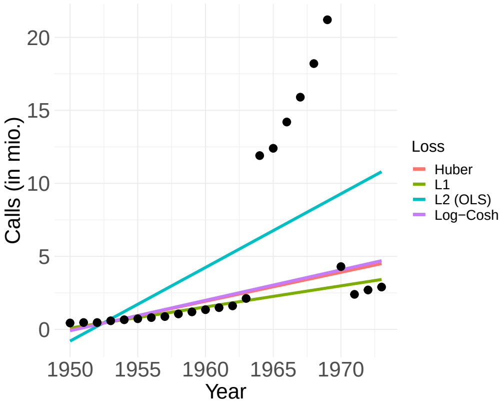
$$L\left(y, f\right) = \frac{c^2}{2} \log \left(1 + \left(\frac{|y - f|}{c} \right)^2 \right), \quad c \in \mathbb{R}$$

- Particularly robust toward outliers (controllable via c).
- Analytic properties: differentiable, but not convex!



TELEPHONE DATA

We now illustrate the effect of using robust loss functions. The telephone data set contains the number of calls (in 10mio units) made in Belgium between 1950 and 1973 ($n = 24$). Outliers are due to a change in measurement without re-calibration for 6 years.



LOG-BARRIER LOSS

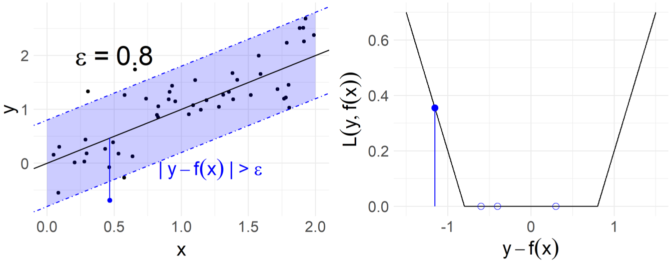
$$L(y, f) = \begin{cases} -\epsilon^2 \cdot \log\left(1 - \left(\frac{|y-f|}{\epsilon}\right)^2\right) & \text{if } |y - f| \leq \epsilon \\ \infty & \text{if } |y - f| > \epsilon \end{cases}$$

- Behaves like $L2$ loss for small residuals
- We use this if we don't want residuals larger than ϵ at all
- No guarantee that the risk minimization problem has a solution
- Plot shows log-barrier loss for $\epsilon = 2$:

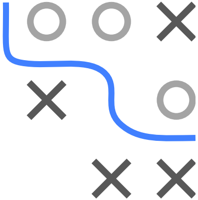
ϵ-INSENSITIVE LOSS

$$L(y, f) = \begin{cases} 0 & \text{if } |y - f| \leq \epsilon \\ |y - f| - \epsilon & \text{otherwise} \end{cases}, \quad \epsilon \in \mathbb{R}_+$$

- Modification of $L1$ loss, errors below ϵ accepted without penalty
- Used in SVM regression
- Properties: convex and not differentiable for $y - f \in \{-\epsilon, \epsilon\}$

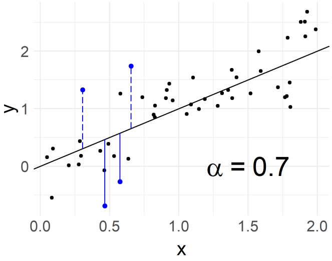


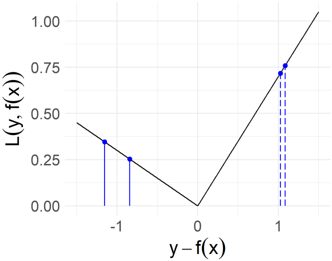
QUANTILE LOSS / PINBALL LOSS



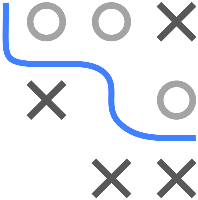
$$L(y,f) = \begin{cases} (1-\alpha)(f-y) & \text{if } y < f \\ \alpha(y-f) & \text{if } y \geq f \end{cases}, \quad \alpha \in (0,1)$$

- Extension of $L1$ loss (equal to $L1$ for $\alpha = 0.5$).
- Weighs either positive or negative residuals more strongly
- $\alpha < 0.5$ ($\alpha > 0.5$) penalty to over-estimation (under-estimation)
- Risk minimizer is (conditional) α-quantile (median for $\alpha = 0.5$)





QUANTILE LOSS / PINBALL LOSS / 2



We simulate $n = 200$ samples from a heteroskedastic LM using the DGP $y = 1 + 0.2x + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.5 + 0.5x)$ and $x \sim \mathcal{U}[0, 10]$. Using the quantile loss, we estimate the conditional α-quantiles for $\alpha \in \{0.05, 0.5, 0.95\}$.