# Learning from MOM's principles

Guillaume Lecué

joint works with Geoffrey Chinot, Matthieu Lerasle and Timothée Mathieu

CNRS, ENSAE

19 December 2017 – CIRM, Marseille

# Test of robustness of the LASSO

$$Y = \langle X, t^* \rangle + \mathcal{N}(0,1) \text{ and } (X_1, Y_1), \cdots, (X_N, Y_N) \overset{i.i.d.}{\sim} (X, Y)$$

# Test of robustness of the LASSO

$$Y = \langle X, t^* \rangle + \mathcal{N}(0,1) \text{ and } (X_1, Y_1), \cdots, (X_N, Y_N) \overset{i.i.d.}{\sim} (X, Y)$$

| | |
|---|---|
| $Y_1$ | $X_1^\top$ |
| $Y_2$ | $X_2^\top$ |
| $\cdots$ | $\cdots$ |
| $Y_{100}$ | $X_{100}^\top$ |
| $\cdots$ | $\cdots$ |
| $Y_n$ | $X_n^\top$ |
| $\cdots$ | $\cdots$ |
| $Y_N$ | $X_N^\top$ |

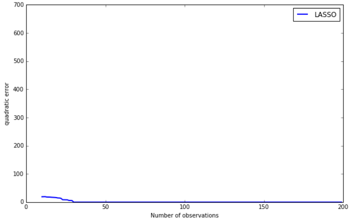
$$\underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, t \rangle)^2 + \sqrt{\frac{2 \log d}{N}} \, \|t\|_1 \right)$$

# Test of robustness of the LASSO

$$Y = \langle X, t^* \rangle + \mathcal{N}(0,1) \text{ and } (X_1, Y_1), \cdots, (X_N, Y_N) \overset{i.i.d.}{\sim} (X, Y)$$

| | |
|---|---|
| $Y_1$ | $X_1^\top$ |
| $Y_2$ | $X_2^\top$ |
| $\cdots$ | $\cdots$ |
| $\cdots$ | $\cdots$ |
| $Y_n$ | $X_n^\top$ |
| $\cdots$ | $\cdots$ |
| $Y_N$ | $X_N^\top$ |

$$\left.\right\} \underset{t \in \mathbb{R}^d}{\text{argmin}} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, t \rangle)^2 + \sqrt{\frac{2 \log d}{N}} \|t\|_1 \right)$$



# Test of robustness of the LASSO

$$Y = \langle X, t^* \rangle + \mathcal{N}(0,1) \text{ and } (X_1, Y_1), \cdots, (X_N, Y_N) \overset{i.i.d.}{\sim} (X, Y)$$

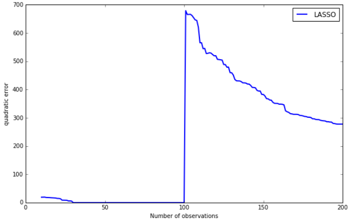
| | |
|---|---|
| $Y_1$ | $X_1^\top$ |
| $Y_2$ | $X_2^\top$ |
| $\cdots$ | $\cdots$ |
| $\tilde{Y}_{100} = 1\bar{M}$ | $\tilde{X}_{100}^\top = (1)_1^d$ |
| $\cdots$ | $\cdots$ |
| $Y_n$ | $X_n^\top$ |
| $\cdots$ | $\cdots$ |
| $Y_N$ | $X_N^\top$ |

$$\Bigg\} \operatorname*{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, t \rangle)^2 + \sqrt{\frac{2 \log d}{N}} \|t\|_1 \right)$$

# Test of robustness of the LASSO

$$Y = \langle X, t^* \rangle + \mathcal{N}(0,1) \text{ and } (X_1, Y_1), \cdots, (X_N, Y_N) \overset{i.i.d.}{\sim} (X, Y)$$

$$\left.\begin{array}{|c|c|}
\hline
Y_1 & X_1^\top \\
Y_2 & X_2^\top \\
\cdots & \cdots \\
\color{red}{\tilde{Y}_{100} = 1\bar{M}} & \color{red}{\tilde{X}_{100}^\top = (1)_1^d} \\
\cdots & \cdots \\
Y_{\color{blue}{n}} & X_{\color{blue}{n}}^\top \\
\cdots & \cdots \\
Y_N & X_N^\top \\
\hline
\end{array}\right\} \operatorname*{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{\color{blue}{n}} \sum_{i=1}^{\color{blue}{n}} (Y_i - \langle X_i, t \rangle)^2 + \sqrt{\frac{2 \log d}{N}} \|t\|_1 \right)$$

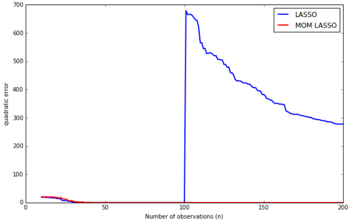


# Test of robustness of the LASSO

$$Y = \langle X, t^* \rangle + \mathcal{N}(0,1) \text{ and } (X_1, Y_1), \cdots, (X_N, Y_N) \overset{i.i.d.}{\sim} (X, Y)$$

| | |
|---|---|
| $Y_1$ | $X_1^\top$ |
| $Y_2$ | $X_2^\top$ |
| $\cdots$ | $\cdots$ |
| $\tilde{Y}_{100} = 1\bar{M}$ | $\tilde{X}_{100}^\top = (1)_1^d$ |
| $\cdots$ | $\cdots$ |
| $Y_n$ | $X_n^\top$ |
| $\cdots$ | $\cdots$ |
| $Y_N$ | $X_N^\top$ |

$$\underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, t \rangle)^2 + \sqrt{\frac{2 \log d}{N}} \|t\|_1 \right)$$



# Test of robustness in classification

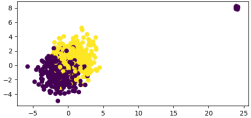
Dataset made of :

- 600 informative data $\overset{i.i.d.}{\sim} (Y, X)$ s.t. $\mathcal{L}(X|Y=1) = \mathcal{N}((1,1), 1.4I)$, $\mathcal{L}(X|Y=-1) = \mathcal{N}((-1,-1), 1.4I)$ and $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1)$.

# Test of robustness in classification

Dataset made of :

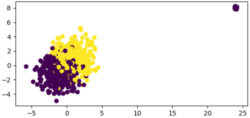
- 600 informative data $\overset{i.i.d.}{\sim} (Y, X)$ s.t. $\mathcal{L}(X|Y=1) = \mathcal{N}((1,1), 1.4I)$, $\mathcal{L}(X|Y=-1) = \mathcal{N}((-1,-1), 1.4I)$ and $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1)$.
- 30 outliers data in the top corner: $Y = -1$ and $X \sim \mathcal{N}((24, 8), 0.1)$



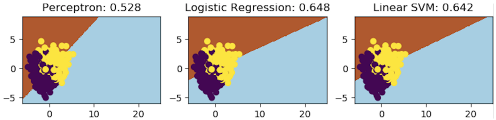
# Test of robustness in classification

Dataset made of :

- 600 informative data $\overset{i.i.d.}{\sim} (Y, X)$ s.t. $\mathcal{L}(X|Y=1) = \mathcal{N}((1,1), 1.4I)$, $\mathcal{L}(X|Y=-1) = \mathcal{N}((-1,-1), 1.4I)$ and $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1)$.
- 30 outliers data in the top corner: $Y = -1$ and $X \sim \mathcal{N}((24, 8), 0.1)$



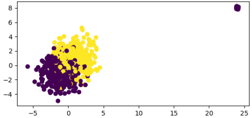
Classical procedures (Perceptron, Logistic regression, SVM):



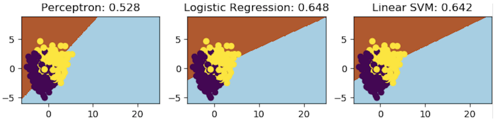
# Test of robustness in classification

Dataset made of :

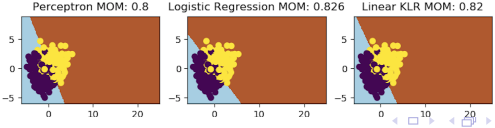
- 600 informative data $\overset{i.i.d.}{\sim} (Y, X)$ s.t. $\mathcal{L}(X|Y=1) = \mathcal{N}((1,1), 1.4I)$, $\mathcal{L}(X|Y=-1) = \mathcal{N}((-1,-1), 1.4I)$ and $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1)$.
- 30 outliers data in the top corner: $Y = -1$ and $X \sim \mathcal{N}((24,8), 0.1)$



Classical procedures (Perceptron, Logistic regression, SVM):




Their MOM (Median Of Means) version:



# Robust statistics : motivations

- Huge datasets are likely to be corrupted by outliers
- heavy-tailed data are common in practice (like in finance)
- Robust theory has been a central issue for a long time

# Robust statistics : motivations

- Huge datasets are likely to be corrupted by outliers
- heavy-tailed data are common in practice (like in finance)
- Robust theory has been a central issue for a long time

Huber's loss function has been designed for that

$$\hat{t} \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} \rho_\kappa(Y_i - \langle X_i, t \rangle) \text{ where } \rho_\kappa(t) = \begin{cases} t^2 & \text{if } |t| \leq \kappa \\ 2\kappa|t| - \kappa^2 & \text{if } |t| > \kappa. \end{cases}$$

is robust to outliers **in the $Y_i$'s** but <u>**not in the $X_i$'s**</u>.

# Robust statistics : motivations

- Huge datasets are likely to be corrupted by outliers
- heavy-tailed data are common in practice (like in finance)
- Robust theory has been a central issue for a long time

Huber's loss function has been designed for that

$$\hat{t} \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} \rho_\kappa(Y_i - \langle X_i, t \rangle) \text{ where } \rho_\kappa(t) = \begin{cases} t^2 & \text{if } |t| \leq \kappa \\ 2\kappa|t| - \kappa^2 & \text{if } |t| > \kappa. \end{cases}$$

is robust to outliers **in the $Y_i$'s** but **<u>not in the $X_i$'s</u>**.

[Huber and Ronchetti, "Robust Statistics"]:

*"..we can act as if the $X_i$'s are free of gross error"*

**The leverage point problem** $\rightsquigarrow$ preprocessing

Construct procedures robust to outliers in the $X_i$'s

**A benchmark result:** Let $(X_i, Y_i)_{i=1}^N$ be

- i.i.d. $\sim (X, Y)$
- $Y = \langle X, t^* \rangle + \zeta$ where $X \sim \mathcal{N}(0, I_{d\times d})$ and $\zeta \sim \mathcal{N}(0, \sigma^2)$ ind. of $X$,

then OLS $\hat{t} \in \underset{t\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^N (Y_i - \langle X_i, t\rangle)^2$ satisfies with probability at least $1 - c_0 \exp(-c_1 d)$,

$$\left\|\hat{t} - t^*\right\|_2^2 \lesssim \frac{\sigma^2 d}{N}$$

when $N \gtrsim d$.

**A benchmark result:** Let $(X_i, Y_i)_{i=1}^N$ be

- i.i.d. $\sim (X, Y)$
- $Y = \langle X, t^* \rangle + \zeta$ where $X \sim \mathcal{N}(0, I_{d\times d})$ and $\zeta \sim \mathcal{N}(0, \sigma^2)$ ind. of $X$,

then OLS $\hat{t} \in \underset{t\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^N (Y_i - \langle X_i, t\rangle)^2$ satisfies with probability at least $1 - c_0 \exp(-c_1 d)$,

$$\left\|\hat{t} - t^*\right\|_2^2 \lesssim \frac{\sigma^2 d}{N}$$

when $N \gtrsim d$.

**Question**

Is it possible to construct an estimator satisfying the very same result when 1) the dataset is corrupted by **outliers** and 2) under **weak moment assumption**

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathbb{E}(Y - \langle X, t \rangle)^2$.

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2$.

**Dataset:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2$.

**Dataset:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- $\mathcal{O}$ stands for *outliers*: **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}(Y - \langle X, t \rangle)^2$.

**Dataset:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- $\mathcal{O}$ stands for *outliers*: **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- $\mathcal{I}$ stands for *informative*:

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathbb{E}(Y - \langle X, t \rangle)^2$.

**Dataset:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- $\mathcal{O}$ stands for *outliers*: **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- $\mathcal{I}$ stands for *informative*:
  1. $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2$.

**Dataset:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- $\mathcal{O}$ stands for *outliers*: **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- $\mathcal{I}$ stands for *informative*:
  1. $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent
  2. $\forall i \in \mathcal{I}, t \in \mathbb{R}^d$,

$$\mathbb{E}\langle X_i, t \rangle^2 = \mathbb{E}\langle X, t \rangle^2 \text{ and } \mathbb{E}(Y_i - \langle X_i, t \rangle)^2 = \mathbb{E}(Y - \langle X, t \rangle)^2$$

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}(Y - \langle X, t \rangle)^2$.

**Dataset:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- $\mathcal{O}$ stands for *outliers*: **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- $\mathcal{I}$ stands for *informative*:
  1. $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent
  2. $\forall i \in \mathcal{I}, t \in \mathbb{R}^d$,

$$\mathbb{E}\langle X_i, t \rangle^2 = \mathbb{E}\langle X, t \rangle^2 \text{ and } \mathbb{E}(Y_i - \langle X_i, t \rangle)^2 = \mathbb{E}(Y - \langle X, t \rangle)^2$$

- $\zeta_i := Y_i - \langle X_i, t^* \rangle$, assume $\forall t \in \mathbb{R}^d$, $\operatorname{var}(\zeta_i \langle X_i, t \rangle) \leq \sigma^2 \mathbb{E}\langle X_i, t \rangle^2$

# From i.i.d. to the $\mathcal{O} \cup \mathcal{I}$ framework

**Aim:** $(X, Y)$ a r.v., estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}(Y - \langle X, t\rangle)^2$.

**Dataset:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- $\mathcal{O}$ stands for *outliers*: **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- $\mathcal{I}$ stands for *informative*:
  1. $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent
  2. $\forall i \in \mathcal{I}, t \in \mathbb{R}^d$,

$$\mathbb{E}\langle X_i, t\rangle^2 = \mathbb{E}\langle X, t\rangle^2 \text{ and } \mathbb{E}(Y_i - \langle X_i, t\rangle)^2 = \mathbb{E}(Y - \langle X, t\rangle)^2$$

- $\zeta_i := Y_i - \langle X_i, t^*\rangle$, assume $\forall t \in \mathbb{R}^d$, $\mathrm{var}(\zeta_i \langle X_i, t\rangle) \leq \sigma^2 \mathbb{E}\langle X_i, t\rangle^2$
- $\forall t \in \mathbb{R}^d$, $\left\|\langle X_i, t\rangle\right\|_{L_2} \leq \theta_1 \left\|\langle X_i, t\rangle\right\|_{L_1}$ (small ball assumption from [Koltchinskii & Mendelson], [van de Geer & Muro], [Oliveira])

# Result for the MOM OLS

In the $\mathcal{O} \cup \mathcal{I}$ framework, the MOM OLS $\tilde{t}_d$ with number of blocks $K = d$ where

$$\tilde{t}_d \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \sup_{t' \in \mathbb{R}^d} \mathrm{MOM}_{K=d}(\ell_t - \ell_{t'})$$

is such that with probability at least $1 - c_0 \exp(-c_1 d)$,

$$\left\| \tilde{t}_d - t^* \right\|_2^2 \lesssim \frac{\sigma^2 d}{N}$$

when $N \gtrsim d$ and $d \gtrsim |\mathcal{O}|$.

# Result for the MOM OLS

In the $\mathcal{O} \cup \mathcal{I}$ framework, the MOM OLS $\tilde{t}_d$ with number of blocks $K = d$ where

$$\tilde{t}_d \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \sup_{t' \in \mathbb{R}^d} \mathrm{MOM}_{K=d}(\ell_t - \ell_{t'})$$

is such that with probability at least $1 - c_0 \exp(-c_1 d)$,

$$\left\| \tilde{t}_d - t^* \right\|_2^2 \lesssim \frac{\sigma^2 d}{N}$$

when $N \gtrsim d$ and $d \gtrsim |\mathcal{O}|$.

**Conclusion**

It is possible to recover the same result in the $\mathcal{O} \cup \mathcal{I}$ framework as in the i.i.d. Gaussian with independent noise framework.

# Construction of MOM estimators I: MOM's principle

**Aim:** Estimate the mean of a real-valued random variable $\mathbb{E}Z$ from $Z_1, \ldots, Z_N \overset{i.i.d.}{\sim} Z$.

# Construction of MOM estimators I: MOM's principle

**Aim:** Estimate the mean of a real-valued random variable $\mathbb{E}Z$ from $Z_1, \dots, Z_N \overset{i.i.d.}{\sim} Z$.

$$
\left.
\begin{array}{l}
\left.\begin{array}{c} Z_1 \\ Z_2 \\ \vdots \\ Z_{N/K} \end{array}\right\} \dfrac{1}{|B_1|}\displaystyle\sum_{i \in B_1} Z_i = P_{B_1} Z \\
\left.\begin{array}{c} Z_{N/K+1} \\ \vdots \\ \vdots \end{array}\right\} \dfrac{1}{|B_2|}\displaystyle\sum_{i \in B_2} Z_i = P_{B_2} Z \\
\left.\begin{array}{c} \vdots \\ Z_{N-1} \\ Z_N \end{array}\right\} \dfrac{1}{|B_K|}\displaystyle\sum_{i \in B_K} Z_i = P_{B_K} Z
\end{array}
\right\}
\mathrm{Median}(P_{B_1} Z, \cdots, P_{B_K} Z) = MOM_K(Z)
$$

# Construction of MOM estimators II: MOM's principle

**Refs:**

* [Nemirovsky, Yudin. 1983]
* [Jerrum, Valiant, Vazirani. 1986]
* [Alon, Matias, Szegedy. 1999]
* [Devroye, Lerasle, Lugosi, Oliveira. 2016]
* [Lugosi, Mendelson, 2017]

# Construction of MOM estimators II: MOM's principle

**Refs:**

* [Nemirovsky, Yudin. 1983]
* [Jerrum, Valiant, Vazirani. 1986]
* [Alon, Matias, Szegedy. 1999]
* [Devroye, Lerasle, Lugosi, Oliveira. 2016]
* [Lugosi, Mendelson, 2017]

**Key idea:** $MOM_K(Z)$ is a subgaussian estimator of $\mathbb{E}Z$ under a $L_2$-moment assumption: if $\|Z\|_{L_2} < \infty$ then with probability at least $1 - c_0 \exp(-c_1 K)$,

$$|MOM_K(Z) - \mathbb{E}Z| \lesssim \sigma \sqrt{\frac{K}{N}}.$$

# Construction of MOM estimators II: MOM's principle

**Refs:**

* [Nemirovsky, Yudin. 1983]
* [Jerrum, Valiant, Vazirani. 1986]
* [Alon, Matias, Szegedy. 1999]
* [Devroye, Lerasle, Lugosi, Oliveira. 2016]
* [Lugosi, Mendelson, 2017]

**Key idea:** *$MOM_K(Z)$ is a subgaussian estimator of $\mathbb{E}Z$ under a $L_2$-moment assumption*: if $\|Z\|_{L_2} < \infty$ then with probability at least $1 - c_0 \exp(-c_1 K)$,

$$|MOM_K(Z) - \mathbb{E}Z| \lesssim \sigma\sqrt{\frac{K}{N}}.$$

Adaptation to $K$ via Lepski's method:

- $\hat{I}_K = \left[MOM_K(Z) - \sigma\sqrt{K/N}, MOM_K(Z) + \sigma\sqrt{K/N}\right]$
- $\hat{K} = \min\left(K : \cap_{k=K}^N \hat{I}_k \neq \emptyset\right)$
- $\tilde{\mu} \in \cap_{k=\hat{K}}^N I_k$

# Construction of MOM estimators III: MOM's principle

**Aim:** We are given:

- $(X, Y)$, $F$ and $f^* \in \underset{f \in F}{\operatorname{argmin}}\, R(f)$ where $R(f) = \mathbb{E}\ell_f(X, Y)$ like $\ell_f(x, y) = (y - f(x))^2, \log(1 + e^{-yf(x)}), (1 - yf(x))_+, \rho_\kappa(y - f(x))$
- $(X_1, Y_1), \dots, (X_N, Y_N)$ some data.

# Construction of MOM estimators III: MOM's principle

**Aim:** We are given:

- $(X, Y)$, $F$ and $f^* \in \underset{f\in F}{\text{argmin}}\, R(f)$ where $R(f) = \mathbb{E}\ell_f(X, Y)$ like $\ell_f(x, y) = (y - f(x))^2, \log(1 + e^{-yf(x)}), (1 - yf(x))_+, \rho_\kappa(y - f(x))$
- $(X_1, Y_1), \dots, (X_N, Y_N)$ some data.

We want to

- Estimate $f^*$: w.h.p. $\left\|\hat{f} - f^*\right\|_{L_2}^2 \leq \textit{rate}$
- Predict $Y$: w.h.p. $R(\hat{f}) \leq \inf_{f\in F} R(f) + \textit{residue}$

# Construction of MOM estimators III: MOM's principle

**Aim:** We are given:

- $(X, Y)$, $F$ and $f^* \in \underset{f \in F}{\text{argmin}}\, R(f)$ where $R(f) = \mathbb{E}\ell_f(X, Y)$ like $\ell_f(x, y) = (y - f(x))^2, \log(1 + e^{-yf(x)}), (1 - yf(x))_+, \rho_\kappa(y - f(x))$
- $(X_1, Y_1), \dots, (X_N, Y_N)$ some data.

We want to

- Estimate $f^*$: w.h.p. $\left\| \hat{f} - f^* \right\|_{L_2}^2 \leq \mathit{rate}$
- Predict $Y$: w.h.p. $R(\hat{f}) \leq \inf_{f \in F} R(f) + \mathit{residue}$

**Classical approach via ERM**: $\hat{f} \in \underset{f \in F}{\text{argmin}}\, R_N(f)$ where

$$R_N(f) = P_N \ell_f = \frac{1}{N} \sum_{i=1}^{N} \ell_f(X_i, Y_i)$$

# Construction of MOM estimators III: MOM's principle

**Aim:** We are given:

- $(X, Y)$, $F$ and $f^* \in \underset{f \in F}{\text{argmin}}\, R(f)$ where $R(f) = \mathbb{E}\ell_f(X, Y)$ like
  $\ell_f(x, y) = (y - f(x))^2, \log(1 + e^{-yf(x)}), (1 - yf(x))_+, \rho_\kappa(y - f(x))$
- $(X_1, Y_1), \ldots, (X_N, Y_N)$ some data.

We want to

- Estimate $f^*$: w.h.p. $\left\|\hat{f} - f^*\right\|_{L_2}^2 \leq \textit{rate}$
- Predict $Y$: w.h.p. $R(\hat{f}) \leq \inf_{f \in F} R(f) + \textit{residue}$

**Classical approach via ERM**: $\hat{f} \in \underset{f \in F}{\text{argmin}}\, R_N(f)$ where

$$R_N(f) = P_N \ell_f = \frac{1}{N} \sum_{i=1}^{N} \ell_f(X_i, Y_i)$$

**Main idea**

Replace the (non-robust) empirical mean $P_N \ell_f$ by a MOM $MOM_K(\ell_f)$ to estimate $R(f) = P\ell_f$

# Construction of MOM estimators IV: MOM's principle

1) **MOM minimizer:** $\bar{f} \in \operatorname*{argmin}_{f \in F} MOM_K(\ell_f)$ where

$$MOM_K(\ell_f) = Median(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f).$$

# Construction of MOM estimators IV: MOM's principle

1) **MOM minimizer:** $\bar{f} \in \operatorname*{argmin}_{f \in F} MOM_K(\ell_f)$ where

$$MOM_K(\ell_f) = Median(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f).$$

slow rates but efficient algorithms

# Construction of MOM estimators IV: MOM's principle

1) **MOM minimizer:** $\bar{f} \in \underset{f \in F}{\mathrm{argmin}}\ MOM_K(\ell_f)$ where

$$MOM_K(\ell_f) = Median(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f).$$

slow rates but efficient algorithms

2) **Le Cam's Test-estimator** based on the test: "$f$ is better than $g$ when $MOM_K(\ell_f - \ell_g) < 0$"

# Construction of MOM estimators IV: MOM's principle

1) **MOM minimizer:** $\bar{f} \in \operatorname*{argmin}_{f \in F} MOM_K(\ell_f)$ where

$$MOM_K(\ell_f) = Median(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f).$$

> slow rates but efficient algorithms

2) **Le Cam's Test-estimator** based on the test: "$f$ is better than $g$ when $MOM_K(\ell_f - \ell_g) < 0$"

> fast minimax rates but no algorithms

# Construction of MOM estimators IV: MOM's principle

1) **MOM minimizer:** $\bar{f} \in \underset{f \in F}{\operatorname{argmin}} \, MOM_K(\ell_f)$ where

$$MOM_K(\ell_f) = Median(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f).$$

slow rates but efficient algorithms

2) **Le Cam's Test-estimator** based on the test: "$f$ is better than $g$ when $MOM_K(\ell_f - \ell_g) < 0$"

fast minimax rates but no algorithms

see also the "tournament estimator" from [Lugosi & Mendelson]

# Construction of MOM estimators IV: MOM's principle

1) **MOM minimizer:** $\bar{f} \in \operatorname*{argmin}_{f \in F} MOM_K(\ell_f)$ where

$$MOM_K(\ell_f) = Median(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f).$$

> slow rates but efficient algorithms

2) **Le Cam's Test-estimator** based on the test: "$f$ is better than $g$ when $MOM_K(\ell_f - \ell_g) < 0$"

> fast minimax rates but no algorithms

see also the "tournament estimator" from [Lugosi & Mendelson]

3) **Minmax MOM estimator:**

$$\tilde{f} \in \operatorname*{argmin}_{f \in F} \sup_{g \in F} MOM_K(\ell_f - \ell_g)$$

# Construction of MOM estimators IV: MOM's principle

1) **MOM minimizer:** $\bar{f} \in \operatorname*{argmin}_{f \in F} MOM_K(\ell_f)$ where

$$MOM_K(\ell_f) = Median(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f).$$

> slow rates but efficient algorithms

2) **Le Cam's Test-estimator** based on the test: "$f$ is better than $g$ when $MOM_K(\ell_f - \ell_g) < 0$"

> fast minimax rates but no algorithms

see also the "tournament estimator" from [Lugosi & Mendelson]

3) **Minmax MOM estimator:**

$$\tilde{f} \in \operatorname*{argmin}_{f \in F} \sup_{g \in F} MOM_K(\ell_f - \ell_g)$$

> fast minimax rates and efficient algorithms

# Minmax MOM estimator. Statistical properties I

**Aims:** $(X, Y)$, estimate $f^* \in \operatorname*{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$ and predict $Y$

# Minmax MOM estimator. Statistical properties I

**Aims:** $(X, Y)$, estimate $f^* \in \underset{f \in F}{\operatorname{argmin}} \mathbb{E}(Y - f(X))^2$ and predict $Y$

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

# Minmax MOM estimator. Statistical properties I

**Aims:** $(X, Y)$, estimate $f^* \in \underset{f\in F}{\text{argmin}}\, \mathbb{E}(Y - f(X))^2$ and predict $Y$

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i\in\mathcal{O}} \cup \{(X_i, Y_i)\}_{i\in\mathcal{I}}$$

where:

- **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$

# Minmax MOM estimator. Statistical properties I

**Aims:** $(X, Y)$, estimate $f^* \in \underset{f \in F}{\operatorname{argmin}} \, \mathbb{E}(Y - f(X))^2$ and predict $Y$

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- on the *informative* data:

# Minmax MOM estimator. Statistical properties I

**Aims:** $(X, Y)$, estimate $f^* \in \underset{f \in F}{\operatorname{argmin}}\, \mathbb{E}(Y - f(X))^2$ and predict $Y$

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- on the *informative* data:
  1. $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent

# Minmax MOM estimator. Statistical properties I

**Aims:** $(X, Y)$, estimate $f^* \in \underset{f \in F}{\operatorname{argmin}}\, \mathbb{E}(Y - f(X))^2$ and predict $Y$

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- on the *informative* data:
  1. $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent
  2. $\forall i \in \mathcal{I}, f \in F$ $\quad \|f(X_i) - f^*(X_i)\|_{L_2} = \|f(X) - f^*(X)\|_{L_2}$
     $\|Y_i - f(X_i)\|_{L_2} = \|Y - f(X)\|_{L_2}$

# Minmax MOM estimator. Statistical properties I

**Aims:** $(X, Y)$, estimate $f^* \in \underset{f \in F}{\text{argmin}}\, \mathbb{E}(Y - f(X))^2$ and predict $Y$

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

$$\{(X_1, Y_1), \cdots, (X_N, Y_N)\} = \{(X_i, Y_i)\}_{i \in \mathcal{O}} \cup \{(X_i, Y_i)\}_{i \in \mathcal{I}}$$

where:

- **no assumption** on the $(X_i, Y_i), i \in \mathcal{O}$
- on the *informative* data:
  1. $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent
  2. $\forall i \in \mathcal{I}, f \in F$ $\quad \|f(X_i) - f^*(X_i)\|_{L_2} = \|f(X) - f^*(X)\|_{L_2}$
     $\quad \|Y_i - f(X_i)\|_{L_2} = \|Y - f(X)\|_{L_2}$
- $\zeta_i := Y_i - f^*(X_i)$, we assume that for all $f \in F$

$$\mathrm{var}(\zeta_i(f(X_i) - f^*(X_i))) \le \sigma^2 \mathbb{E}(f(X_i) - f^*(X_i))^2$$

- $\forall f \in F$, $\|f(X_i) - f^*(X_i)\|_{L_2} \le \theta_1 \|f(X_i) - f^*(X_i)\|_{L_1}$ (SBA)

# Minmax MOM estimator. Statistical properties II

Two fixed points measuring the complexity of the problem:

$$r_Q(\gamma_Q) = \inf\left\{ r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{N}{2},\ \mathbb{E} \sup_{\substack{g \in F - f^* \\ \|g\|_{L_P^2} \leq r}} \left| \sum_{i \in J} \epsilon_i g(X_i) \right| \leqslant \gamma_Q |J| r \right\}$$

$$r_M(\gamma_M) = \inf\left\{ r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{N}{2},\ \mathbb{E} \sup_{\substack{g \in F - f^* \\ \|g\|_{L_P^2} \leq r}} \left| \sum_{i \in J} \epsilon_i \zeta_i g(X_i) \right| \leq \gamma_M |J| r^2 \right\}$$

where $\zeta_i = Y_i - f^*(X_i)$.

# Minmax MOM estimator. Statistical properties II

Two fixed points measuring the complexity of the problem:

$$r_Q(\gamma_Q) = \inf\left\{ r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{N}{2}, \ \mathbb{E} \sup_{\substack{g \in F - f^* \\ \|g\|_{L_P^2} \leq r}} \left| \sum_{i \in J} \epsilon_i g(X_i) \right| \leqslant \gamma_Q |J| r \right\}$$

$$r_M(\gamma_M) = \inf\left\{ r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{N}{2}, \ \mathbb{E} \sup_{\substack{g \in F - f^* \\ \|g\|_{L_P^2} \leq r}} \left| \sum_{i \in J} \epsilon_i \zeta_i g(X_i) \right| \leq \gamma_M |J| r^2 \right\}$$

where $\zeta_i = Y_i - f^*(X_i)$.
Let

$$r^* = \max\{r_Q(\gamma_Q), r_M(\gamma_M)\}.$$

$(r^*)^2$ is the minimax rate of convergence in the i.i.d. framework with Gaussian design and Gaussian noise independent of the design [L. & Mendelson].

# Minmax MOM estimator. Statistical properties III

**Theorem**

In the $\mathcal{O} \cup \mathcal{I}$ framework. Let $K \in \left[\max(N(r^*)^2/\sigma^2, |\mathcal{O}|), N\right]$. With probability at least $1 - c_0 \exp(-c_1 K)$, the minmax MOM estimator

$$\hat{f}_K \in \operatorname*{argmin}_{f \in F} \sup_{g \in F} MOM_K(\ell_f - \ell_g)$$

satisfies

$$\left\| \hat{f}_K - f^* \right\|_{L_2}^2 \leq c_3 \frac{\sigma^2 K}{N} \text{ and } R(\hat{f}_K) \leq \inf_{f \in F} R(f) + \frac{c_4 \sigma^2 K}{N}.$$

# Minmax MOM estimator. Statistical properties III

**Theorem**

In the $\mathcal{O} \cup \mathcal{I}$ framework. Let $K \in \left[\max(N(r^*)^2/\sigma^2, |\mathcal{O}|), N\right]$. With probability at least $1 - c_0 \exp(-c_1 K)$, the minmax MOM estimator

$$\hat{f}_K \in \underset{f \in F}{\operatorname{argmin}} \sup_{g \in F} MOM_K(\ell_f - \ell_g)$$

satisfies

$$\left\| \hat{f}_K - f^* \right\|_{L_2}^2 \leq c_3 \frac{\sigma^2 K}{N} \text{ and } R(\hat{f}_K) \leq \inf_{f \in F} R(f) + \frac{c_4 \sigma^2 K}{N}.$$

In particular, for $K = \max(N(r^*)^2, |\mathcal{O}|)$,

$$\left\| \hat{f}_K - f^* \right\|_{L_2}^2 \leq R(\hat{f}_K) - \inf_{f \in F} R(f) \leq c_4 \max\left( (r^*)^2, \frac{\sigma^2 |\mathcal{O}|}{N} \right)$$

$= c_4 (r^*)^2$ (the minimax rate) when $\sigma^2 |\mathcal{O}| \leq N(r^*)^2$.

# Minmax MOM estimator. Statistical properties III

**Theorem**

In the $\mathcal{O} \cup \mathcal{I}$ framework. Let $K \in \left[\max(N(r^*)^2/\sigma^2, |\mathcal{O}|), N\right]$. With probability at least $1 - c_0 \exp(-c_1 K)$, the minmax MOM estimator

$$\hat{f}_K \in \operatorname*{argmin}_{f \in F} \sup_{g \in F} MOM_K(\ell_f - \ell_g)$$

satisfies

$$\left\| \hat{f}_K - f^* \right\|_{L_2}^2 \leq c_3 \frac{\sigma^2 K}{N} \text{ and } R(\hat{f}_K) \leq \inf_{f \in F} R(f) + \frac{c_4 \sigma^2 K}{N}.$$

In particular, for $K = \max(N(r^*)^2, |\mathcal{O}|)$,

$$\left\| \hat{f}_K - f^* \right\|_{L_2}^2 \leq R(\hat{f}_K) - \inf_{f \in F} R(f) \leq c_4 \max\left( (r^*)^2, \frac{\sigma^2 |\mathcal{O}|}{N} \right)$$

$= c_4 (r^*)^2$ (the minimax rate) when $\sigma^2 |\mathcal{O}| \leq N(r^*)^2$.
(then, adaptation to $K$ via Lepski's method).

# Regularized minmax MOM estimators

$$\hat{f}_K \in \underset{f \in F}{\text{argmin}} \sup_{g \in F} MOM_K(\ell_f - \ell_g) + \lambda(\|f\| - \|g\|)$$

General results:

- sparsity oracle inequalities and sparse estimation rates (when $\|\cdot\|$ has some sparsity inducing power)

# Regularized minmax MOM estimators

$$\hat{f}_K \in \underset{f \in F}{\mathrm{argmin}} \ \sup_{g \in F} MOM_K(\ell_f - \ell_g) + \lambda(\|f\| - \|g\|)$$

General results:

- sparsity oracle inequalities and sparse estimation rates (when $\|\cdot\|$ has some sparsity inducing power)
- ”complexity”-based oracle inequality and estimation rate (always).

# Regularized minmax MOM estimators

$$\hat{f}_K \in \underset{f \in F}{\text{argmin}} \ \sup_{g \in F} MOM_K(\ell_f - \ell_g) + \lambda(\|f\| - \|g\|)$$

General results:

- sparsity oracle inequalities and sparse estimation rates (when $\|\cdot\|$ has some sparsity inducing power)
- "complexity"-based oracle inequality and estimation rate (always).

Example: **MOM version of the LASSO:**

$$\hat{t}_K \in \underset{t \in \mathbb{R}^d}{\text{argmin}} \ \sup_{t' \in \mathbb{R}^d} MOM_K(\ell_t - \ell_{t'}) + \lambda_K \left(\|t\|_1 - \|t'\|_1\right)$$

where $\ell_t(x, y) = (y - \langle x, t \rangle)^2$ and

$$\lambda_K \sim \sigma \sqrt{\frac{1}{N} \log\left(\frac{\sigma^2 d}{K}\right)}$$

# MOM version of the LASSO

**Aim:** Estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathbb{E}(Y - \langle X, t \rangle)^2$ w.r.t. $s = \|t^*\|_0$.

# MOM version of the LASSO

**Aim:** Estimate $t^* \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2$ w.r.t. $s = \|t^*\|_0$.

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

- No assumption on $|\mathcal{O}|$ observations s.t. $|\mathcal{O}| \leq N/10$

# MOM version of the LASSO

**Aim:** Estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}(Y - \langle X, t \rangle)^2$ w.r.t. $s = \|t^*\|_0$.

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

- No assumption on $|\mathcal{O}|$ observations s.t. $|\mathcal{O}| \leq N/10$
- $(X_i, Y_i)_{i \in \mathcal{I}} \overset{i.i.d.}{\sim} (X, Y)$:

# MOM version of the LASSO

**Aim:** Estimate $t^* \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2$ w.r.t. $s = \|t^*\|_0$.

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

- No assumption on $|\mathcal{O}|$ observations s.t. $|\mathcal{O}| \leq N/10$
- $(X_i, Y_i)_{i \in \mathcal{I}} \overset{i.i.d.}{\sim} (X, Y)$:
  1. $X$ is isotropic
  2. $\forall t \in \mathbb{R}^d, p \in [c_0 \log(d)], j \in [d]$: $||X^{(j)}||_{L^p} \leq L\sqrt{p}||X^{(j)}||_{L^2}$

# MOM version of the LASSO

**Aim:** Estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}(Y - \langle X, t \rangle)^2$ w.r.t. $s = \|t^*\|_0$.

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

- No assumption on $|\mathcal{O}|$ observations s.t. $|\mathcal{O}| \leq N/10$
- $(X_i, Y_i)_{i \in \mathcal{I}} \overset{i.i.d.}{\sim} (X, Y)$:
  1. $X$ is isotropic
  2. $\forall t \in \mathbb{R}^d, p \in [c_0 \log(d)], j \in [d]$: $||X^{(j)}||_{L^p} \leq L\sqrt{p}||X^{(j)}||_{L^2}$
  3. $\zeta = Y - \langle X, t^* \rangle \in L^{q_0}$ for some $q_0 > 2$

# MOM version of the LASSO

**Aim:** Estimate $t^* \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}(Y - \langle X, t \rangle)^2$ w.r.t. $s = \|t^*\|_0$.

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

- No assumption on $|\mathcal{O}|$ observations s.t. $|\mathcal{O}| \leq N/10$
- $(X_i, Y_i)_{i \in \mathcal{I}} \overset{i.i.d.}{\sim} (X, Y)$:
  1. $X$ is isotropic
  2. $\forall t \in \mathbb{R}^d, p \in [c_0 \log(d)], j \in [d]$: $||X^{(j)}||_{L^p} \leq L\sqrt{p}||X^{(j)}||_{L^2}$
  3. $\zeta = Y - \langle X, t^* \rangle \in L^{q_0}$ for some $q_0 > 2$
  4. $\forall t \in \mathbb{R}^d, \mathrm{var}(\zeta \langle X, t \rangle) \leq \sigma^2 \left\| \langle X, t \rangle \right\|_{L^2}^2$, $\left\| \langle X, t \rangle \right\|_{L^2} \leq \theta_0 \left\| \langle X, t \rangle \right\|_{L^1}$

# MOM version of the LASSO

**Aim:** Estimate $t^* \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2$ w.r.t. $s = \|t^*\|_0$.

**The $\mathcal{O} \cup \mathcal{I}$ framework:**

- No assumption on $|\mathcal{O}|$ observations s.t. $|\mathcal{O}| \leq N/10$
- $(X_i, Y_i)_{i \in \mathcal{I}} \overset{i.i.d.}{\sim} (X, Y)$:
  1. $X$ is isotropic
  2. $\forall t \in \mathbb{R}^d, p \in [c_0 \log(d)], j \in [d]$: $||X^{(j)}||_{L^p} \leq L\sqrt{p}||X^{(j)}||_{L^2}$
  3. $\zeta = Y - \langle X, t^* \rangle \in L^{q_0}$ for some $q_0 > 2$
  4. $\forall t \in \mathbb{R}^d, \mathrm{var}(\zeta \langle X, t \rangle) \leq \sigma^2 \left\| \langle X, t \rangle \right\|_{L^2}^2$, $\left\| \langle X, t \rangle \right\|_{L^2} \leq \theta_0 \left\| \langle X, t \rangle \right\|_{L^1}$

**Theorem**

In the $\mathcal{O} \cup \mathcal{I}$ framework. Let $K \in [\max(s \log(d/s), |\mathcal{O}|), N]$. With probability at least $1 - c_0 \exp(-c_1 K)$, the MOM LASSO $\hat{t}_K$ satisfies

$$\left\| \hat{t}_K - t^* \right\|_2^2 \leq c_3 \frac{\sigma^2 K}{N} \qquad = \max\left( \frac{\sigma^2 s \log(d/s)}{N}, \frac{\sigma^2 |\mathcal{O}|}{N} \right)$$

for $K = \max(s \log(d/s), |\mathcal{O}|)$. (adaptation via Lepski's method)

# Algorithms

# Descent methods for the MOM minimizer I

**Problem:** $u \in \mathbb{R}^d \to MOM_K(\ell_u)$ is not convex (in general) where

$$MOM_K(\ell_u) = Median\left(\frac{1}{|B_1|}\sum_{i\in B_1}(Y_i - \langle X_i, u\rangle)^2, \cdots, \frac{1}{|B_K|}\sum_{i\in B_K}(Y_i - \langle X_i, u\rangle)^2\right)$$

# Descent methods for the MOM minimizer I

**Problem:** $u \in \mathbb{R}^d \to MOM_K(\ell_u)$ is not convex (in general) where

$$MOM_K(\ell_u) = Median\left(\frac{1}{|B_1|}\sum_{i\in B_1}(Y_i - \langle X_i, u\rangle)^2, \cdots, \frac{1}{|B_K|}\sum_{i\in B_K}(Y_i - \langle X_i, u\rangle)^2\right)$$

still there is a natural way to choose a descent direction:

# Descent methods for the MOM minimizer I

**Problem:** $u \in \mathbb{R}^d \to MOM_K(\ell_u)$ is not convex (in general) where

$$MOM_K(\ell_u) = Median\left(\frac{1}{|B_1|}\sum_{i\in B_1}(Y_i - \langle X_i, u\rangle)^2, \cdots, \frac{1}{|B_K|}\sum_{i\in B_K}(Y_i - \langle X_i, u\rangle)^2\right)$$

still there is a natural way to choose a descent direction:

Partition $\mathbb{R}^d = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K$ where, for all $k \in [K]$,

$$\mathcal{C}_k = \left\{u \in \mathbb{R}^d : MOM_K(\ell_u) = P_{B_k}\ell_u\right\}$$

# Descent methods for the MOM minimizer I

**Problem:** $u \in \mathbb{R}^d \to MOM_K(\ell_u)$ is not convex (in general) where

$$MOM_K(\ell_u) = Median\left(\frac{1}{|B_1|}\sum_{i\in B_1}(Y_i - \left\langle X_i, u\right\rangle)^2, \cdots, \frac{1}{|B_K|}\sum_{i\in B_K}(Y_i - \left\langle X_i, u\right\rangle)^2\right)$$

still there is a natural way to choose a descent direction:

Partition $\mathbb{R}^d = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K$ where, for all $k \in [K]$,

$$\mathcal{C}_k = \left\{u \in \mathbb{R}^d : MOM_K(\ell_u) = P_{B_k}\ell_u\right\}$$

Given a point $u_t \in \mathbb{R}^d$:

1. find $k \in [K]$, such that $MOM_K(\ell_{u_t}) = P_{B_k}\ell_{u_t}$ (i.e. $u_t \in \mathcal{C}_k$)

# Descent methods for the MOM minimizer I

**Problem:** $u \in \mathbb{R}^d \to MOM_K(\ell_u)$ is not convex (in general) where

$$MOM_K(\ell_u) = Median\left(\frac{1}{|B_1|}\sum_{i\in B_1}(Y_i - \langle X_i, u\rangle)^2, \cdots, \frac{1}{|B_K|}\sum_{i\in B_K}(Y_i - \langle X_i, u\rangle)^2\right)$$

still there is a natural way to choose a descent direction:

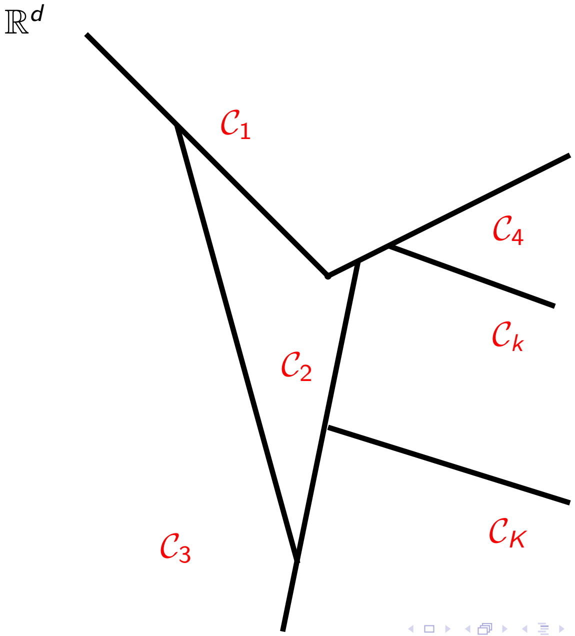
Partition $\mathbb{R}^d = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K$ where, for all $k \in [K]$,

$$\mathcal{C}_k = \left\{u \in \mathbb{R}^d : MOM_K(\ell_u) = P_{B_k}\ell_u\right\}$$

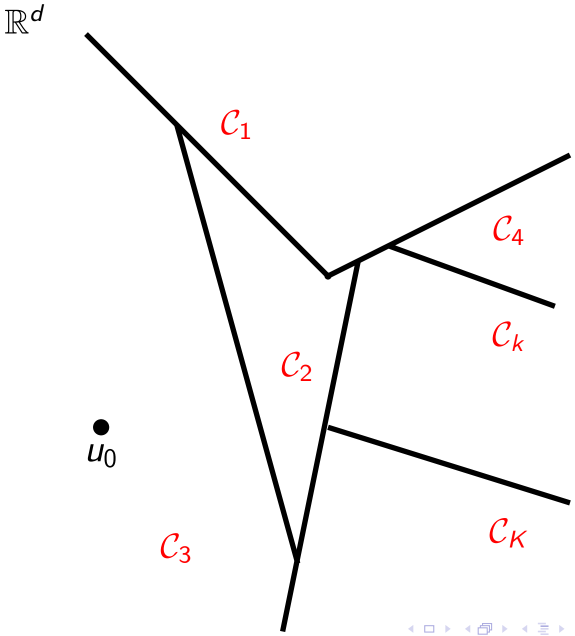
Given a point $u_t \in \mathbb{R}^d$:

1. find $k \in [K]$, such that $MOM_K(\ell_{u_t}) = P_{B_k}\ell_{u_t}$ (i.e. $u_t \in \mathcal{C}_k$)
2. descent direction: $\nabla_t := \nabla(u \to P_{B_k}\ell_u)_{|u=u_t}$
3. $u_{t+1} = u_t - \eta_t \nabla_t$

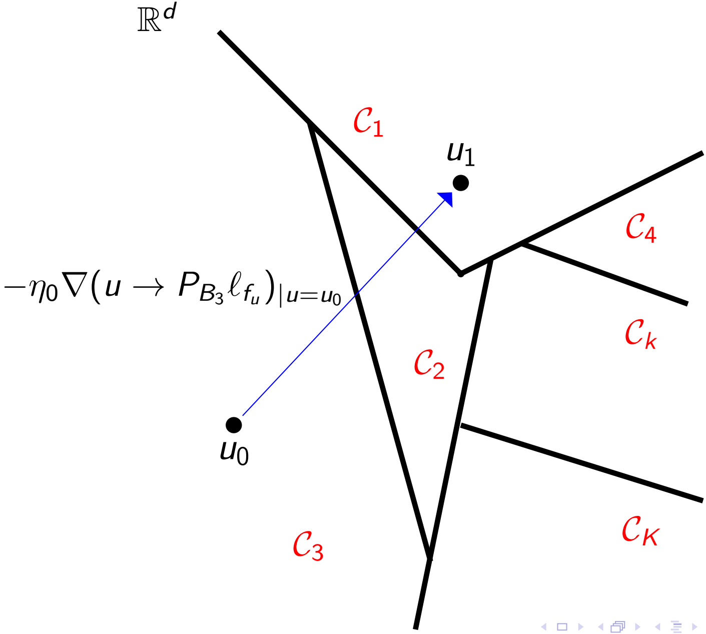
# Descent methods for the MOM minimizer II



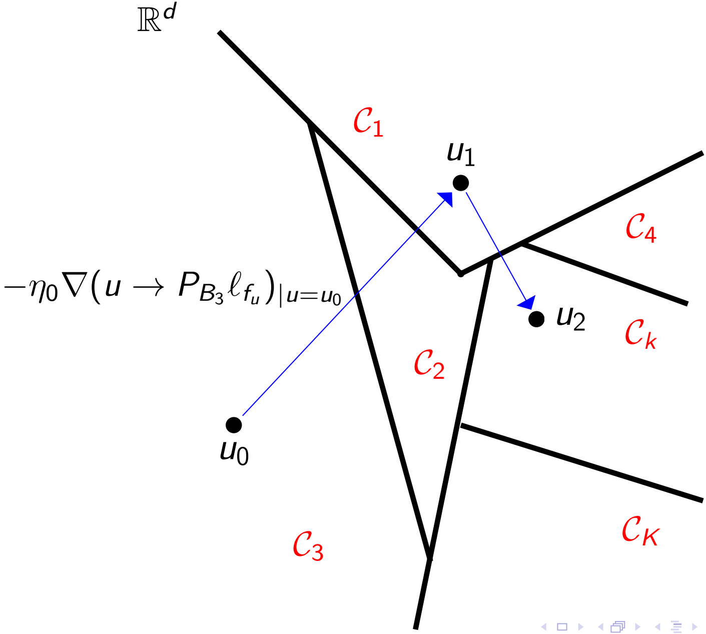
# Descent methods for the MOM minimizer II



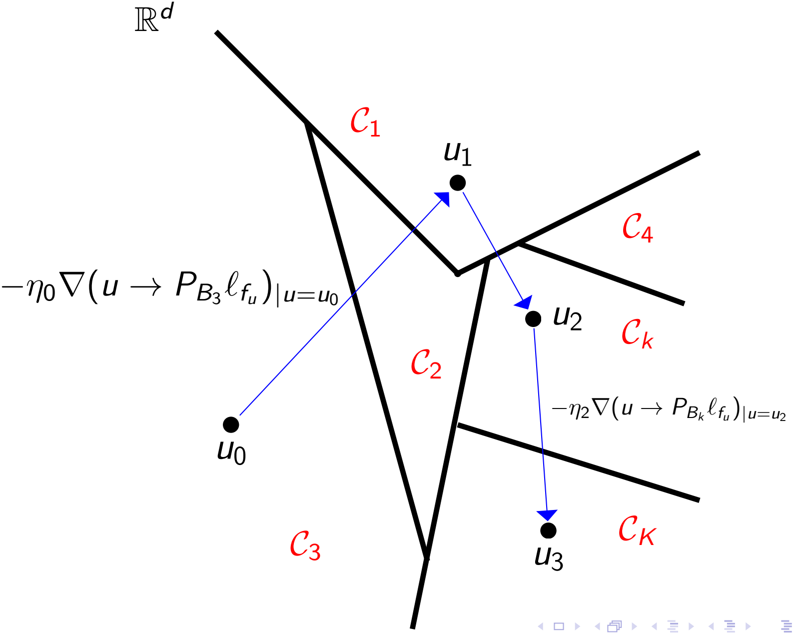
# Descent methods for the MOM minimizer II



# Descent methods for the MOM minimizer II



# Descent methods for the MOM minimizer II



# MOM GD = BGD

$$\left.\begin{array}{c}(X_1, Y_1)\\(X_2, Y_2)\\\vdots\\(X_{N/K}, Y_{N/K})\end{array}\right\} P_{B_1}\ell_{u_t}$$

$$\left.\begin{array}{c}(X_{N/K+1}, Y_{N/K+1})\\\vdots\\\vdots\end{array}\right\} P_{B_2}\ell_{u_t}$$

$$\left.\begin{array}{c}\vdots\\\vdots\\(X_{N-1}, Y_{N-1})\\(X_N, Y_N)\end{array}\right\} P_{B_K}\ell_{u_t}$$

MOM version of the gradient descent = **Block Gradient Descent with a particular choice of block**

1. find $k \in [K]$, s.t. $MOM_K(\ell_{u_t}) = P_{B_k}\ell_{u_t}$

# MOM GD = BGD

$$\left.\begin{array}{c}(X_1, Y_1)\\(X_2, Y_2)\\\vdots\\(X_{N/K}, Y_{N/K})\end{array}\right\} P_{B_1}\ell_{u_t}$$

$$\left.\begin{array}{c}(X_{N/K+1}, Y_{N/K+1})\\\vdots\end{array}\right\} P_{B_2}\ell_{u_t}$$

$$\left.\begin{array}{c}\vdots\\(X_{N-1}, Y_{N-1})\\(X_N, Y_N)\end{array}\right\} P_{B_K}\ell_{u_t}$$

MOM version of the gradient descent = **Block Gradient Descent with a particular choice of block**

1. find $k \in [K]$, s.t. $MOM_K(\ell_{u_t}) = P_{B_k}\ell_{u_t}$
2. descent direction: $\nabla_t := \nabla(u \to P_{B_k}\ell_u)_{|u=u_t}$
3. $u_{t+1} = u_t - \eta_t \nabla_t$

# MOM GD = BGD

$$\left.\begin{array}{c}(X_1, Y_1)\\(X_2, Y_2)\\ \vdots \\(X_{N/K}, Y_{N/K})\end{array}\right\} P_{B_1}\ell_{u_t}$$

$$\left.\begin{array}{c}(X_{N/K+1}, Y_{N/K+1})\\ \vdots \\ \vdots\end{array}\right\} P_{B_2}\ell_{u_t}$$

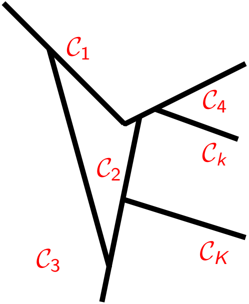
$$\left.\begin{array}{c}\vdots \\(X_{N-1}, Y_{N-1})\\(X_N, Y_N)\end{array}\right\} P_{B_K}\ell_{u_t}$$

MOM version of the gradient descent = **Block Gradient Descent with a particular choice of block**

1. find $k \in [K]$, s.t. $MOM_K(\ell_{u_t}) = P_{B_k}\ell_{u_t}$
2. descent direction: $\nabla_t := \nabla(u \to P_{B_k}\ell_u)_{|u=u_t}$
3. $u_{t+1} = u_t - \eta_t \nabla_t$

**Idea:** Choose the descent block according to its centrality via the median operator $\rightsquigarrow$ "remove outliers" and closer to $\mathbb{E}\ell_{u_t}$

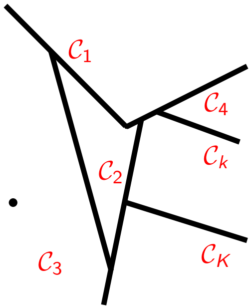
# Pb of local minima => Random blocks




Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\operatorname*{argmin}_{u \in \mathbb{R}^d} P_{B_k} \ell_u$$

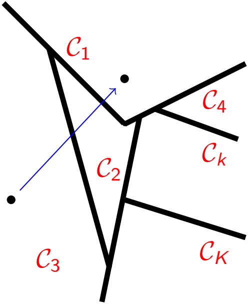
# Pb of local minima => Random blocks




Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\operatorname*{argmin}_{u \in \mathbb{R}^d} P_{B_k} \ell_u$$

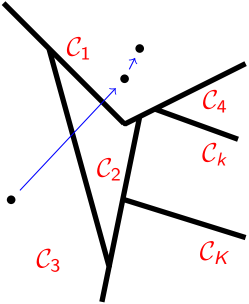
# Pb of local minima => Random blocks




Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\underset{u \in \mathbb{R}^d}{\text{argmin}}\ P_{B_k} \ell_u$$

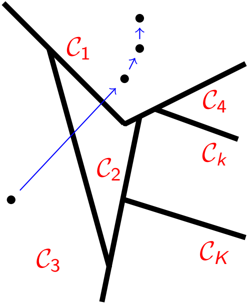
# Pb of local minima => Random blocks




Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\operatorname*{argmin}_{u\in\mathbb{R}^d} P_{B_k}\ell_u$$

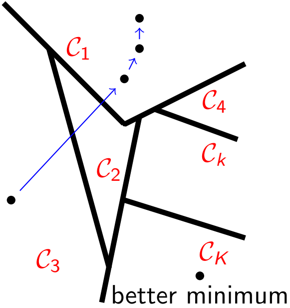
# Pb of local minima => Random blocks




Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \; P_{B_k} \ell_u$$

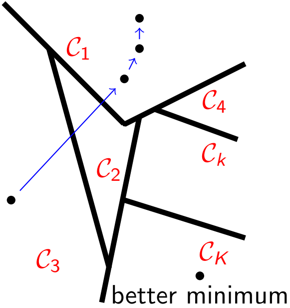
# Pb of local minima => Random blocks




Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\operatorname*{argmin}_{u \in \mathbb{R}^d} P_{B_k} \ell_u$$

# Pb of local minima => Random blocks



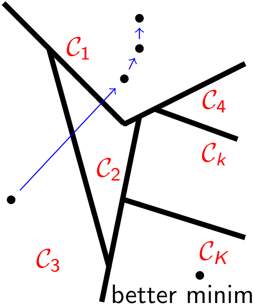

Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\operatorname*{argmin}_{u\in\mathbb{R}^d} P_{B_k}\ell_u$$

**Solution:** choose the blocks of data at random at every step:

1. random partition: $\{1,\dots,N\} = B_1 \cup \dots \cup B_K$
2. median block: $P_{B_k}\ell_{u_t} = MOM_K(\ell_{u_t})$
3. descent direction: $\nabla_t := \nabla(u \to P_{B_k}\ell_u)_{|u=u_t}$
4. $u_{t+1} = u_t - \eta_t \nabla_t$

# Pb of local minima => Random blocks




Local minima if a cell $\mathcal{C}_k$ contains a minimum from

$$\operatorname*{argmin}_{u \in \mathbb{R}^d} P_{B_k} \ell_u$$

**Solution:** choose the blocks of data at random at every step:

1. random partition: $\{1, \dots, N\} = B_1 \cup \dots \cup B_K$
2. median block: $P_{B_k} \ell_{u_t} = MOM_K(\ell_{u_t})$
3. descent direction: $\nabla_t := \nabla(u \to P_{B_k} \ell_u)_{|u = u_t}$
4. $u_{t+1} = u_t - \eta_t \nabla_t$

**MOM GD with random blocks = BSGD with a particular choice of the descent blocks**

# Convergence of the MOM GD with random blocks

**Theorem**

Let $\mathcal{D}_N = \{(X_i, Y_i)_{i=1}^N\}$. Assume that

1. $\|\nabla_u \ell_u(x,y)\|_2 \leq L$
2. $\hat{u} \in \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathbb{E}_{B_1 \cup \cdots \cup B_K} [MOM_K(\ell_u) | \mathcal{D}_N]$ is such that $\forall \epsilon > 0$,

$$\inf_{\|\hat{u}-u\|_2 \geq \epsilon} \left\langle \hat{u} - u, \mathbb{E}[\nabla_u \ell_u(x,y) | \mathcal{D}_N] \right\rangle > 0$$

3. $\sum_t \eta_t^2 < \infty$ and $\sum_t \eta_t = \infty$
4. for $\lambda_d$-almost all $u \in \mathbb{R}^d$, there exists an open set $B$ such that $u \in B$ and for all partition $B_1 \cup \cdots \cup B_K$ and $v \in B$, $\ell_u$ and $\ell_v$ have the same median block.

Then, for almost all dataset $\mathcal{D}_N$,

$$\|u_T - \hat{u}\|_2 \xrightarrow[T \to \infty]{a.s} 0$$

# A descent/ascent algorithm for the minmax MOM estimator

**Idea:** Alternate between ascent (for the max) and descent (for the min).

# A descent/ascent algorithm for the minmax MOM estimator

**Idea:** Alternate between ascent (for the max) and descent (for the min).
Example for the **minmax MOM version of the LASSO:**

$$\hat{u} \in \operatorname*{argmin}_{u \in \mathbb{R}^d} \sup_{u' \in \mathbb{R}^d} MOM_K(\ell_u - \ell_{u'}) + \lambda_K \left( \|u\|_1 - \|u'\|_1 \right)$$

where $\ell_u(x, y) = (y - \langle x, u \rangle)^2$ and $\lambda_K \sim \sigma \sqrt{(1/N) \log(\sigma^2 d / K)}$.

# A descent/ascent algorithm for the minmax MOM estimator

**Idea:** Alternate between ascent (for the max) and descent (for the min).
Example for the **minmax MOM version of the LASSO:**

$$\hat{u} \in \operatorname*{argmin}_{u\in\mathbb{R}^d} \sup_{u'\in\mathbb{R}^d} MOM_K(\ell_u - \ell_{u'}) + \lambda_K\left(\|u\|_1 - \|u'\|_1\right)$$

where $\ell_u(x,y) = (y - \langle x, u\rangle)^2$ and $\lambda_K \sim \sigma\sqrt{(1/N)\log(\sigma^2 d/K)}$.
At iteration $(u_t, u_{t'})$ we do:

- u1 random partition: $\{1,\dots,N\} = B_1 \cup \dots \cup B_K$
- u2 median block: $P_{B_k}(\ell_{u_t} - \ell_{u'_t}) = MOM_K(\ell_{u_t} - \ell_{u'_t})$
- u3 descent direction: $\nabla_t := \nabla(u \to P_{B_k}\ell_u)_{|u=u_t} = -2\mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k u_t)$
- u4 $u_{t+1} = \mathrm{prox}_{\lambda_K\|\cdot\|_1}(u_t - \eta_t \nabla_t)$

# A descent/ascent algorithm for the minmax MOM estimator

**Idea:** Alternate between ascent (for the max) and descent (for the min).
Example for the **minmax MOM version of the LASSO:**

$$\hat{u} \in \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \sup_{u' \in \mathbb{R}^d} MOM_K(\ell_u - \ell_{u'}) + \lambda_K \left(\|u\|_1 - \|u'\|_1\right)$$

where $\ell_u(x, y) = (y - \langle x, u \rangle)^2$ and $\lambda_K \sim \sigma \sqrt{(1/N) \log(\sigma^2 d/K)}$.
At iteration $(u_t, u_{t'})$ we do:

- u1 random partition: $\{1, \ldots, N\} = B_1 \cup \cdots \cup B_K$
- u2 median block: $P_{B_k}(\ell_{u_t} - \ell_{u'_t}) = MOM_K(\ell_{u_t} - \ell_{u'_t})$
- u3 descent direction: $\nabla_t := \nabla(u \to P_{B_k} \ell_u)_{|u=u_t} = -2\mathbb{X}_k^\top (\mathbb{Y}_k - \mathbb{X}_k u_t)$
- u4 $u_{t+1} = \mathrm{prox}_{\lambda_K \|\cdot\|_1}(u_t - \eta_t \nabla_t)$
- $u'1$ random partition: $\{1, \ldots, N\} = B_1 \cup \cdots \cup B_K$
- $u'2$ median block: $P_{B_{k'}}(\ell_{u_{t+1}} - \ell_{u'_t}) = MOM_K(\ell_{u_{t+1}} - \ell_{u'_t})$
- $u'3$ ascent direction: $\nabla'_t := -\nabla(u \to P_{B_{k'}} \ell_u)_{|u=u'_t} = 2\mathbb{X}_{k'}^\top (\mathbb{Y}_{k'} - \mathbb{X}_{k'} u'_t)$
- $u'4$ $u'_{t+1} = \mathrm{prox}_{\lambda_K \|\cdot\|_1}(u'_t + \eta_t \nabla'_t)$

# Simulations: effect of random blocks on local minima

$N = 200$ i.i.d. copies of $(X, Y)$ where

$$Y = \langle X, t^* \rangle + \zeta, \quad X \sim \mathcal{N}(0, I_{d \times d}) \quad \zeta \sim \mathcal{N}(0, 1) \text{ ind. of } X$$

where $d = 500$ and $\|t^*\|_0 = 20$.

# Simulations: effect of random blocks on local minima

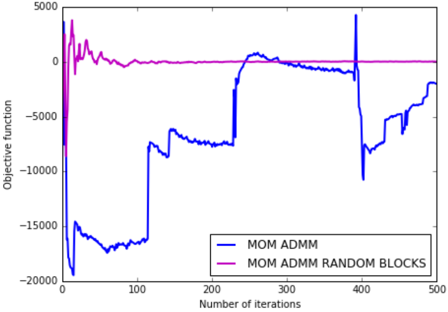
$N = 200$ i.i.d. copies of $(X, Y)$ where

$$Y = \left\langle X, t^* \right\rangle + \zeta, \quad X \sim \mathcal{N}(0, I_{d\times d}) \quad \zeta \sim \mathcal{N}(0,1) \text{ ind. of } X$$

where $d = 500$ and $\left\| t^* \right\|_0 = 20$.

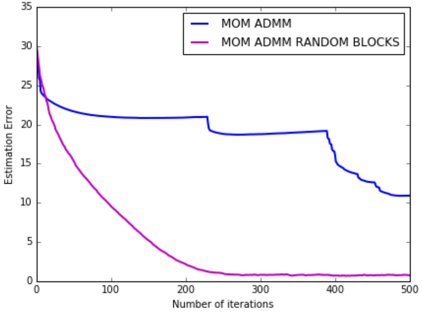
**Objective function**

$MOM_K(\ell_u - \ell_{u'}) + \lambda_K \left( \|u\|_1 - \|u'\|_1 \right)$




**Estimation error**

$\left\| \hat{t} - t^* \right\|_2$



# Adaptation of classical algorithms to their MOM version

| **Objective function** | **Estimation error** |
| --- | --- |
| $MOM_K(\ell_u - \ell_{u'}) + \lambda_K \left(\|u\|_1 - \|u'\|_1\right)$ | $\left\|\hat{t} - t^*\right\|_2$ |

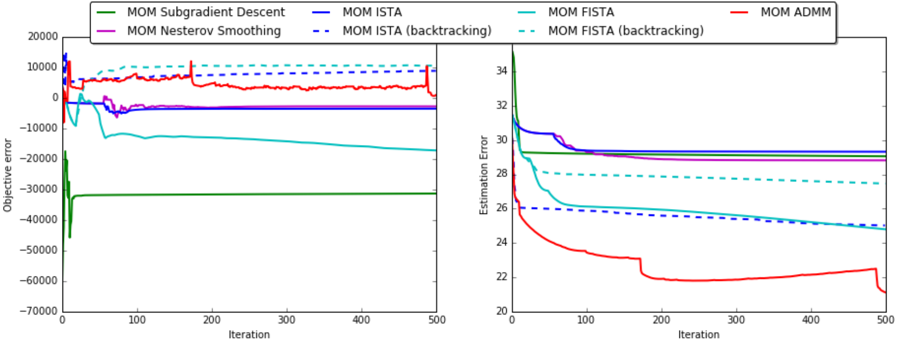
# Adaptation of classical algorithms to their MOM version

**Objective function**

$$MOM_K(\ell_u - \ell_{u'}) + \lambda_K \left( \|u\|_1 - \|u'\|_1 \right)$$

**Estimation error**

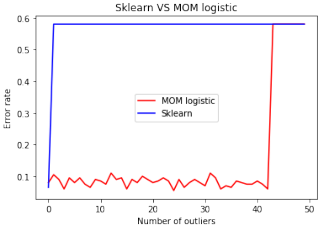
$$\left\| \hat{t} - t^* \right\|_2$$




(non random blocks)

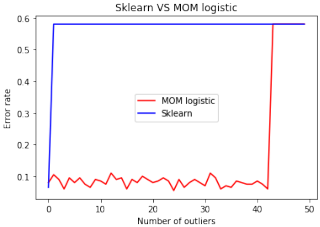
# Test of robustness of minmax MOM estimators

**Logistic Vs MOM logistic** $N = 1000,\ d = 50,\ K = 100$

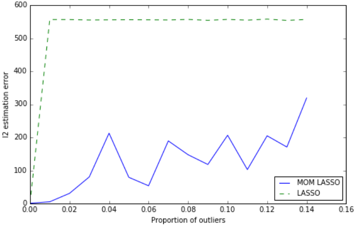


# Test of robustness of minmax MOM estimators

**Logistic Vs MOM logistic** $N = 1000,\ d = 50,\ K = 100$




**LASSO Vs MOM LASSO** $N = 200,\ d = 500,\ s = 10$, adaptive choice of $K$ and $\lambda$



# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

1. split the dataset into $V$ disjoints blocks $\mathcal{D}_1, \dots, \mathcal{D}_V$

# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

1. split the dataset into $V$ disjoints blocks $\mathcal{D}_1, \ldots, \mathcal{D}_V$
2. $\forall v \in [V]$, $\cup_{u \neq v} \mathcal{D}_u$ is used to train a family of estimators

$$\mathcal{F}^{(v)} := \left(\hat{f}^{(v)}_{K,\lambda} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda\right). \tag{1}$$

# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

1. split the dataset into $V$ disjoints blocks $\mathcal{D}_1, \dots, \mathcal{D}_V$
2. $\forall v \in [V]$, $\cup_{u \neq v} \mathcal{D}_u$ is used to train a family of estimators

$$\mathcal{F}^{(v)} := \left( \hat{f}_{K,\lambda}^{(v)} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda \right). \tag{1}$$

3. The remaining $\mathcal{D}_v$ of the dataset is used to test the performance of each estimator in $\mathcal{F}^{(v)}$

# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

1. split the dataset into $V$ disjoints blocks $\mathcal{D}_1, \dots, \mathcal{D}_V$
2. $\forall v \in [V]$, $\cup_{u \neq v} \mathcal{D}_u$ is used to train a family of estimators

$$\mathcal{F}^{(v)} := \left( \hat{f}^{(v)}_{K,\lambda} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda \right). \tag{1}$$

3. The remaining $\mathcal{D}_v$ of the dataset is used to test the performance of each estimator in $\mathcal{F}^{(v)}$
4. $B_1^{(v)} \cup \dots \cup B_{K'}^{(v)}$ is a partition of the test set $\mathcal{D}_v$ into $K'$ blocks

# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

1. split the dataset into $V$ disjoints blocks $\mathcal{D}_1, \ldots, \mathcal{D}_V$
2. $\forall v \in [V]$, $\cup_{u \neq v} \mathcal{D}_u$ is used to train a family of estimators
$$\mathcal{F}^{(v)} := \left(\hat{f}^{(v)}_{K,\lambda} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda\right). \tag{1}$$
3. The remaining $\mathcal{D}_v$ of the dataset is used to test the performance of each estimator in $\mathcal{F}^{(v)}$
4. $B_1^{(v)} \cup \cdots \cup B_{K'}^{(v)}$ is a partition of the test set $\mathcal{D}_v$ into $K'$ blocks
5. for all $v \in [V]$ and $f \in \mathcal{F}^{(v)}$,
$$\mathrm{MOM}^{(v)}_{K'}(\ell_f) = \mathrm{Median}\left(P_{B_1^{(v)}}\ell_f, \cdots, P_{B_{K'}^{(v)}}\ell_f\right) \tag{2}$$

# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

1. split the dataset into $V$ disjoints blocks $\mathcal{D}_1, \dots, \mathcal{D}_V$
2. $\forall v \in [V]$, $\cup_{u \neq v} \mathcal{D}_u$ is used to train a family of estimators

$$\mathcal{F}^{(v)} := \left( \hat{f}^{(v)}_{K,\lambda} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda \right). \tag{1}$$

3. The remaining $\mathcal{D}_v$ of the dataset is used to test the performance of each estimator in $\mathcal{F}^{(v)}$
4. $B_1^{(v)} \cup \dots \cup B_{K'}^{(v)}$ is a partition of the test set $\mathcal{D}_v$ into $K'$ blocks
5. for all $v \in [V]$ and $f \in \mathcal{F}^{(v)}$,

$$\mathrm{MOM}_{K'}^{(v)} (\ell_f) = \mathrm{Median} \left( P_{B_1^{(v)}} \ell_f, \cdots, P_{B_{K'}^{(v)}} \ell_f \right) \tag{2}$$

6. $(\hat{K}, \hat{\lambda})$ minimizes the $\mathrm{MomCv}_V$ criteria

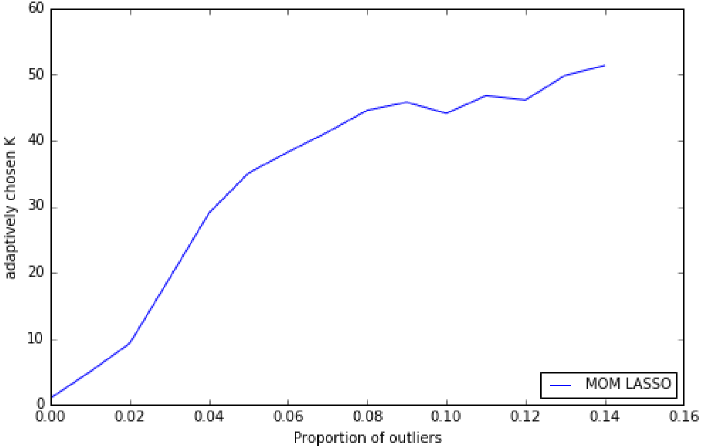
$$(K, \lambda) \in \mathcal{G}_K \times \mathcal{G}_\lambda \to \mathrm{MomCv}_V(K, \lambda) = \mathrm{Median} \left( \mathrm{MOM}_{K'}^{(v)} \left( \ell_{\hat{f}^{(v)}_{K,\lambda}} \right)_{v \in [V]} \right),$$

# Choice of hyper-parameters via MOM CV

**Idea:** The dataset may be corrupted by outliers therefore the Classical CV criteria cannot be trusted to choose hyper-parameters.

1. split the dataset into $V$ disjoints blocks $\mathcal{D}_1, \dots, \mathcal{D}_V$
2. $\forall v \in [V]$, $\cup_{u \neq v} \mathcal{D}_u$ is used to train a family of estimators
$$\mathcal{F}^{(v)} := \left( \hat{f}^{(v)}_{K,\lambda} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda \right). \tag{1}$$
3. The remaining $\mathcal{D}_v$ of the dataset is used to test the performance of each estimator in $\mathcal{F}^{(v)}$
4. $B^{(v)}_1 \cup \dots \cup B^{(v)}_{K'}$ is a partition of the test set $\mathcal{D}_v$ into $K'$ blocks
5. for all $v \in [V]$ and $f \in \mathcal{F}^{(v)}$,
$$\mathrm{MOM}^{(v)}_{K'}(\ell_f) = \mathrm{Median}\left( P_{B^{(v)}_1} \ell_f, \cdots, P_{B^{(v)}_{K'}} \ell_f \right) \tag{2}$$
6. $(\hat{K}, \hat{\lambda})$ minimizes the $\mathrm{MomCv}_V$ criteria
$$(K, \lambda) \in \mathcal{G}_K \times \mathcal{G}_\lambda \to \mathrm{MomCv}_V(K, \lambda) = \mathrm{Median}\left( \mathrm{MOM}^{(v)}_{K'}\left( \ell_{\hat{f}^{(v)}_{K,\lambda}} \right)_{v \in [V]} \right),$$
7. return $\hat{f}_{\hat{K},\hat{\lambda}}$.

# Adaptively chosen number of blocks $K$




$\hat{K}$ increases with $|\mathcal{O}|/N$ because we need at least $K \geq 2|\mathcal{O}|$ to make MOM estimators working.

# An outliers detection algorithm (random blocks)

**Idea:** Outliers should not be selected in the median blocks along the iterations.

# An outliers detection algorithm (random blocks)

**Idea:** Outliers should not be selected in the median blocks along the iterations.

**Definition**

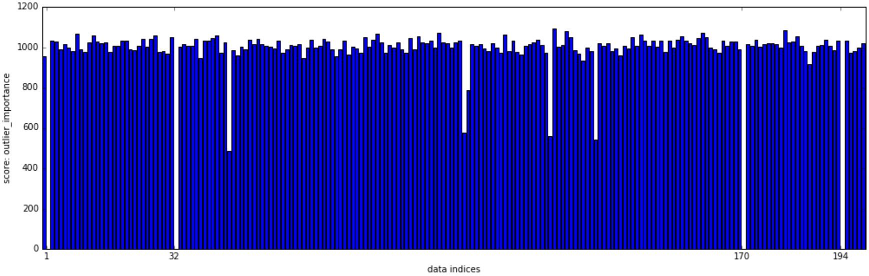
For all $i = 1, \ldots, N$, $\mathbf{Score}((X_i, Y_i))$ = number of times $(X_i, Y_i)$ has been selected in a median block along the iterations.

# An outliers detection algorithm (random blocks)

**Idea:** Outliers should not be selected in the median blocks along the iterations.

**Definition**

For all $i = 1, \ldots, N$, $\textbf{Score}((X_i, Y_i))$ = number of times $(X_i, Y_i)$ has been selected in a median block along the iterations.




outliers are data number 1, 32, 170, 194.

Thanks!

# Alternating sub-gradient descent

**input** : $(t_0, t_0') \in \mathbb{R}^d \times \mathbb{R}^d$ : initial point
$(\eta_p)_p, (\beta_p)_p$: two step size sequences
**output**: approximated solution to the min-max problem

**for** $t = 1, \ldots, T$ **do**
  find $k \in [K]$ such that $MOM_K(\ell_{t_p} - \ell_{t_p'}) = P_{B_k}(\ell_{t_p} - \ell_{t_p'})$

$$t_{p+1} = t_p + 2\eta_p \mathbb{X}_k^\top (\mathbb{Y}_k - \mathbb{X}_k t_p) - \lambda \eta_p \mathrm{sign}(t_p)$$

  find $k \in [K]$ such that $MOM_K(\ell_{t_{p+1}} - \ell_{t_p'}) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t_p'})$

$$t_{p+1}' = t_p' + 2\beta_p \mathbb{X}_k^\top (\mathbb{Y}_k - \mathbb{X}_k t_p') - \lambda \beta_p \mathrm{sign}(t_p')$$

**end**
**Return** $(t_p, t_p')$

# Alternating proximal gradient descent

**input** : $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$ : initial point
$(\eta_k)_k, (\beta_k)_k$: two step size sequences
**output**: approximated solution to the min-max problem

**for** $t = 1, \dots, T$ **do**

  find $k \in [K]$ such that $MOM_K(\ell_{t_p} - \ell_{t'_p}) = P_{B_k}(\ell_{t_p} - \ell_{t'_p})$

$$t_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\left(t_p + 2\eta_k \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t_p)\right)$$

  find $k \in [K]$ such that $MOM_K(\ell_{t_{p+1}} - \ell_{t'_p}) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p})$

$$t'_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\left(t'_p + 2\beta_k \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t'_p)\right)$$

**end**

# MOM ADMM

**input** : $(t_0, t_0') \in \mathbb{R}^d \times \mathbb{R}^d$ : initial point. $\rho$: a parameter
**output**: approximated solution to the min-max problem

**for** $t = 1, \dots, T$ **do**

find $k \in [K]$ such that $MOM_K(\ell_{t_p} - \ell_{t_p'}) = P_{B_k}(\ell_{t_p} - \ell_{t_p'})$

$$
\begin{aligned}
t_{p+1} &= (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d \times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z_p - u_p)\\
z_{p+1} &= \operatorname{prox}_{\lambda \|\cdot\|_1}(t_{p+1} + u_p/\rho)\\
u_{p+1} &= u_p + \rho(t_{p+1} - z_{p+1})
\end{aligned}
$$

find $k \in [K]$ such that $MOM_K(\ell_{t_{p+1}} - \ell_{t_p'}) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t_p'})$

$$
\begin{aligned}
t'_{p+1} &= (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d \times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z'_p - u'_p)\\
z'_{p+1} &= \operatorname{prox}_{\lambda \|\cdot\|_1}(t'_{p+1} + u'_p/\rho)\\
u'_{p+1} &= u'_p + \rho(t'_{p+1} - z'_{p+1})
\end{aligned}
$$

**end**

**Return** $(t_p, t_p')$