

# La moyenne empirique est-elle perfectible ?

Olivier Catoni

CNRS

INRIA - CLASSIC

Département de Mathématiques et Applications,

École Normale Supérieure

45 rue d'Ulm,

75 230 Paris Cedex 05,

`Olivier.Catoni@ens.fr`

Université de Lille 1,

*Vendredi 21 janvier 2011*

# Un rival pour la moyenne empirique

- Critère : **déviations non asymptotiques** des estimateurs.
- Une alternative à la moyenne empirique :

Considérons une **fonction d'influence** croissante (au sens large)

$\psi : \mathbb{R} \rightarrow \mathbb{R}$ , vérifiant

$$-\log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2).$$

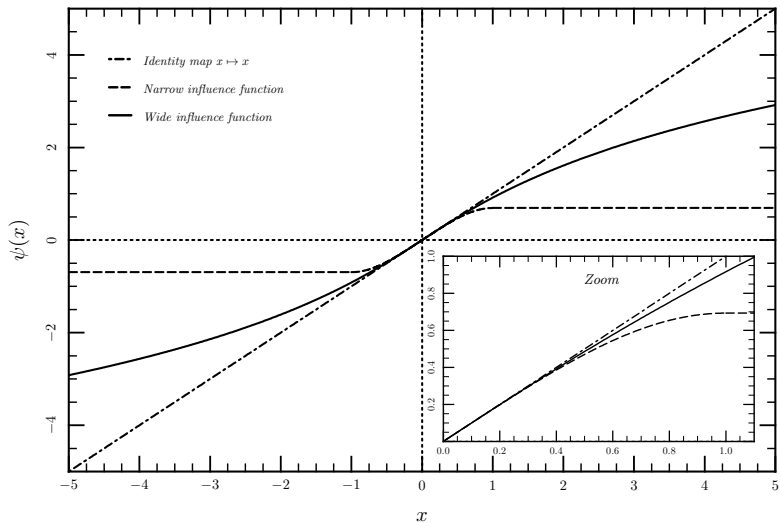
Troncature la plus faible :

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2), & x \geq 0, \\ -\log(1 - x + x^2/2), & x \leq 0. \end{cases}$$

Troncature la plus forte :

$$\psi(x) = \begin{cases} \log(2), & x \geq 1, \\ -\log(1 - x + x^2/2), & 0 \leq x \leq 1, \\ \log(1 + x + x^2/2), & -1 \leq x \leq 0, \\ -\log(2), & x \leq -1. \end{cases}$$

Plot of  $x \mapsto \psi(x)$



# Un estimateur tronqué

Étant donné un échantillon i.i.d.  $Y_1, \dots, Y_n \in \mathbb{R}$  de loi inconnue de moyenne  $m$  et de variance inférieure à  $v$ , considérons l'estimateur de la moyenne  $\hat{\theta}_\alpha$  défini par

$$\sum_{i=1}^n \psi[\alpha(Y_i - \hat{\theta}_\alpha)] = 0.$$

## Proposition

Fixons  $\epsilon \in ]\exp(-n/2), 1[$  et considérons

$$\xi = \frac{2 \log(\epsilon^{-1})}{n}, \quad \eta = \sqrt{\frac{\xi v}{1 - \xi}}, \quad \alpha = \sqrt{\frac{\xi}{v + \eta^2}}.$$

Avec probabilité au moins  $1 - 2\epsilon$ ,

$$|m - \hat{\theta}_\alpha| \leq \eta \simeq \sqrt{\frac{2v \log(\epsilon^{-1})}{n}}.$$

## Comparaison avec la moyenne empirique

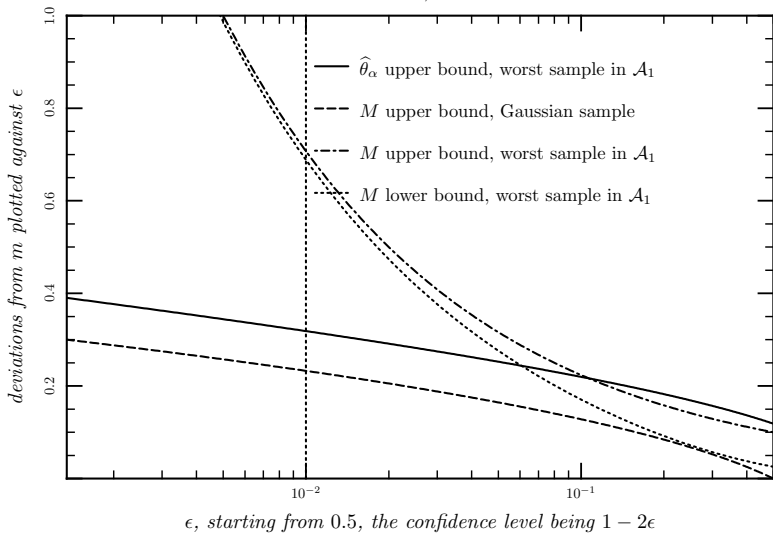
$$\text{Soit } \hat{m} = \frac{1}{n} \sum_{i=1}^n Y_i,$$

$$\mathbb{P}\left(|m - \hat{m}| \geq \sqrt{\frac{v}{2n\epsilon}}\right) \leq 2\epsilon.$$

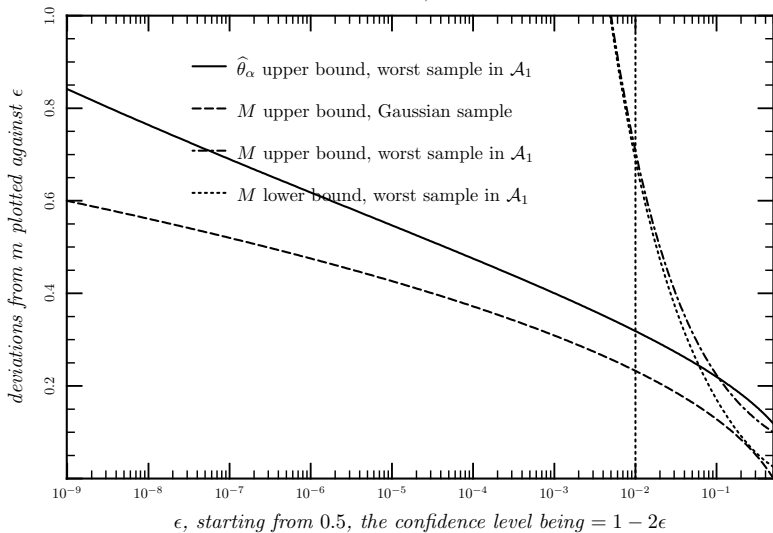
De plus il existe une distribution de variance  $v$  pour laquelle avec probabilité au moins  $2\epsilon$ ,

$$|m - \hat{m}| \geq \sqrt{\frac{v}{2n\epsilon}} \left(1 - \frac{2e\epsilon}{n}\right)^{(n-1)/2}.$$

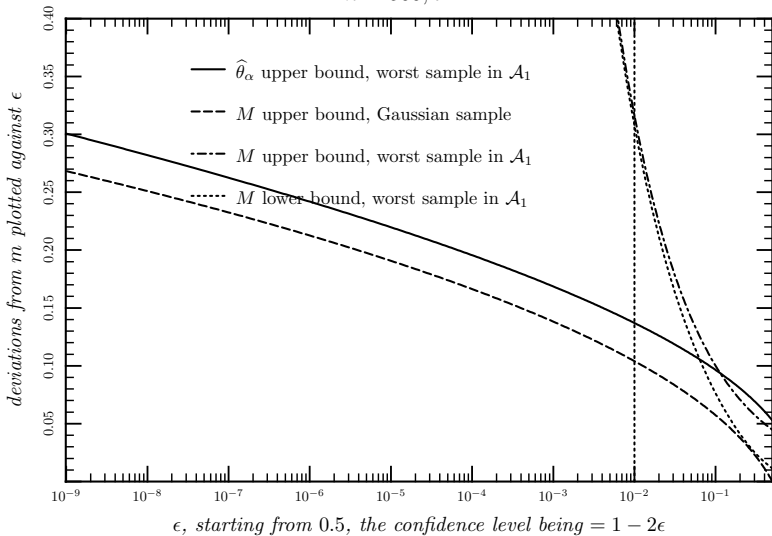
$n = 100, v = 1$



$n = 100, v = 1$



$n = 500, v = 1$

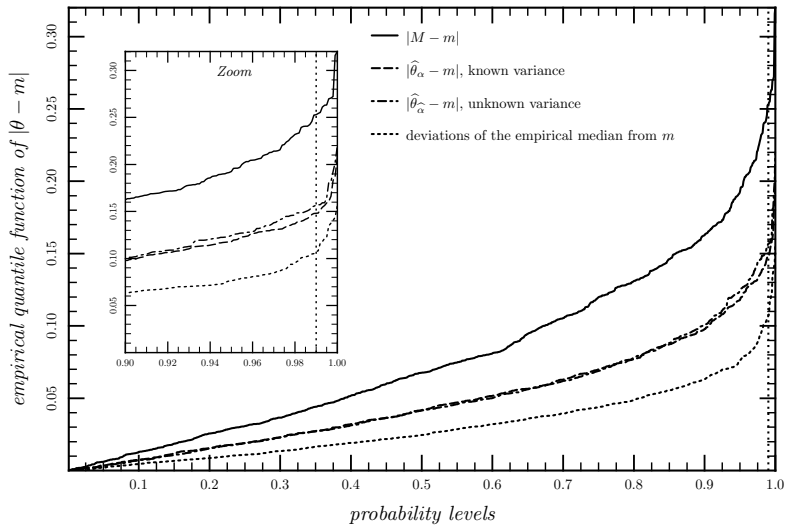


## Quelques expériences

- Loi de  $Y$  : mélange de deux Gaussiennes :  
 $0.99 \mathcal{N}(0, 1) + 0.01 \mathcal{N}(0, 30^2)$  ;
- variance  $\nu = 9.99$  ;
- kurtosis :  $\kappa = 243, 5$  ;
- taille de l'échantillon  $n = 1000$  ;
- niveau  $1 - 2\epsilon = 1 - 10/n = 0.99$  ;
- Quand la variance est inconnue, on la remplace par

$$\hat{V} = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \hat{m})^2.$$

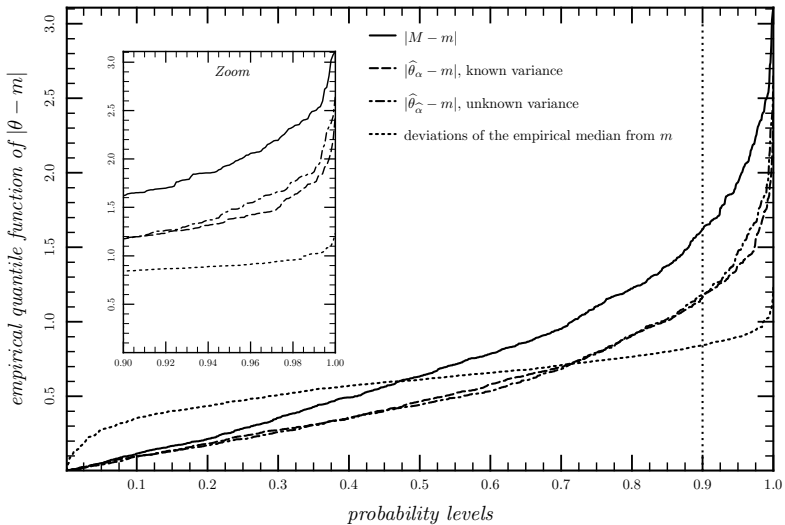
sample size  $n = 1000$ , number of experiments : 1000



- Loi de  $Y$  : mélange de trois Gaussiennes :  
 $0.7 \mathcal{N}(2, 1) + 0.2 \mathcal{N}(-2, 1) + 0.1 \mathcal{N}(0, 30^2)$  ;
- variance  $v = 93.5$  ;
- kurtosis :  $\kappa = 27, 86$  ;
- taille de l'échantillon  $n = 100$  ;
- niveau  $1 - 2\epsilon = 1 - 10/n = 0.90$  ;
- Quand la variance est inconnue, on la remplace par

$$\hat{V} = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \hat{m})^2.$$

sample size  $n = 100$ , number of experiments : 1000



# Estimation de la variance

- Supposons pour simplifier que  $n = km$  ;
- considérons la partition  $\{1, \dots, n\} = \bigsqcup_{\ell=1}^m I_{\ell}$ , où  $I_{\ell} = \{i \in \mathbb{N}; (\ell - 1)k < i \leq \ell k\}$
- et posons

$$Q_{\delta}(\beta) = \frac{1}{m} \sum_{\ell=1}^m \psi \left( \frac{1}{k(k-1)} \sum_{\substack{i,j \in I_{\ell} \\ i < j}} [\beta(Y_i - Y_j)^2 - 2\delta] \right);$$

- supposons connu un majorant  $\kappa$  de la kurtosis :  $\mathbb{E}(Y - m)^4 \leq \kappa v^2$ .

Considérons

$$\chi = \kappa - 1 + 2/(k - 1),$$

$$y = \frac{2 \log(\epsilon^{-1})}{m},$$

$$\xi = 2y(1 + 2y),$$

$$\delta = \sqrt{\frac{2k \log(\epsilon^{-1})}{\chi m}},$$

$$Q_\delta(\hat{\beta}) = -y,$$

$$\hat{v} = \sqrt{\frac{\delta(\delta - \xi)}{\hat{\beta}}}.$$

## Proposition

*Dans le cas où*

$$\log(\epsilon^{-1}) \leq \min \left\{ \frac{m}{4(1 + \sqrt{2})}, \frac{n}{2\chi} \right\},$$

*avec probabilité au moins  $1 - 2\epsilon$ ,*

$$|\log(v) - \log(\hat{v})| \leq -\frac{1}{2} \log \left( 1 - \frac{2y(1 + 2y)}{\delta} \right) \simeq \frac{y}{\delta} \simeq \sqrt{\frac{2\chi \log(\epsilon^{-1})}{n}}.$$

De plus en choisissant  $k = \sqrt{\frac{n}{4 \log(\epsilon^{-1})(\kappa - 1)}}$ , on obtient, avec probabilité au moins  $1 - 2\epsilon$

$$\begin{aligned} & |\log(v) - \log(\hat{v})| \\ & \leq -\frac{1}{2} \log \left[ 1 - 2\sqrt{\frac{2(\kappa - 1) \log(\epsilon^{-1})}{n}} \right] \\ & \quad \times \exp \left( \frac{13}{2} \sqrt{\frac{\log(\epsilon_1^{-1})}{(\kappa - 1)n}} \right) \\ & \quad \simeq \sqrt{\frac{2(\kappa - 1) \log(\epsilon_1^{-1})}{n}}. \end{aligned}$$

# Estimation jointe de la variance et de la moyenne

Considérons  $\hat{\theta}$  défini par

$$\sum_{i=1}^n \psi[\hat{\alpha}(Y_i - \hat{\theta})] = 0,$$

où  $\hat{\alpha}$  est un estimateur de la valeur optimale de  $\alpha$ . Soit  $\hat{v}$  un estimateur de  $v$ . Supposons qu'avec probabilité au moins  $1 - 2\epsilon_1$ ,

$$|\log(v) - \log(\hat{v})| \leq \zeta.$$

Choisissons

$$\eta = \frac{2 \cosh(\zeta/2)^2 [\log(1 + x^{-1}) + \log(\epsilon_2^{-1})]}{n[1 - 1.81x^2 \sinh(\zeta/2)^2]}, \quad \gamma = \frac{\eta}{1 - \eta},$$

$$\alpha = \sqrt{\frac{2 [\log(1 + x^{-1}) + \log(\epsilon_2^{-1})]}{n[1 - 1.81x^2 \sinh(\zeta/2)^2] (1 + \gamma)v}}, \quad \hat{\alpha} = \alpha \sqrt{\frac{v}{\hat{v}}}.$$

## Proposition

Avec probabilité au moins  $1 - 2\epsilon_1 - 2\epsilon_2$ ,

$$|m - \hat{\theta}| \leq \sqrt{\frac{\eta v}{1 - \eta}} \leq \sqrt{\frac{\eta \hat{v}}{1 - \eta}} \exp(\zeta/2) \leq \sqrt{\frac{\eta v}{1 - \eta}} \exp(\zeta).$$

L'optimum en  $x$  est atteint aux alentours de

$$x \simeq [3.62 \log(\epsilon_2^{-1})]^{-1/3} \sinh(\zeta/2)^{-2/3}.$$

# Comparaison avec la moyenne empirique

## Proposition

Pour toute loi de kurtosis inférieure ou égale à  $\kappa$ , la moyenne empirique  $\hat{m}$  satisfait avec probabilité au moins  $1 - 2\epsilon$ ,

$$\begin{aligned} \frac{|m - \hat{m}|}{\sqrt{v}} &\leq \inf_{\lambda \in [0,1]} \sqrt{\frac{2 \log(\lambda^{-1} \epsilon^{-1})}{n}} + \frac{\sqrt{\kappa} \log(\lambda^{-1} \epsilon^{-1})}{3n} \\ &+ \left( \frac{\kappa}{2(1-\lambda)n^3\epsilon} \right)^{1/4} \left[ 1 + \frac{3(n-1)\kappa \log(\lambda^{-1} \epsilon^{-1})^2}{4^3(1+\sqrt{2})^4 n^2} \right]^{1/4} \\ &\underset{\substack{n\epsilon \rightarrow 0 \\ \log(\epsilon^{-1})/n \rightarrow 0}}{\approx} \left( \frac{\kappa}{2n^3\epsilon} \right)^{1/4}. \end{aligned}$$

Pour être plus explicite, on peut se contenter de choisir

$$\lambda = \min \left\{ \frac{1}{2}, 2^{7/4} \left( \frac{n\epsilon}{\kappa} \right)^{1/4} \sqrt{\log \left( \frac{\kappa}{2n\epsilon^5} \right)} \right\}.$$

Début de la preuve :

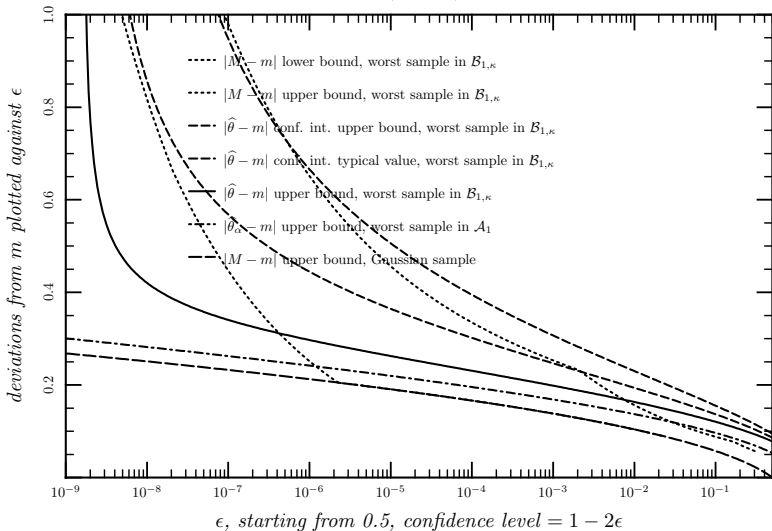
$$\hat{m} = m + \frac{1}{\alpha n} \sum_{i=1}^n \psi[\alpha(Y_i - m)] + \frac{1}{\alpha n} \sum_{i=1}^n G_i.$$

### Proposition

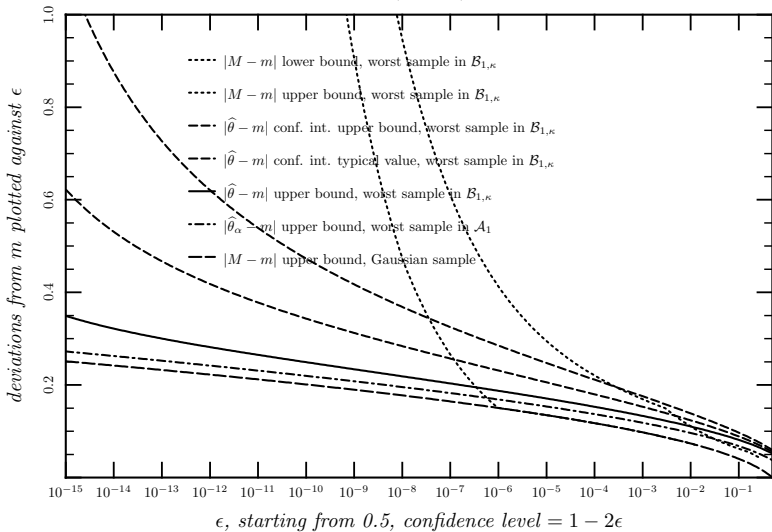
*Supposons que  $\epsilon^{-1} \geq n \geq 16$ . Il existe alors une loi de moyenne  $m$ , de variance  $v$  et de kurtosis  $\kappa$  telle que, avec probabilité au moins  $2\epsilon$ ,*

$$|m - \hat{m}| \geq \max \left\{ \left[ \frac{(\kappa - 1)(1 - 8\epsilon)}{4n\epsilon} \right]^{1/4}, \left[ \frac{(\kappa - 1)}{2n\epsilon} \left[ 1 - \left( \frac{n\epsilon}{16} \right)^{1/4} - 4\epsilon \right] \right]^{1/4} - \sqrt{\frac{\log[16/(n\epsilon)]v}{2n}} \right\} \sqrt{\frac{v}{n}}.$$

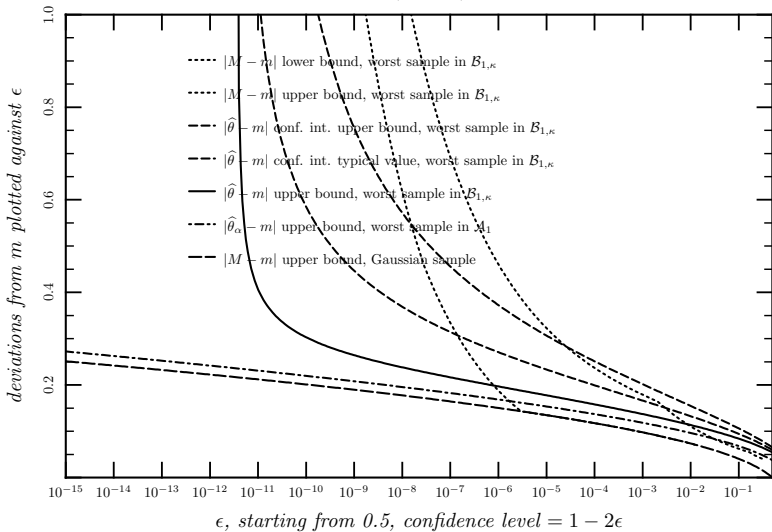
$n = 500, v = 1, \kappa = 3$



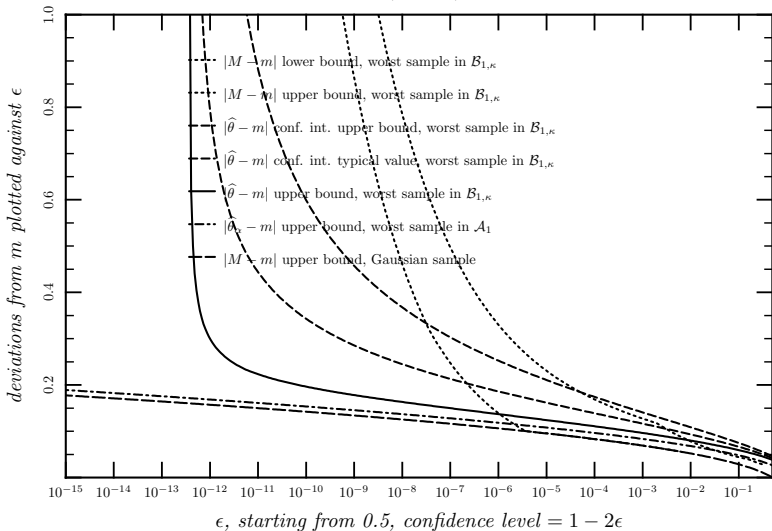
$n = 1000, v = 1, \kappa = 3$



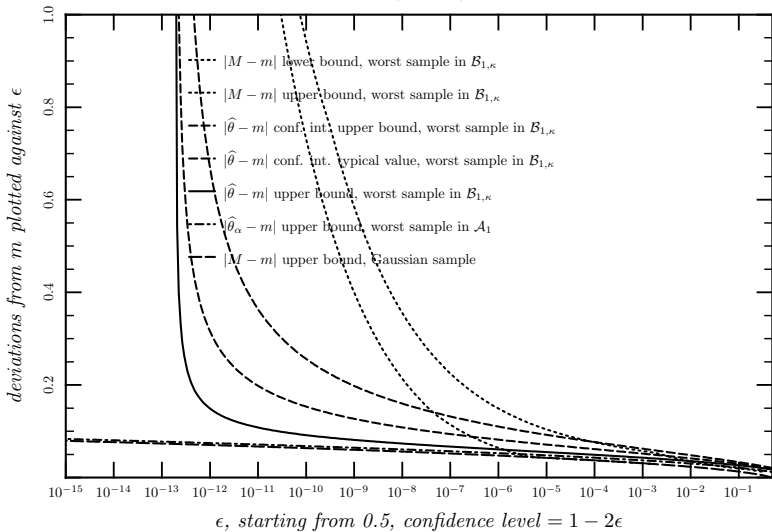
$n = 1000, v = 1, \kappa = 6$



$n = 2000, v = 1, \kappa = 13$



$n = 10000, v = 1, \kappa = 69$



Schémas itératifs :

$$\begin{aligned}\theta_0 &= \widehat{m}, & \theta_{k+1} &= \theta_k + \frac{1}{n\alpha} \sum_{i=1}^n \psi[\alpha(Y_i - \theta_k)], \\ \beta_0 &= \frac{\delta - y}{\widehat{V}}, & \beta_{k+1} &= \frac{(\delta - y)\beta_k}{\delta + Q(\beta_k)}.\end{aligned}$$



**Jean-Yves Audibert and Olivier Catoni.**

Robust linear least squares regression.

2010.



**Olivier Catoni.**

Challenging the empirical mean and empirical variance : a deviation study.

2010.