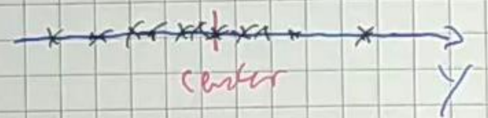
Robust loss functions

- put lower emphasize on outliers ⇒ regression solution is influenced (much) less
- illustration: find the "center" of N 1-dimensional data points



- compute mean: $\bar{y} = \frac{1}{N} \sum_i y_i$
- compute median: $y_{med} = \{ y \mid \#\{y_i < y\} = \#\{y_i > y\} \}$

(is unique, when N is odd: y_{med} middle point in sorted order)

Q: which is better in the presence of outliers - mean or median?

construct toy problem: inlier distribution: $y \sim \mathcal{N}(0, \sigma^2)$

outlier distr. $y \sim \mathcal{N}(0, \tau^2) \quad \tau^2 > \sigma^2$

superposition: "contaminated" distribution: $y \sim (1-\varepsilon)\mathcal{N}(0,\sigma^2) + \varepsilon \mathcal{N}(0,\tau^2)$

$(1-\varepsilon)$: inlier fraction $(0 \leq \varepsilon \leq 1)$

$\frac{\tau}{\sigma} \geq 1$: scale ratio of the two distr.

- setting:
  - choose TS of size N ⇒ determine mean and median
  - repeat this infinitely often with different TS ⇒ determine variance between the means of each TS and medians of each TS

$$\mathrm{Var}_{TS}(\bar{y}) = \frac{\sigma^2}{N}\left(1-\varepsilon + \varepsilon\frac{\tau^2}{\sigma^2}\right), \quad \mathrm{Var}_{TS}(y_{med}) = \frac{\sigma^2}{N} \frac{\pi}{2\left(1-\varepsilon+\varepsilon\frac{\sigma}{\tau}\right)^2}$$

$$\frac{\mathrm{Var}(y_{med})}{\mathrm{Var}(\bar{y})} = \begin{cases} > 1 & \Rightarrow \text{mean is better} \\ < 1 & \Rightarrow \text{median is better} \end{cases}$$

	$\varepsilon = 0$	$\varepsilon = 5\%,\ \frac{\tau}{\sigma} = 4$	$\varepsilon = 5\%,\ \frac{\tau}{\sigma} = 10$	$\varepsilon = 10\%\ \ \frac{\tau}{\sigma} = 3$	$\varepsilon = 10\%\ \frac{\tau}{\sigma} = 10$
$\frac{Var(Y_{med})}{Var(\bar{Y})}$	1.57	1	0.29	1	0.17
	mean better	both are equal	median better	equal	median much better

⇒ conclusion: - median more robust in the presence of outliers
- mean more accurate for the inliers (when outliers are eliminated)

Q: Can we have both? Yes, Huber loss

desirable: results of a machine learning alg. do not change much if we picked a different training set ≙ "variance of an estimator over all possible TSs"

⇒ low variance ≙ low probability to pick unfortunate TS (with high resulting error)

simplest estimator: select a representative for set of 1-D points ("center")

- mean: lowest variance without outliers
- median: lowest variance in presence of outliers

} toy problem: contaminated Gaussian distribution

- Huber loss: is as good as mean w/o outliers, as median with outliers

$\frac{Var(Y_{Huber})}{Var(\bar{Y})}$	1.05		0.21		0.13

Idea of Huber loss:

- penalize inliers like mean
- outliers like median

} in terms of their distance from the representative

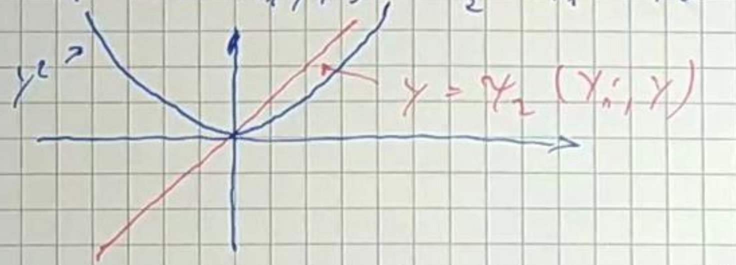
• reformulate computation of mean and median as a optimization problem

$$\hat{Y}_{center} = \arg\min_{Y} \sum_i loss(Y_i, Y)$$

- if $loss(Y_i, Y) = \frac{1}{2}(Y_i - Y)^2 \Rightarrow$ mean: $\frac{\partial}{\partial Y} \sum_i loss(Y_i, Y) = \frac{\partial}{\partial Y} \sum_i \frac{1}{2}(Y_i - Y)^2$

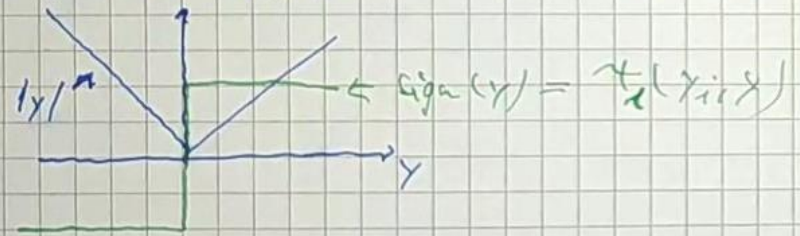
$$= -\sum_i (Y_i - Y) \overset{!}{=} 0$$

$$\hat{Y} = \frac{1}{N} \sum_i Y_i$$



• if $loss(Y_i, Y) = |Y_i - Y| \Rightarrow$ median: $\frac{\partial}{\partial Y} \sum_i |Y_i - Y| = -\sum_i sign(Y_i - Y) \overset{!}{=} 0$

$$= \sum_{i: Y_i - Y > 0} 1 + \sum_{i: Y_i - Y < 0} (-1) \overset{!}{=} 0$$



⇒ positive and negative sums must cancel

$$\#\{Y_i < \hat{Y}\} = \#\{Y_i > \hat{Y}\}$$

introduce: influence function of an instance: $\underbrace{\psi(Y_i, Y)}_{\text{force}} = \frac{\partial}{\partial Y} \underbrace{loss(Y_i, Y)}_{\text{potential}}$

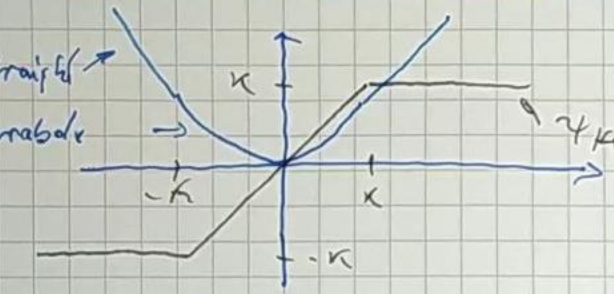
≙ force, how an instance pulls or pushes the representative

squared loss: instances with large distance from the representative exert proportionally larger forces ⇒ good for inliers

absolute loss: all instances have the same force (except for sign)

⇒ good for outliers

combine these behaviors:



$$\psi_{Huber}(Y_i, Y) = \begin{cases} Y_i - Y & \text{if } |Y_i - Y| \le \kappa \leftarrow \text{inlier threshold} \\ \kappa \,\mathrm{sign}(Y_i - Y) & \text{if } |Y_i - Y| \ge \kappa \leftarrow \text{outliers} \end{cases}$$

⇒ find loss by integration ~~Loss Huber $(Y_i - Y) =$~~

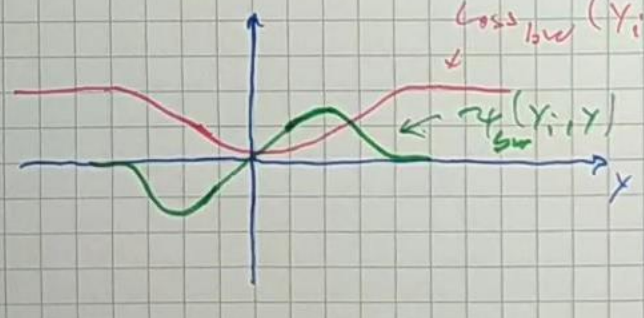
$$\mathrm{Loss}_{Huber}(Y_i, Y) = \begin{cases} \frac{1}{2}(Y_i - Y)^2 & \text{if } |Y_i - Y| \le \kappa \quad \text{squared loss for inliers} \\ \kappa |Y_i - Y| - \frac{\kappa^2}{2} & \text{if } |Y_i - Y| \ge \kappa \end{cases}$$

κ: hyperparameter, if the ~~norm~~ inlier distribution is Gaussian,
optimal choice $\kappa = 1.37\,\sigma$ (σ ≙ std. dev.)

interpretation: the force of outliers doesn't increase with distance ⇒ more stable estimates

Q: can we eliminate the outlier influence entirely ≙ far outliers have <u>zero force</u>

A: Yes, <u>biweight function</u>

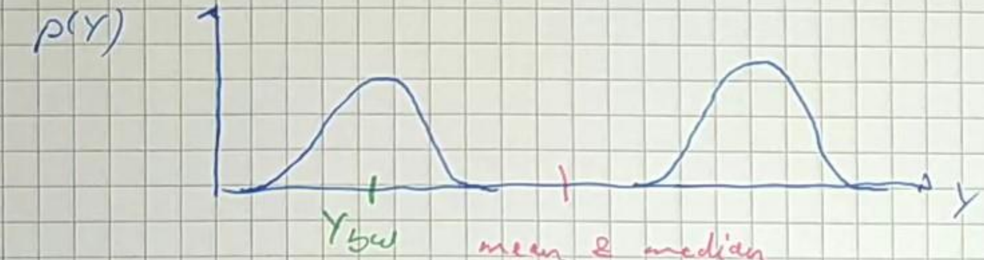


$$\psi_{bw}(Y_i, Y) = \begin{cases} (Y_i - Y)\left(1 - \left(\frac{Y_i - Y}{\kappa}\right)^2\right)^2 & \text{if } |Y_i - Y| \le \kappa \\ 0 & \text{if } |Y_i - Y| > \kappa \end{cases}$$

$$\mathrm{Loss}_{bw}(Y_i, Y) = \begin{cases} 1 - \left(1 - \left(\frac{Y_i - Y}{\kappa}\right)^2\right)^3 & \text{if } |Y_i - Y| \le \kappa \\ 1 & \text{if } |Y_i - Y| \ge \kappa \end{cases}$$

- advantage: only inliers and moderate outliers have influence
- disadvantage: optimization is non-convex, has many local optima

example for behavior: suppose the true distribution is <u>bi-modal</u>



mean & median are <u>bad</u> representatives

bi-weighted loss usually picks either cluster = two representative possibilities ($\hat{=}$ local optima)

$\Rightarrow$ gives a good solution for one camp, but entirely misses the other camp

$\Rightarrow$ in this case, no loss searching for a <u>single</u> representative works well, need a set of <u>diverse</u> representatives or a ~~fu~~ full estimate of the prob. density

unsolved problem: how many representatives are needed?

<u>analytical outlier detection /removal</u>

- if the true model is linear: $Y_i = X_i \beta + \varepsilon_i$, we can derive analytical formula

we learned in context of the bias-variance trade-off: the variance of $\hat{\beta}$

$$\hat{\beta} \sim \mathcal{N}\left(\beta^*, (X^T X)^{-1} \sigma^2\right) \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$\Rightarrow$ exponent of the Gaussian defines a distance

$$d(\hat{\beta}, \beta^*) = \underbrace{\frac{1}{D}}_{\text{normalization for feature space dimension}} (\hat{\beta} - \beta^*)^T \frac{X^T X}{\sigma^2} (\hat{\beta} - \beta^*) \qquad r_i = Y_i - X_i \hat{\beta}$$

approximate σ^2 by the mean squared error: $\sigma^2 \approx \frac{1}{N-D} \sum_i r_i^2 = MSE$

- approximate distance by leave-one-out cross-validation:

$\hat{\beta}$ (full TS) as estimator for β^*

$\hat{\beta}_{-i}$ (solution of TS without instance i) for comparison

$$\Rightarrow \text{Cook's distance } d_i = \frac{(\hat{\beta}_{-i} - \hat{\beta})^T X^T X (\hat{\beta}_{-i} - \hat{\beta})}{D \cdot MSE}$$

d_i large ⇒ instance i has a very big influence on $\hat{\beta}$ ⇒ outlier

d_i small (from an F-distribution) ⇒ normal behavior ⇒ inliers

- we can compute d_i analytically, after training only with full TS

define "hat matrix" $H = X \cdot \underbrace{(X^T X)^{-1} X^T}_{\text{pseudo inverse}}$

it puts the hat on the Y_is : $\hat{Y}_i = X_i \hat{\beta} = X_i (X^T X)^{-1} X^T Y_i = H \cdot Y_i$

interpretation: Y_i : noisy response $\hat{Y}_i$: corrected response on the regression line

$$d_i = \frac{r_i^2}{D \cdot MSE} \cdot \frac{H_{ii}}{(1 - H_{ii})^2}$$

H_{ii} : diagonal elems of H ≙ "leverage" of instance i

- outlier detection: place threshold (hyperparameter) on d_i (ex.: $d_{max} = 1$, $d_{max} = \frac{4}{N}$)

→ for general non-linear problems, outlier detection is a largely unsolved problem
- many heuristic solutions, that work sometimes
- learn the inlier prob. density and define outliers as improbable under this density ⇒ hard in high-dimensions ⇒ Advanced Machine Learning