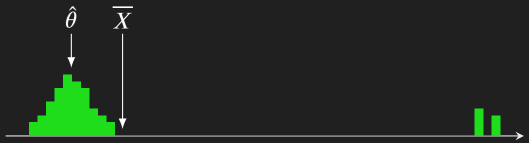
# M-Estimation and Median of Means for Robust Machine Learning

# Name:  Timothée Mathieu (Laboratoire de Mathématiques d'Orsay)



# >>> Outline

1. What is robustness ?
2. Robust Mean Estimation
3. Robust Machine Learning
4. Conclusion

## >>> What is robustness ?

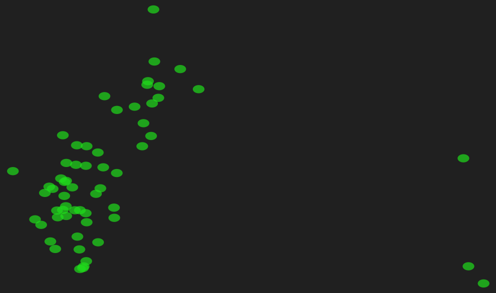
Dataset is either

In [1]: From an heavy-tailed distribution

Out[1]: $X \sim Pareto(\alpha, \beta)$, $X \sim \mathcal{T}(3)$, or more generally $X$ satisfies only moments conditions.

In [2]: Corrupted by outliers

Out[2]: $X_1, \dots, X_n$ contains abnormal data.



Goal: estimate some parameters of this dataset as efficiently as if the data were Gaussian, or at least uncorrupted and light-tailed.

## >>> Corrupted Dataset

Formally,

Adversarial corruption:

$X_1, \dots, X_n$ are corrupted if there exist $X_1', \dots, X_n' \in \mathbb{R}^d$ i.i.d. following a law $P$ that have been modified by an "adversary" to obtain $X_1, \dots, X_n$. The adversary can modify at most $|\mathcal{O}|$ points.

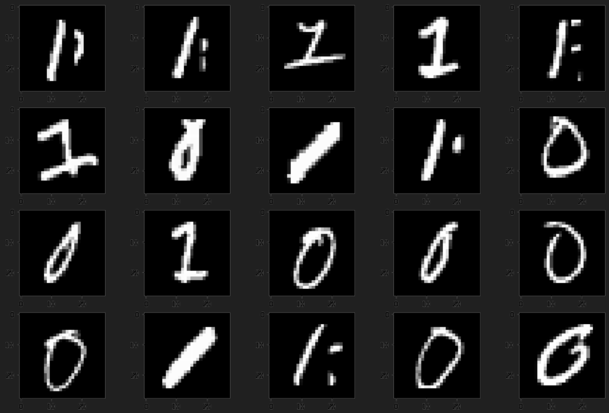
* $X_i = X_i'$ for $i \notin \mathcal{O}$ called inliers
* $X_i$ for $i \in \mathcal{O}$ are called outliers
* We don't make any assumptions on $(X_i)_{i \in \mathcal{O}}$
* $(X_i)_{i \notin \mathcal{O}}$ and $(X_i)_{i \in \mathcal{O}}$ may not be independent
* $(X_i)_{i \notin \mathcal{O}}$ may not be jointly independent.

  Ex: if the adversary modify $|\mathcal{O}|$ points closer to $0$.

>>> Examples

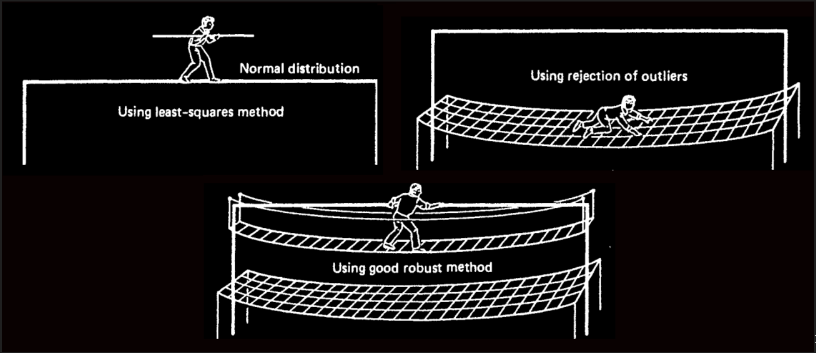
Here are some results of outlier detection on real datasets.

Handwritten 0-1 digits.



Rk: these outliers are not detected as such by a CNN.

# Outlier detection + rejection is not always optimal

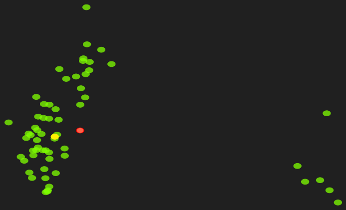


[1]

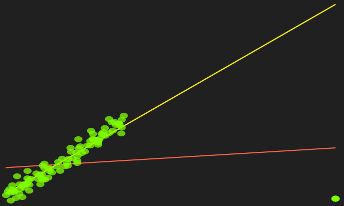
[1] Frank R Hampel et al. *Robust statistics: the approach based on influence functions.*

# >>> Tasks considered in this work

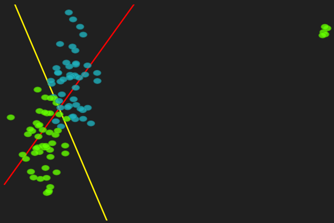
Yellow=Robust   Red=Not Robust



(a) Mean estimation

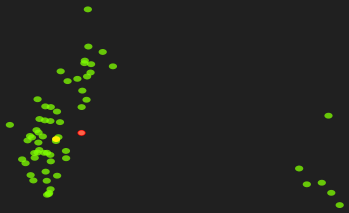


(b) Regression



(c) Classification

# >>> Robust Mean Estimation



## >>> Empirical mean

**Gaussian data:** $X_1, \dots, X_n$ i.i.d $\mathcal{N}(\mu, \sigma^2)$.

Empirical mean gives optimal rate of convergence: for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \sigma\sqrt{\frac{t}{n}}\right) \leq 2\exp(-t)$$

**Heavy-tail data:** if we only have $\mathbb{E}[X^2] < \infty$, Chebychev inequality is tight[2].

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \sigma\sqrt{\frac{t}{n}}\right) \leq \frac{2}{t}$$

**Corrupted data:** one outlier is sufficient to make $\frac{1}{n}\sum_{i=1}^{n} X_i$ arbitrarily large.

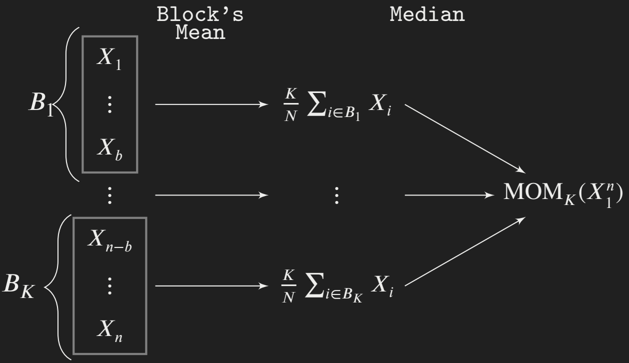
---

[2] Olivier Catoni. "Challenging the empirical mean and empirical variance: A deviation study". In: *Ann. Inst. H. Poincaré Probab. Statist.* (2012).

# >>> Robust Mean Estimation

**Median of Means**[3]. Let $X_1, \dots, X_n$ be i.i.d, let $K \in \mathbb{N}$ and suppose that $K$ divides $n$. Let $B_1, \dots, B_K$ be an equi-partition of $\{1, \dots, n\}$. Define

$$\mathrm{MOM}_K(X_1^n) = \mathrm{Med}\left( \frac{1}{|B_k|} \sum_{i \in B_k} X_i, \quad 1 \le k \le K \right). \tag{1}$$




[3] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons, Inc., New York, 1983.

By property of the median, we have

$$\mathbb{P}\left(\mathrm{MOM}_K(X_1^n) - \mathbb{E}[X] > \varepsilon\right) = \mathbb{P}\left(\mathrm{Med}\left(\frac{1}{|B_k|}\sum_{i\in B_k} X_i - \mathbb{E}[X]\right) > \varepsilon\right)$$
$$\leq \mathbb{P}\left(\sum_{k=1}^{K} \mathbb{1}\left\{\frac{1}{|B_k|}\sum_{i\in B_k} X_i - \mathbb{E}[X] > \varepsilon\right\} \geq \frac{K}{2}\right).$$

This is the deviations of a sum of bounded, i.i.d rvs. Then, via Hoeffding's inequality,

**Theorem (Deviation Median of Means)**

> *Let $X_1, \ldots, X_n, X$ be i.i.d real-valued random variables, with finite variance $\sigma^2$. Then, for all $K \in \{1, \ldots, n\}$,*
>
> $$\mathbb{P}\left(|\mathrm{MOM}_K(X_1^n) - \mathbb{E}[X]| > 2\sigma\sqrt{\frac{K}{n}}\right) \leq 2e^{-K/8} \quad (2)$$

## >>> M-estimators

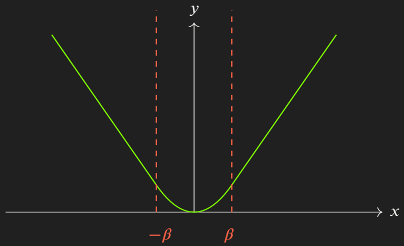
$T(\widehat{P}_n)$ such that

$$T(\widehat{P}_n) \in \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho(X_i - \theta)$$

where $\rho$ is even, convex, non-negative and $\rho(0) = 0$.

**Examples:**

* If $\rho(x) = x^2$: $T(\widehat{P}_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$.
* If $\rho(x) = |x|$: $T(\widehat{P}_n)$ is the sample median.
* **Huber:** Let $\beta > 0$, $\rho_H(x) = \frac{1}{2} x^2 \, \mathbb{1}\{|x| \leq \beta\} + \beta(x - \beta/2) \, \mathbb{1}\{|x| > \beta\}$.



## >>> Deviations of Huber estimator

**Theorem (Deviations of Huber estimator[4])**

> *Suppose that* $\varepsilon_n = |\mathcal{O}|/n \leq 1/32$ *and* $\mathbb{E}_P\left[|X - \mathbb{E}_P[X]|^q\right] < \infty$, $q \geq 2$.
>
> *Then, there exist* $C_1, C_2, C_3, \beta > 0$ *s.t. with probability* $1 - 4e^{-t} - e^{-n/32}$,
>
> $$|T_H(X_1^n) - \mathbb{E}_P[X]| \leq C_1 \sigma \sqrt{\frac{t}{n}} + C_2 \varepsilon_n^{1-1/q} \mathbb{E}[|X - \mathbb{E}[X]|^q]^{1/q}$$
>
> *for any* $t \leq C_3 n^{\frac{q-2}{2q-2}} \left(\frac{\mathbb{E}\left[|X-\mathbb{E}[X]|^2\right]}{\mathbb{E}[|X-\mathbb{E}[X]|^q]}\right)^{\frac{1}{q-1}}$.

we recover a sub-gaussian behavior when $\varepsilon_n = 0$ with only $2$ moments.

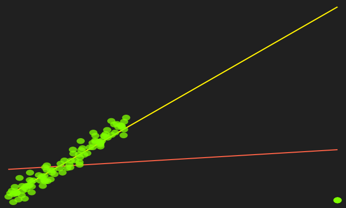
**Remark :** the theorem can be extended to weaker moment assumptions, multivariate setting and to more general $\rho$.

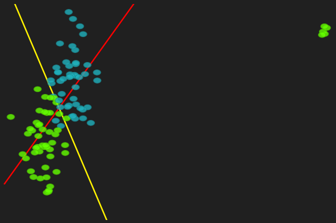
[4] Timothée Mathieu. *Concentration study of M-estimators using the influence function.* 2021. arXiv: 2104.04416 [math.ST].

## >>> Wrap up of Robust Mean estimation part

* We achieve sub-gaussian rates, even if the data are heavy-tailed
* We achieve optimal rates in a corrupted setting
* MoM and Huber's estimator are fast to compute and can be used in place of the sample mean

>>> Robust Machine Learning





Collaboration with: Matthieu Lerasle, Guillaume Lecué, Stanislav Minsker.

## Risk Minimization framework.

Let $(X, Y)$ be a (feature, label) couple.
We are searching for $f \in \mathcal{F}$ such that $f(X) \simeq Y$.

Let

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)]. \tag{3}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function.

## Empirical Risk minimization

The population law is unknown, we can't compute $\mathbb{E}[\ell(f(X), Y)]$.
Instead we use a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

**Problem:** the empirical mean is not robust and $\ell(f(X), Y)$ can be heavy-tailed/corrupted.

## >>> Robust Risk Minimization

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \widehat{E}\left[\ell(f(X_i), Y_i); 1 \leq i \leq n\right]$$

where $\widehat{E}$ is a robust estimator of the mean (Huber or MOM).

Examples:

In [1]: Robust Logistic Regression

Out[1]:

$$\widehat{\beta}_{MOM,K} \in \arg\min_{\beta \in \mathbb{R}^d} \mathrm{MOM}_K\left(\log(1 + \exp(-\langle \beta, X_i \rangle Y_i))\right)$$

In [2]: Robust Least Squares Regression

Out[2]:

$$\widehat{\beta}_{MOM,K} \in \arg\min_{\beta \in \mathbb{R}^d} \mathrm{MOM}_K\left((\beta^T X_i - Y_i)^2\right)$$

These problems can also be stated as Huber minimizations.

## MOM Risk minimization

For the Logistic Regression problem, we have

**Theorem ([5] and[6])**

> *Suppose $X$ has covariance $\Sigma_X$. Assume $n > K > 8|\mathcal{O}|$. Then, with probability greater than $1-2e^{-K/32}$,*
>
> $$R(\widehat{\beta}_{MOM,K}) \leq \inf_{\beta\in\mathbb{R}^d} R(\beta) + 4\sqrt{\|\Sigma_X\|_{op}}\max\left(4\sqrt{\frac{d}{n}}, 2\sqrt{\frac{K}{n}}\right)$$

**Example:** If $\Sigma_X = \sigma^2 I_d$, its largest eigenvalue is $\|\Sigma_X\|_{op} = \sigma^2$.

**Proof idea:** reduce the problem to studying the maximum deviation of a sum of indicator functions.

We get a similar result for Huber risk minimizer, and for more general ERM problems.

---

[5] Matthieu Lerasle, Timothée Mathieu, and Guillaume Lecue. “Robust classification via MOM minimization”. In: *Machine Learning* (2020).

[6] Stanislav Minsker and Timothée Mathieu. “Excess risk bounds in robust empirical risk minimization”. In: *To appear in Information and Inference: A Journal of the IMA* (2020).

# >>> MOM risk minimization in practice

Algorithm to find $\widehat{\beta}_{MOM,K}$ such that

$$\widehat{\beta}_{MOM,K} \in \arg\min_{\beta \in S^{d-1}} \mathrm{MOM}_K\left(\log(1+\exp(-\langle \beta, X_i \rangle Y_i))\right)$$

At iteration $t$,

* Construct the blocks $B_1, \ldots, B_K$ uniformly at random.
* Compute the losses

$$\ell_i(\beta_t) = \log(1+\exp(-\langle \beta_t, X_i \rangle Y_i)).$$

* Find the block $B_{\text{med}}$ such that

$$\mathrm{MOM}_K\left(\ell_i(\beta_t)\right) = \frac{1}{|B_{\text{med}}|}\sum_{i \in B_{\text{med}}} \ell_i(\beta_t).$$

* Do a gradient step of size $\eta_t > 0$ on $B_{\text{med}}$

$$\beta_{t+1} = \beta_t - \eta_t \frac{1}{|B_{\text{med}}|}\sum_{i \in B_{\text{med}}} \nabla_\beta \ell_i(\beta_t).$$

More generally: iterative reweighting algorithm to estimate Huber minimizers or MOM minimizers.

## >>> Robust module in scikit-learn-extra

Implementation in scikit-learn-extra python library of

* Robust linear classification
* Robust linear regression
* Robust KMeans clustering

https://github.com/scikit-learn-contrib/scikit-learn-extra

# >>> Illustrations on Real dataset; Example 1

Prediction of area burnt
by Forest Fires.

* X - x-axis spatial coordinate
* Y - y-axis spatial coordinate
* month - month of the year
* day - day of the week
* FFMC - FFMC index
* DMC - DMC index
* DC - DC index
* ISI - ISI index
* temp - the temperature
* RH - relative humidity in %
* wind - wind speed in km/h
* rain - outside rain in mm/m2
* area - the burned area of the forest (in ha).

## >>> Evaluation

Main problem in evaluation robust ML algorithms:

**There can be outliers in the test set**

**Solutions:** Robust cross validation

* robust score function to estimate the error on a fold (i.e. median absolute error)
* robust aggregation of the error on all the folds (i.e. median of means squared errors on folds)

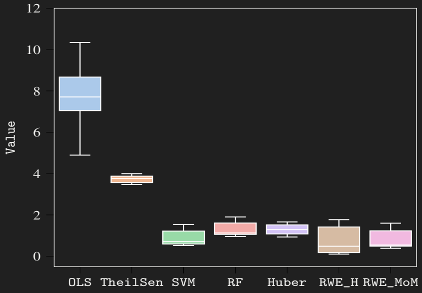
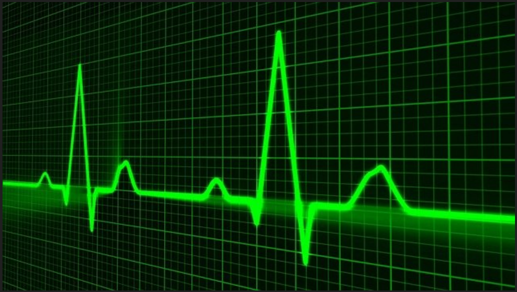



**Figure:** Median Residuals for 10 folds

# >>> Illustrations on Real dataset; Example 2

Prediction of Heart Diseases.



Binary classification problem: predict heart disease or not using features such as cholesterol, age, sex...

The outliers in the target variables $Y$ have a bounded influence.

There can still be highly influencial outliers in the feature space $X$.

Features :

* age
* sex
* chest pain type (4 values)
* resting blood pressure
* serum cholestoral in mg/dl
* fasting blood sugar > 120 mg/dl
* resting electrocardiographic results (values 0,1,2)
* maximum heart rate achieved
* exercise induced angina
* oldpeak = ST depression induced by exercise relative to rest
* the slope of the peak exercise ST segment
* number of major vessels (0-3) colored by flourosopy
* thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

## >>> Evaluation

Classification problem: accuracy is a robust score. Classical CV is alright.

| Algorithm | Accuracy ( in %) |
| --- | --- |
| Logistic Regression | 85.8 |
| Random Forest | 84.5 |
| SVM | 83.8 |
| RWE | 85.1 |
| Polynomial Feature + RWE | 86.8 |
| Polynomial Feature + Logistic Regression | 79.8 |
| Polynomial Feature + outlier removal via RWE + Logistic Regression | 86.7 |

Remark: the outliers in the features are not apparent in one or two dimension visualizations.

## >>> Conclusion

**Other applications**: Robust Multivariate Mean Estimation, Robust Kernel Methods (Robust MMD Computation and two sample testing), Robust KMeans, Outlier detection.

**All codes of algorithms**: on Github.

**Future works**

* Prove the algorithms
* More advanced robust CV
* Robust preprocessing and feature engineering
* Adaptive choice of parameters
* Robust Time series
* ...

Thank you for your attention