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) $\psi : \mathbb{R} \to \mathbb{R}$, vérifiant

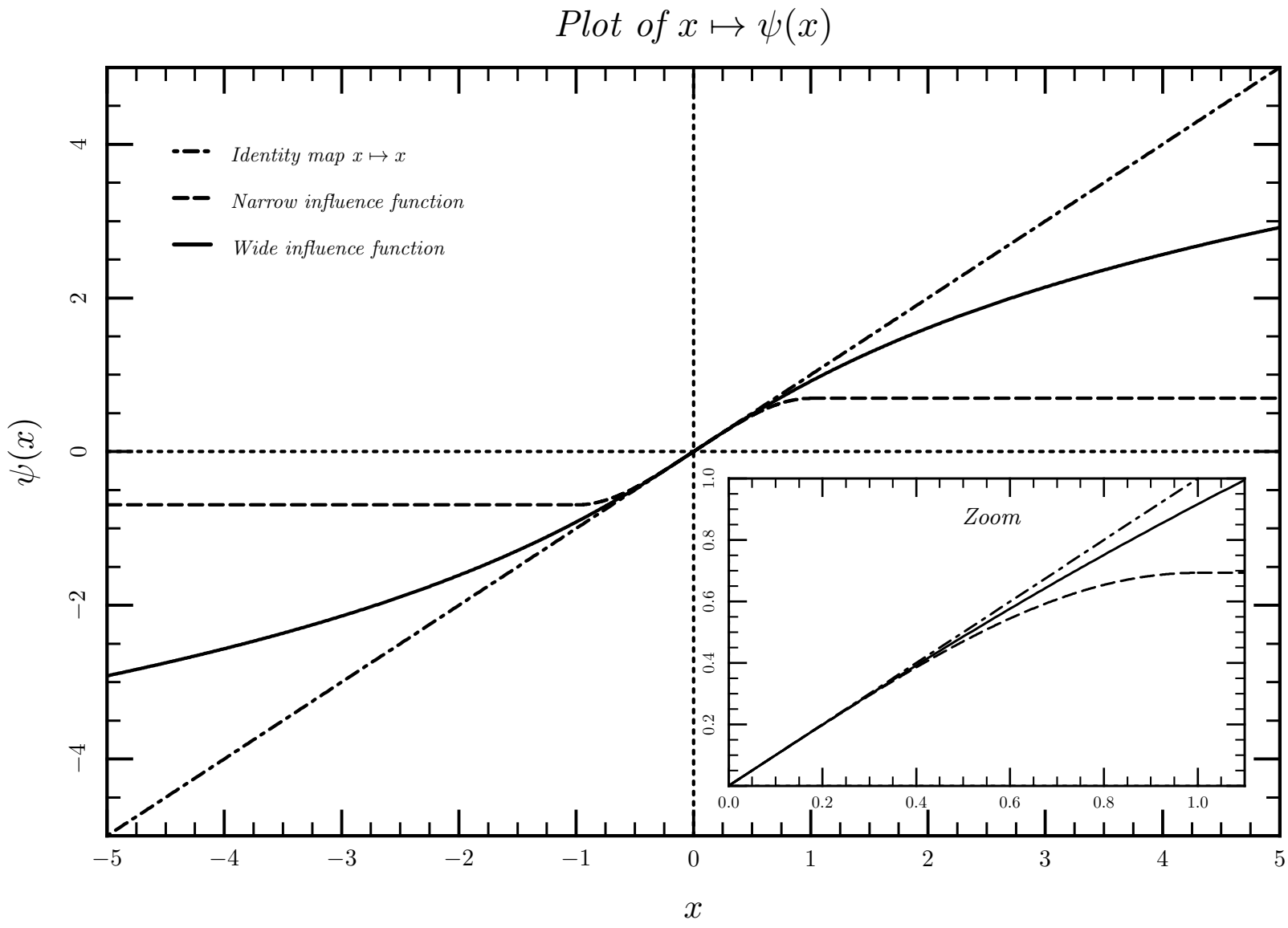
$$-\log\bigl(1 - x + x^2/2\bigr) \leq \psi(x) \leq \log\bigl(1 + x + x^2/2\bigr).$$

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 $x \mapsto \psi(x)$
Identity map $x \mapsto x$
Narrow influence function
Wide influence function
Zoom
$\psi(x)$
x
−5
−4
−3
−2
−1
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0

Un estimateur tronqué

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

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

Proposition

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

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

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

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

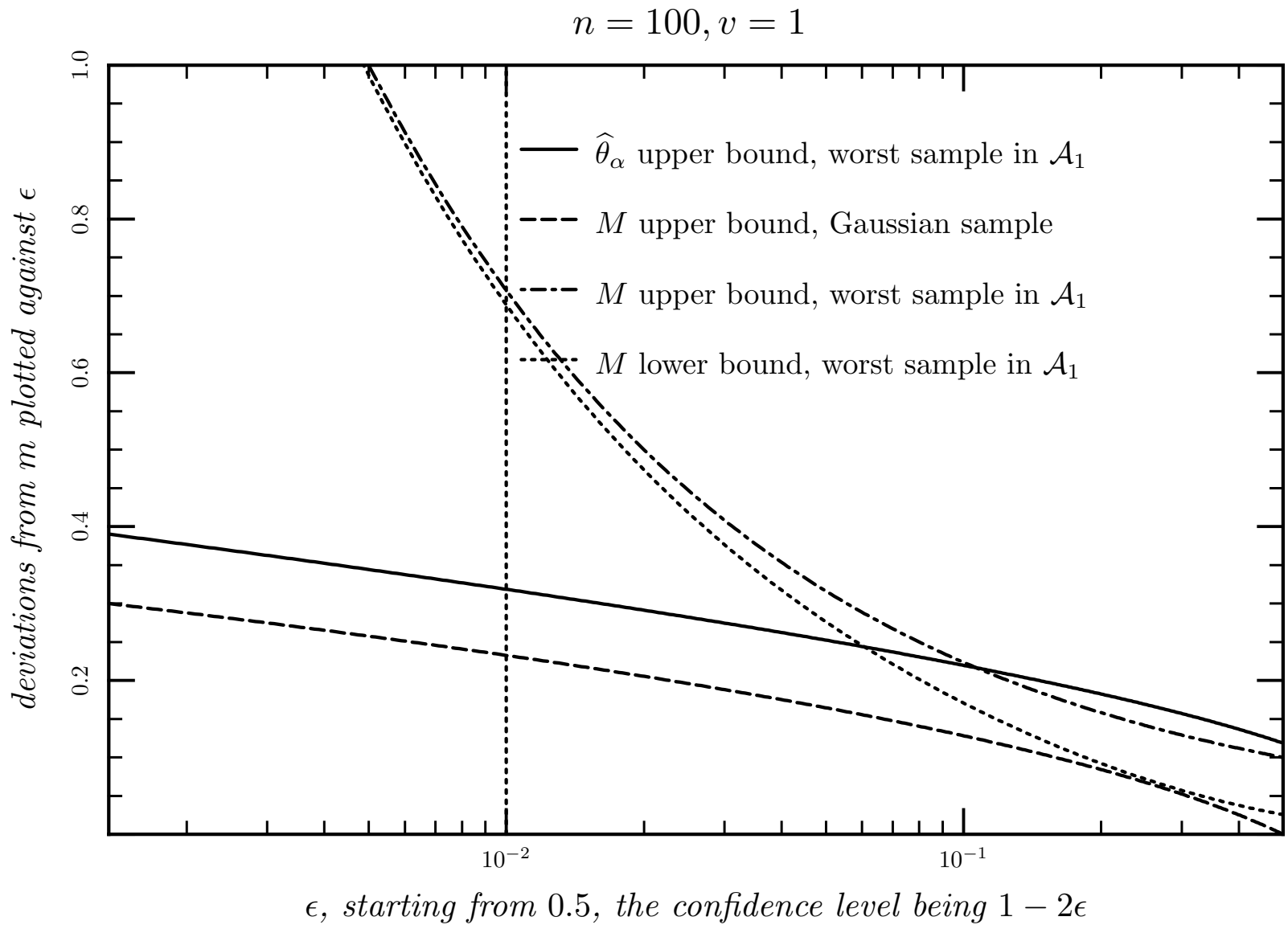
Comparaison avec la moyenne empirique

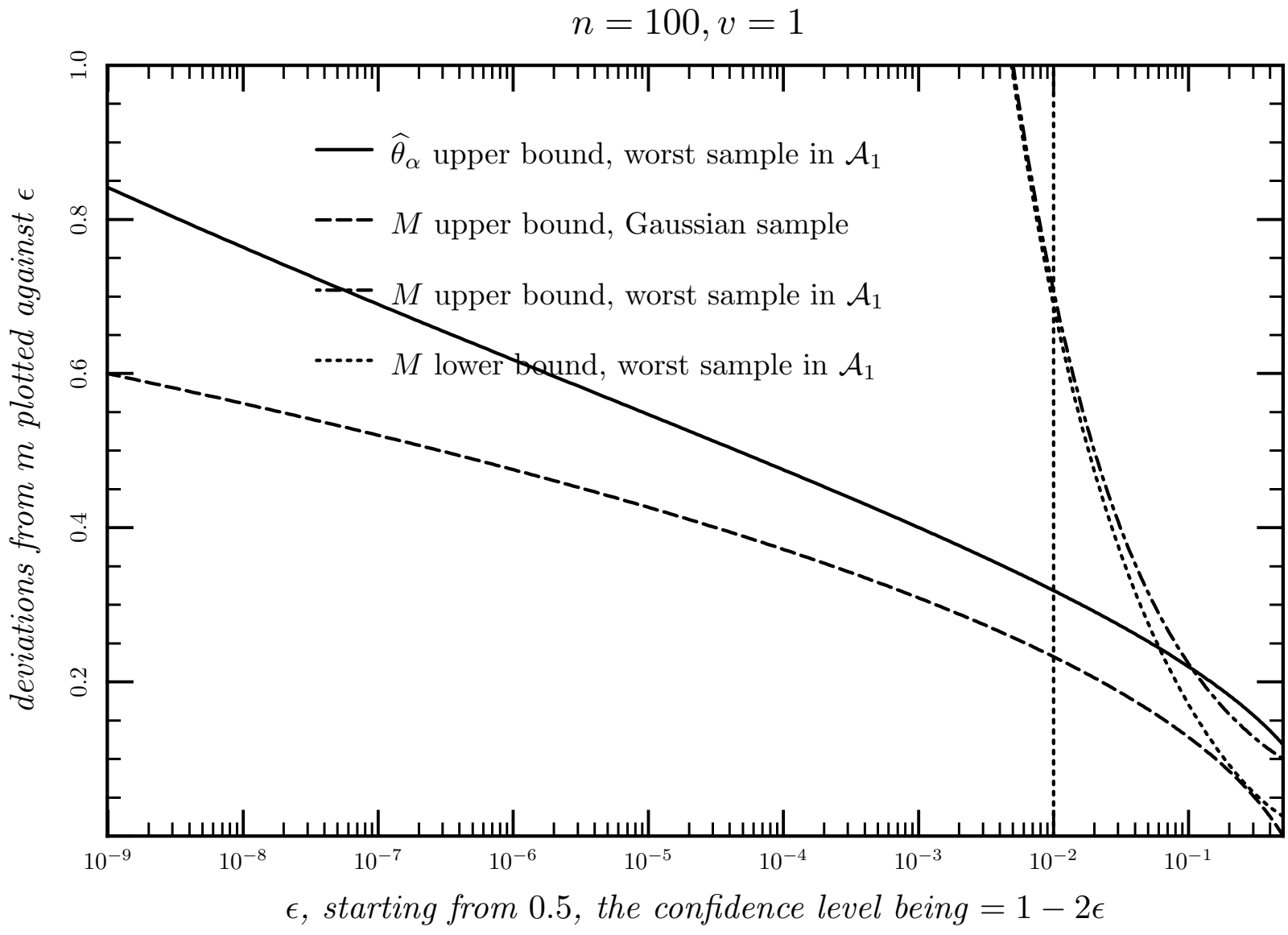
Soit $\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

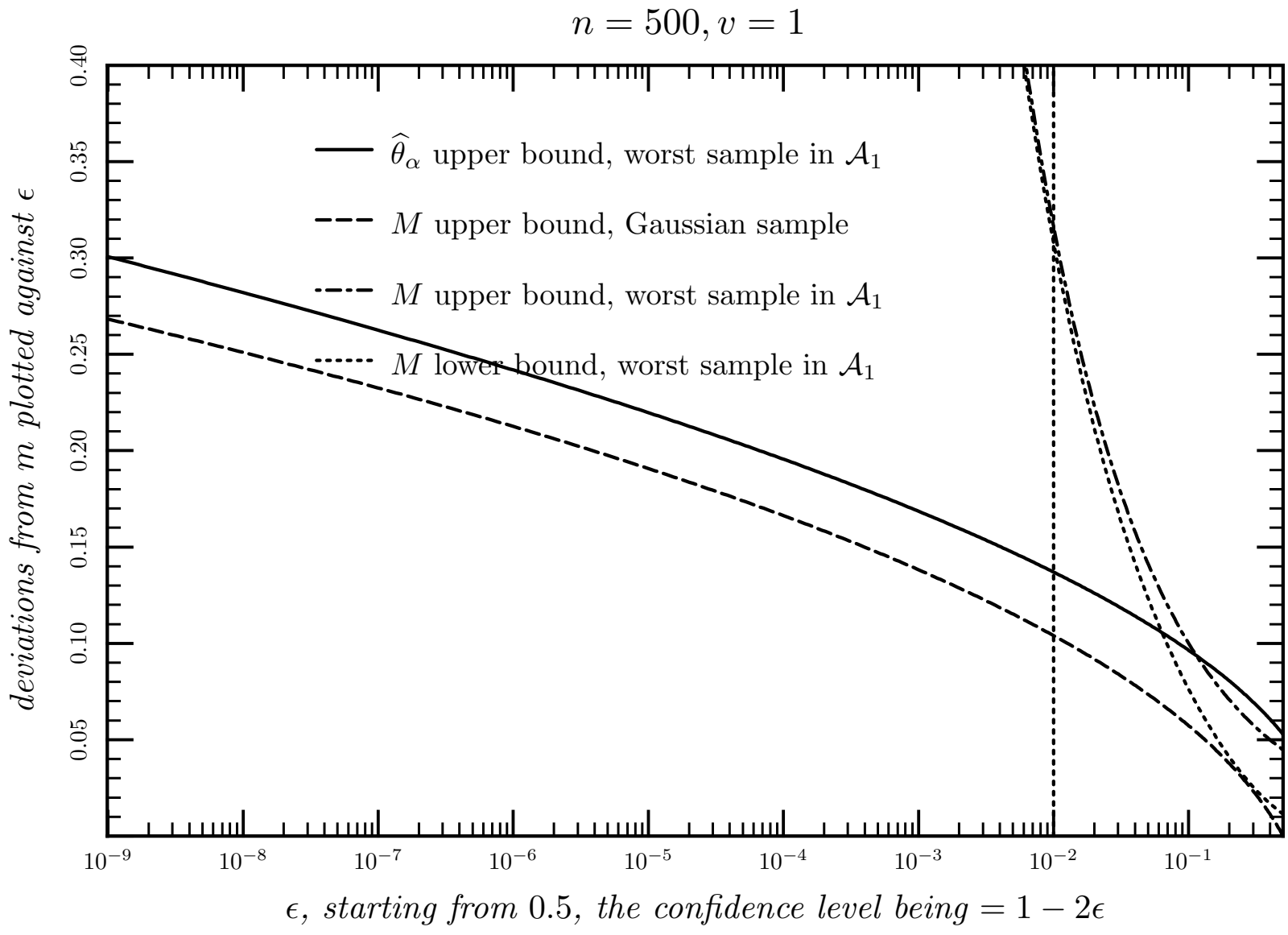
$$\mathbb{P}\left(|m - \widehat{m}| \geq \sqrt{\frac{v}{2n\epsilon}} \right) \leq 2\epsilon.$$

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

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

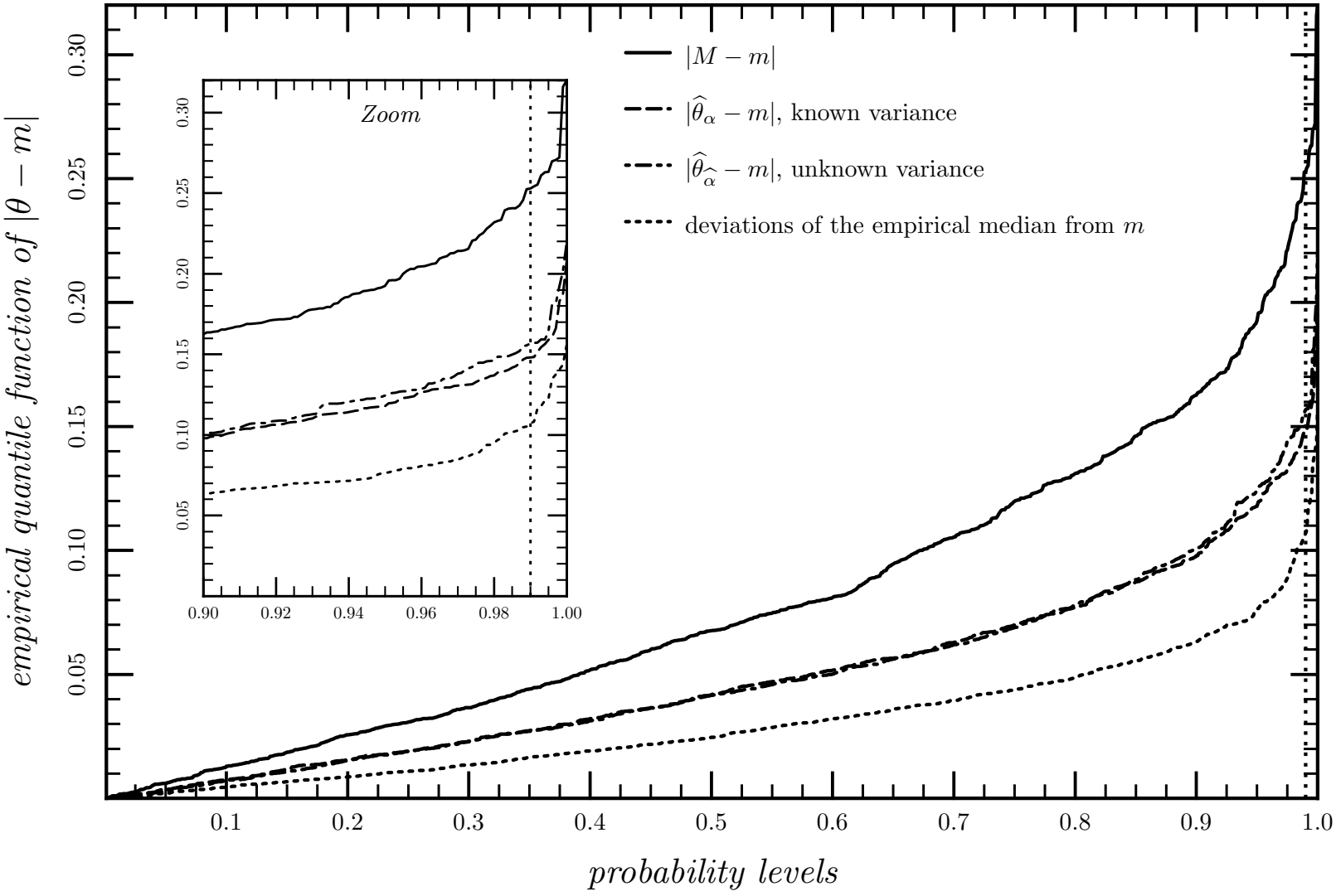

$n = 100, v = 1$
1.0
0.8
0.6
0.4
0.2
deviations from m plotted against ϵ
$\widehat{\theta}_\alpha$ upper bound, worst sample in $\mathcal{A}_1$
M upper bound, Gaussian sample
M upper bound, worst sample in $\mathcal{A}_1$
M lower bound, worst sample in $\mathcal{A}_1$
10^{-2}
10^{-1}
ϵ, starting from 0.5, the confidence level being $1 - 2\epsilon$


$n = 100, v = 1$
$\widehat{\theta}_\alpha$ upper bound, worst sample in $\mathcal{A}_1$
M upper bound, Gaussian sample
M upper bound, worst sample in $\mathcal{A}_1$
M lower bound, worst sample in $\mathcal{A}_1$
deviations from m plotted against ϵ
1.0
0.8
0.6
0.4
0.2
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
ϵ, *starting from* 0.5, *the confidence level being* $= 1 - 2\epsilon$

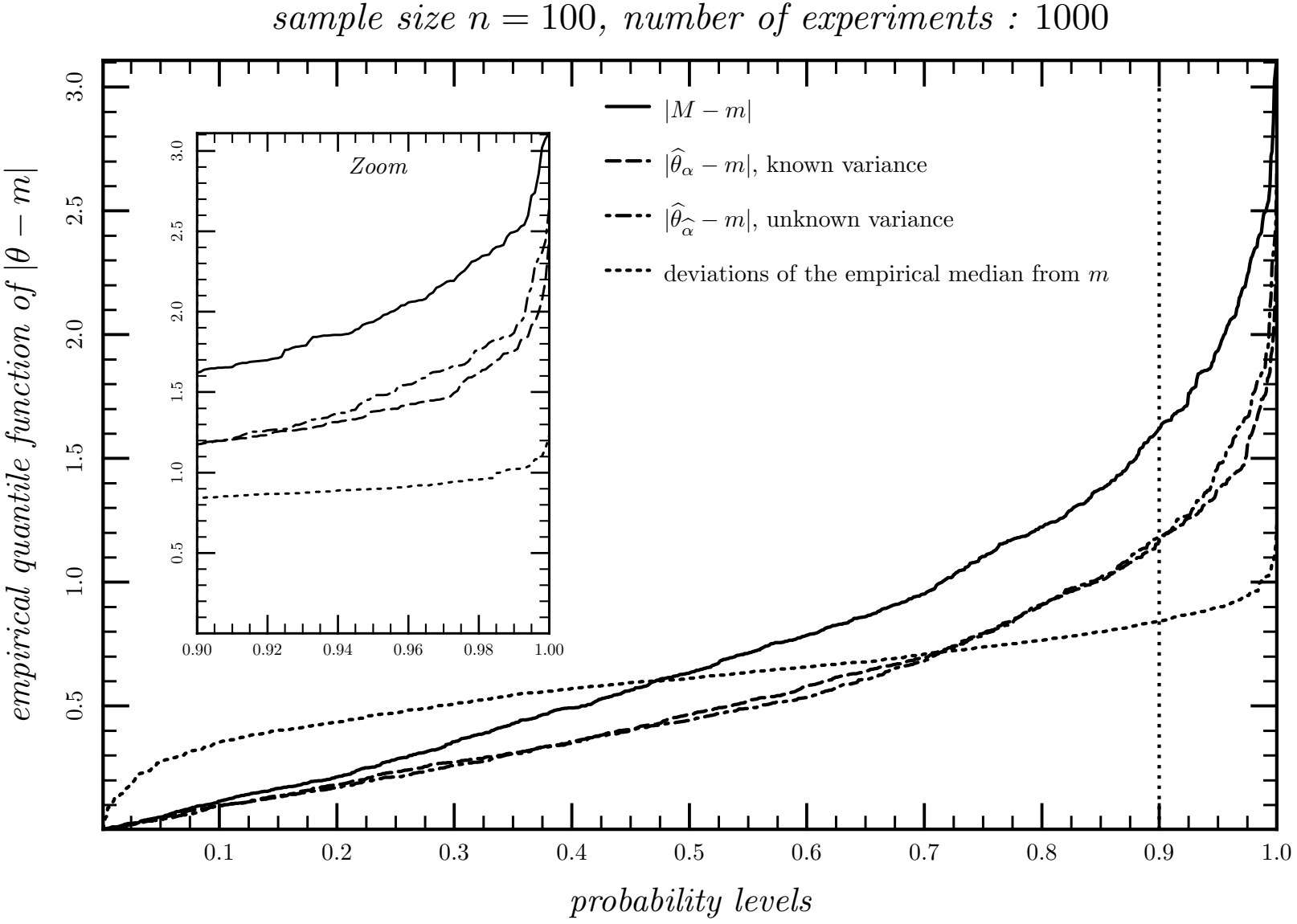

$n = 500, v = 1$
$\widehat{\theta}_\alpha$ upper bound, worst sample in $\mathcal{A}_1$
M upper bound, Gaussian sample
M upper bound, worst sample in $\mathcal{A}_1$
M lower bound, worst sample in $\mathcal{A}_1$
deviations from m plotted against ϵ
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
ϵ, *starting from* 0.5, *the confidence level being* $= 1 - 2\epsilon$

Quelques expériences

- Loi de Y : mélange de deux Gaussiennes : $0.99\ \mathcal{N}(0,1) + 0.01\ \mathcal{N}(0, 30^2)$;
- variance $v = 9.99$;
- 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 $\widehat{V} = \dfrac{1}{n-1} \sum_{i=1}^{n} (Y_i - \widehat{m})^2$.


sample size $n = 1000$, number of experiments : 1000
empirical quantile function of $|\theta - m|$
probability levels
$|M - m|$
$|\widehat{\theta}_\alpha - m|$, known variance
$|\widehat{\theta}_{\widehat{\alpha}} - m|$, unknown variance
deviations of the empirical median from m
Zoom
0.05
0.10
0.15
0.20
0.25
0.30
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.90
0.92
0.94
0.96
0.98
1.00

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


sample size $n = 100$, number of experiments : 1000
empirical quantile function of $|\theta - m|$
probability levels
$|M - m|$
$|\widehat{\theta}_\alpha - m|$, known variance
$|\widehat{\theta}_{\widehat{\alpha}} - m|$, unknown variance
deviations of the empirical median from m
Zoom
0.5
1.0
1.5
2.0
2.5
3.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.90
0.92
0.94
0.96
0.98
1.00

Estimation de la variance

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

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

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

Considérons

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

$$y = \frac{2\log(\epsilon^{-1})}{m}, \qquad Q_\delta(\widehat{\beta}) = -y,$$

$$\xi = 2y(1+2y), \qquad \widehat{v} = \sqrt{\frac{\delta(\delta-\xi)}{\widehat{\beta}}}.$$

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 $1-2\epsilon$,

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

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

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

Estimation jointe de la variance et de la moyenne

Considérons $\widehat{\theta}$ défini par

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

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

$$\left|\log(v) - \log(\widehat{v})\right| \leq \zeta.$$

Choisissons

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

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

Proposition

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

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

L'optimum en x est atteint aux alentours de

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

Comparaison avec la moyenne empirique

Proposition

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

$$\frac{|m-\widehat{m}|}{\sqrt{v}} \leq \inf_{\lambda\in[0,1]} \sqrt{\frac{2\log(\lambda^{-1}\epsilon^{-1})}{n}} + \frac{\sqrt{\kappa}\log(\lambda^{-1}\epsilon^{-1})}{3n} + \left(\frac{\kappa}{2(1-\lambda)n^3\epsilon}\right)^{1/4}\left[1+\frac{3(n-1)\kappa\log(\lambda^{-1}\epsilon^{-1})^2}{4^3(1+\sqrt{2})^4n^2}\right]^{1/4}$$

$$\underset{\substack{n\epsilon\to 0\\ \log(\epsilon^{-1})/n\to 0}}{\simeq} \left(\frac{\kappa}{2n^3\epsilon}\right)^{1/4}.$$

Pour être plus explicite, on peut se contenter de choisir

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

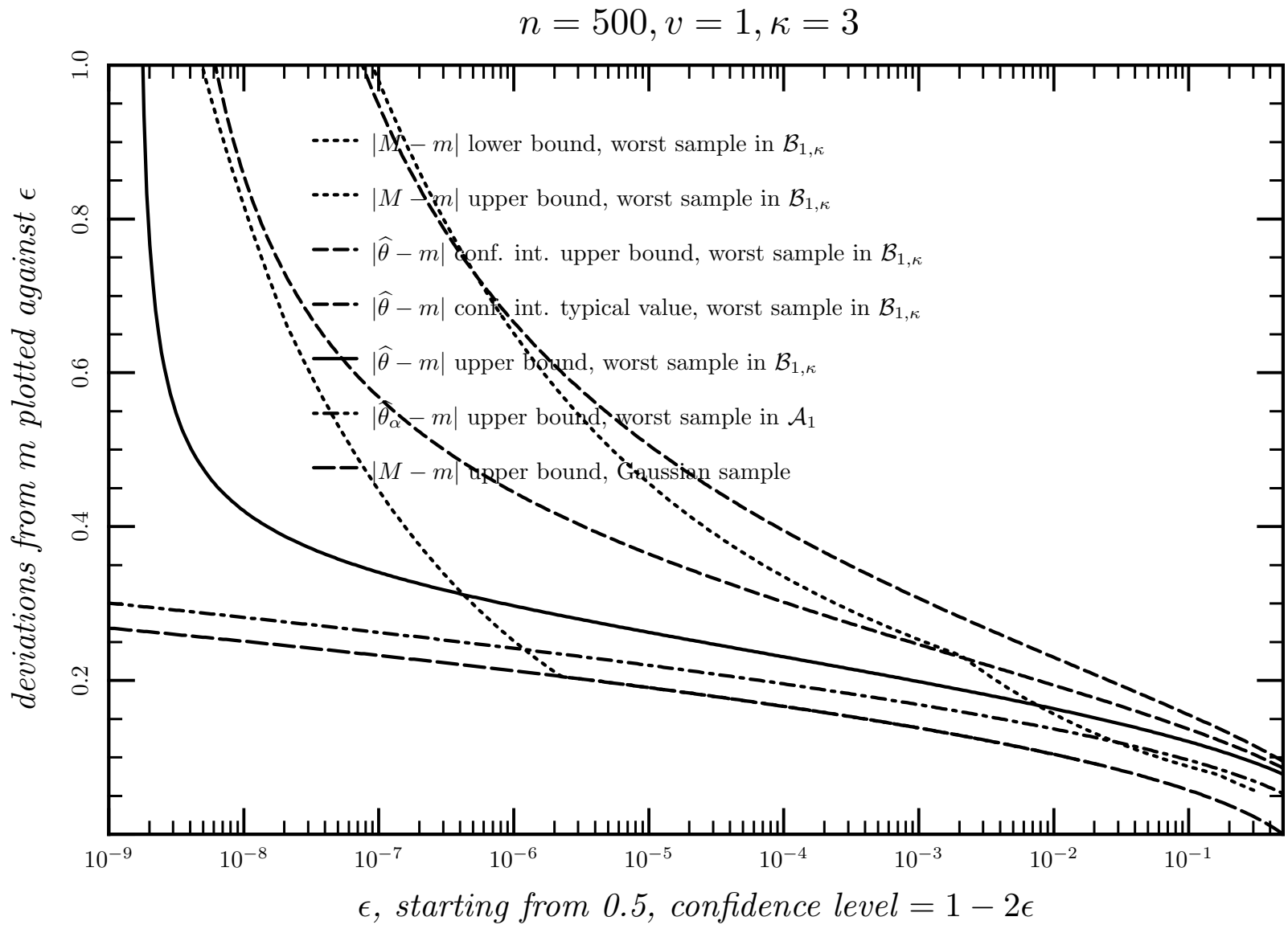
Début de la preuve :

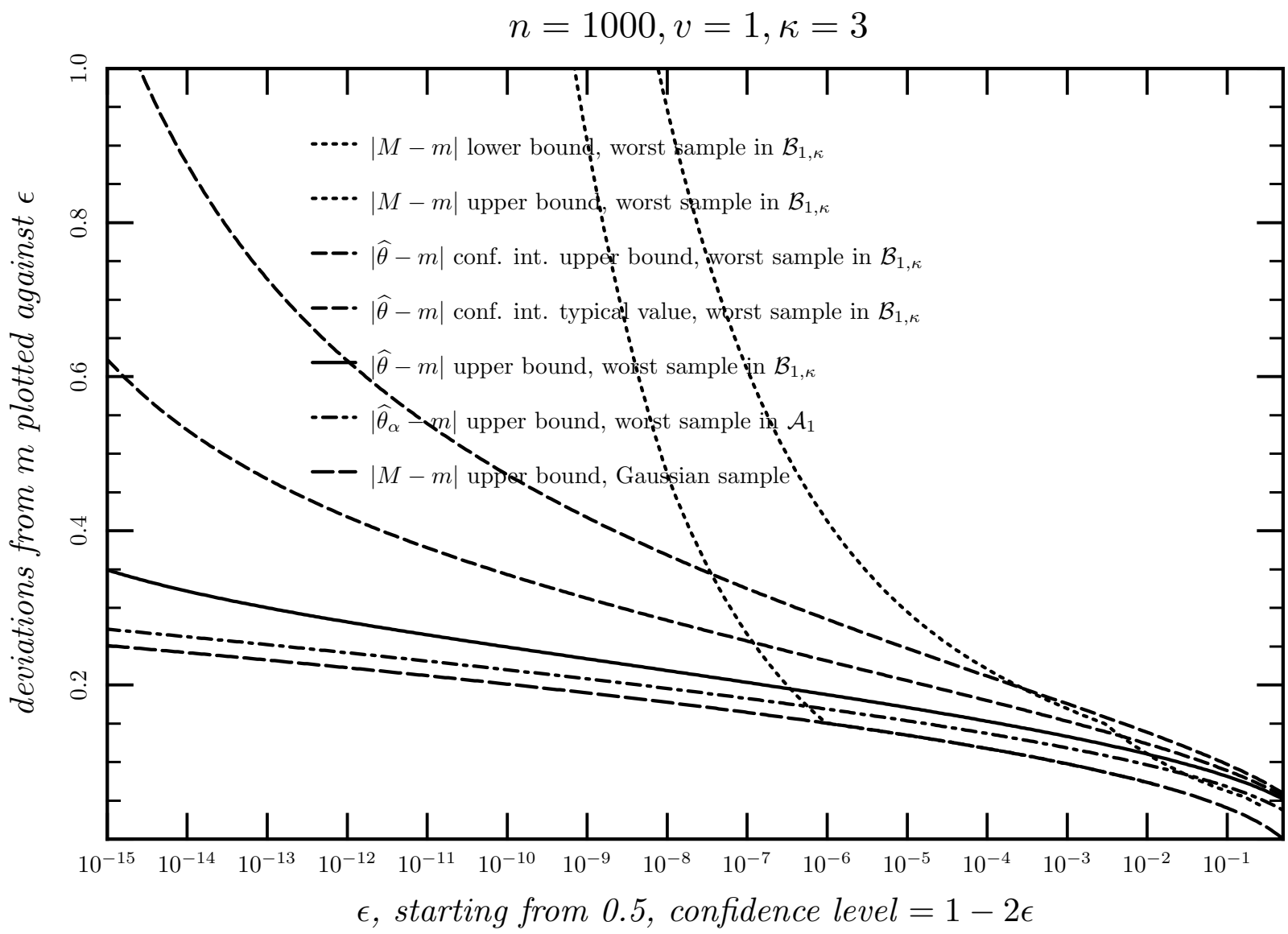
$$\widehat{m} = m + \frac{1}{\alpha n}\sum_{i=1}^{n}\psi\big[\alpha(Y_i - m)\big] + \frac{1}{\alpha n}\sum_{i=1}^{n} G_i.$$

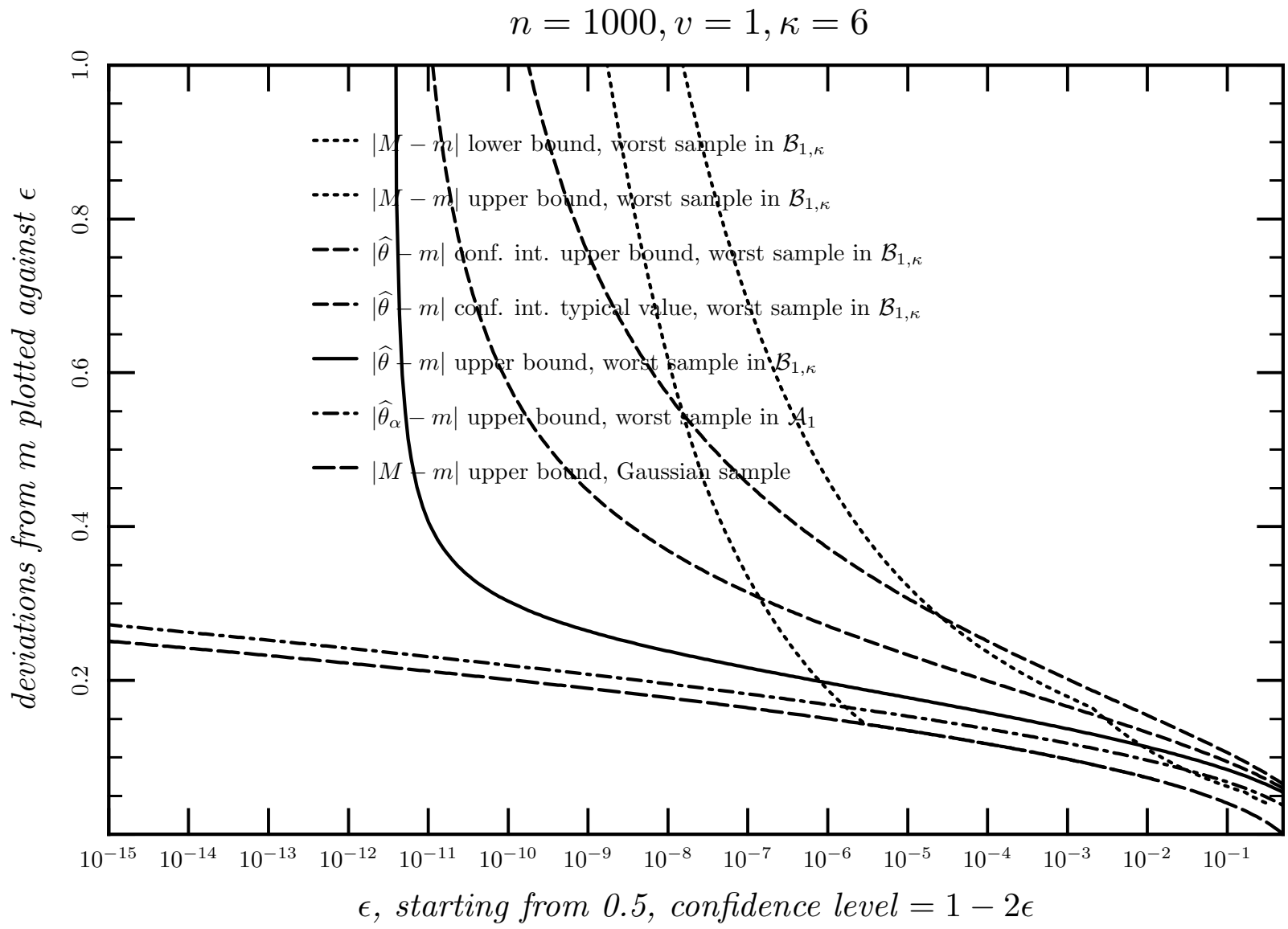
Proposition

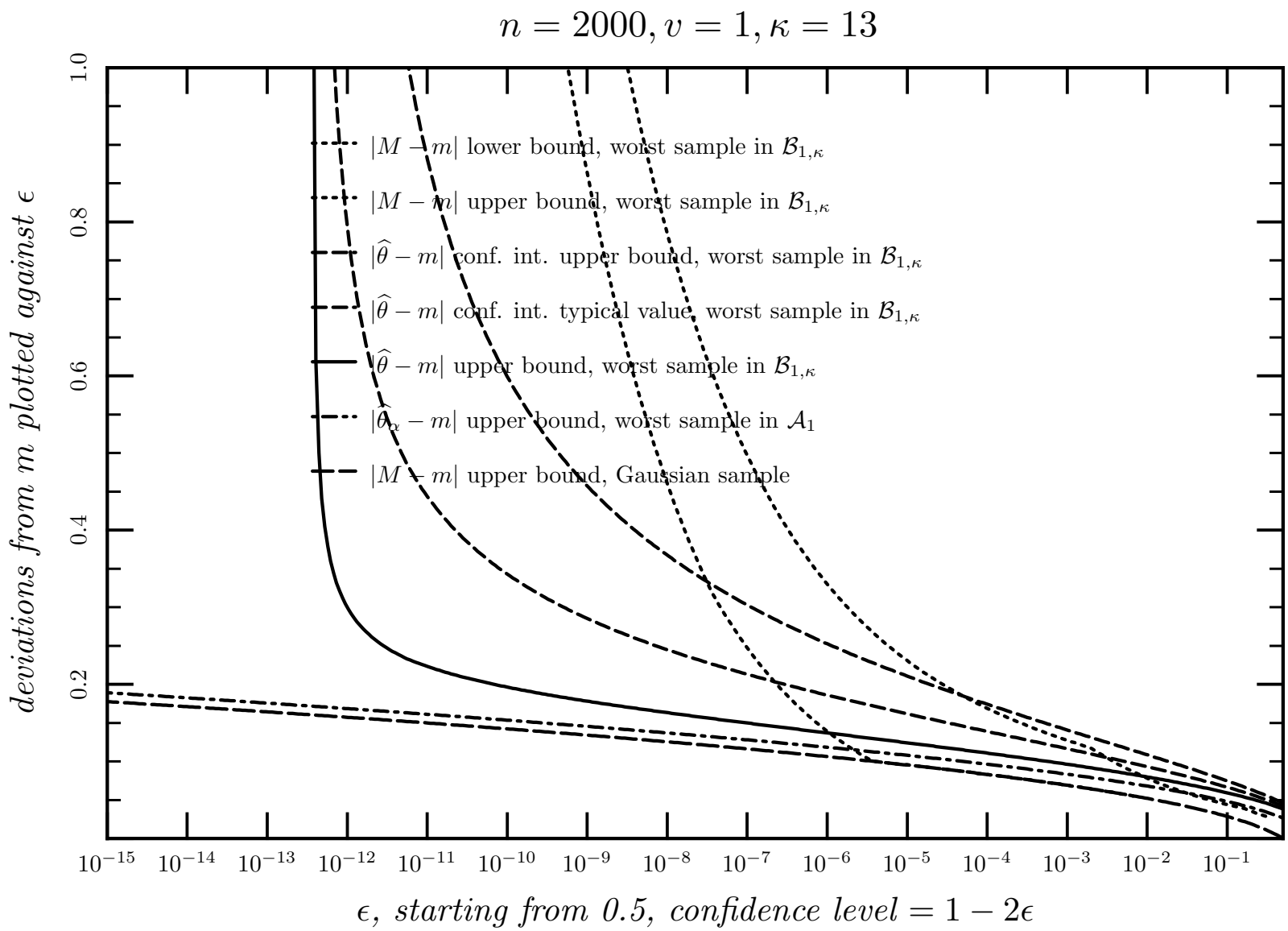
Supposons que $\epsilon^{-1} \geq n \geq 16$. *Il existe alors une loi de moyenne* m, *de variance* v *et de kurtosis* κ *telle que, avec probabilité au moins* 2ϵ,

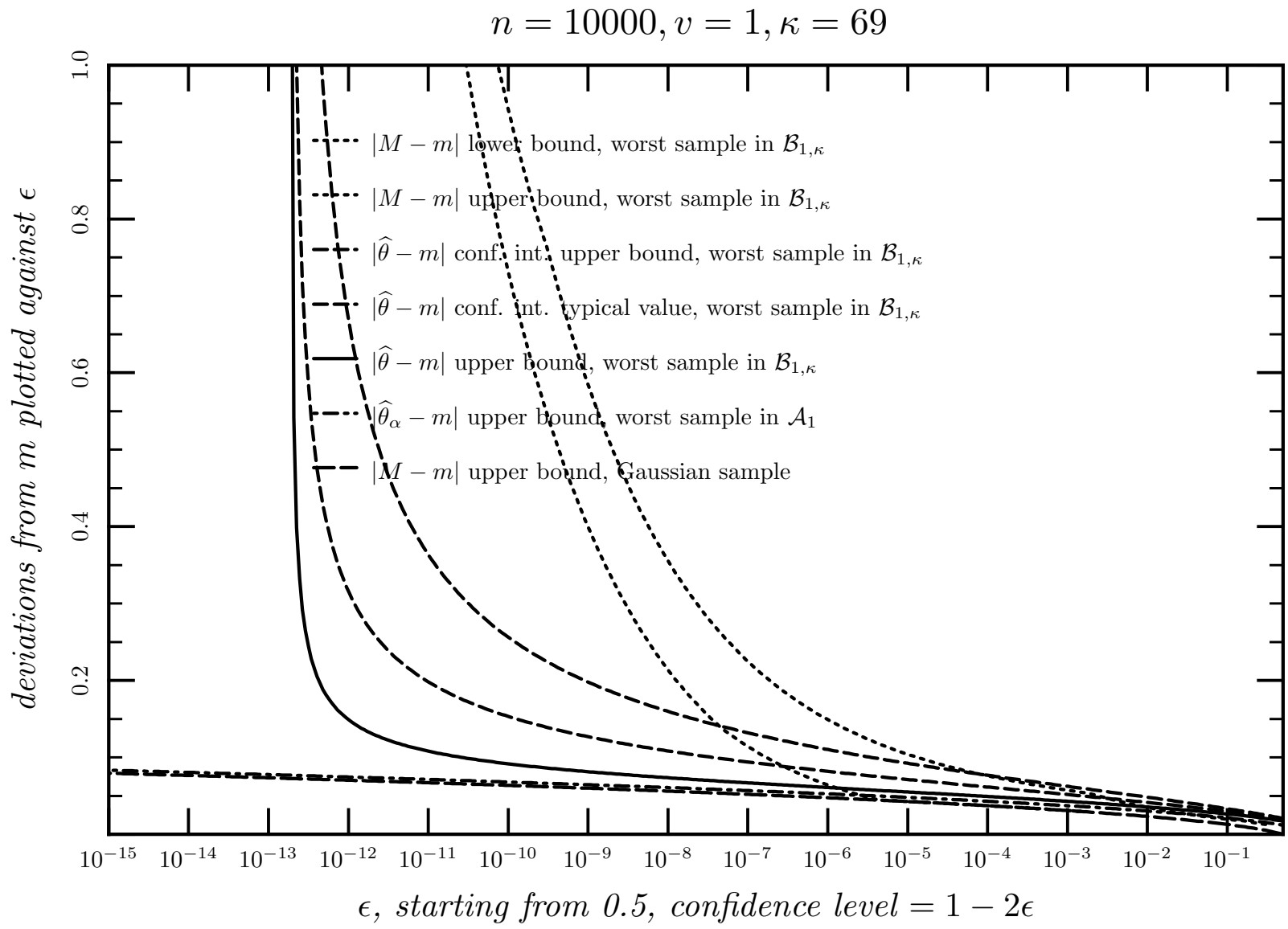
$$|m - \widehat{m}| \geq \max\Bigg\{ \left[\frac{(\kappa-1)(1-8\epsilon)}{4n\epsilon}\right]^{1/4}, \left[\frac{(\kappa-1)}{2n\epsilon}\left[1 - \left(\frac{n\epsilon}{16}\right)^{1/4} - 4\epsilon\right]\right]^{1/4} - \sqrt{\frac{\log\big[16/(n\epsilon)\big]v}{2n}}\Bigg\}\sqrt{\frac{v}{n}}.$$


$n = 500, v = 1, \kappa = 3$
$|M - m|$ lower bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|M - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. typical value, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta}_\alpha - m|$ upper bound, worst sample in $\mathcal{A}_1$
$|M - m|$ upper bound, Gaussian sample
deviations from m plotted against ϵ
1.0
0.8
0.6
0.4
0.2
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
ϵ, starting from 0.5, confidence level $= 1 - 2\epsilon$


$n = 1000, v = 1, \kappa = 3$
$|M - m|$ lower bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|M - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. typical value, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta}_\alpha - m|$ upper bound, worst sample in $\mathcal{A}_1$
$|M - m|$ upper bound, Gaussian sample
deviations from m plotted against ϵ
1.0
0.8
0.6
0.4
0.2
10^{-15} 10^{-14} 10^{-13} 10^{-12} 10^{-11} 10^{-10} 10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
ϵ, starting from 0.5, confidence level $= 1 - 2\epsilon$


$n = 1000, v = 1, \kappa = 6$
$|M - m|$ lower bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|M - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. typical value, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta}_\alpha - m|$ upper bound, worst sample in $\mathcal{A}_1$
$|M - m|$ upper bound, Gaussian sample
deviations from m plotted against ϵ
1.0
0.8
0.6
0.4
0.2
10^{-15} 10^{-14} 10^{-13} 10^{-12} 10^{-11} 10^{-10} 10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
ϵ, starting from 0.5, confidence level $= 1 - 2\epsilon$


$n = 2000, v = 1, \kappa = 13$
$|M - m|$ lower bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|M - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. typical value, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta}_\alpha - m|$ upper bound, worst sample in $\mathcal{A}_1$
$|M - m|$ upper bound, Gaussian sample
1.0
0.8
0.6
0.4
0.2
deviations from m plotted against ϵ
10^{-15} 10^{-14} 10^{-13} 10^{-12} 10^{-11} 10^{-10} 10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
ϵ, starting from 0.5, confidence level $= 1 - 2\epsilon$


$n = 10000, v = 1, \kappa = 69$
$|M - m|$ lower bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|M - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ conf. int. typical value, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta} - m|$ upper bound, worst sample in $\mathcal{B}_{1,\kappa}$
$|\widehat{\theta}_\alpha - m|$ upper bound, worst sample in $\mathcal{A}_1$
$|M - m|$ upper bound, Gaussian sample
deviations from m plotted against ϵ
1.0
0.8
0.6
0.4
0.2
10^{-15} 10^{-14} 10^{-13} 10^{-12} 10^{-11} 10^{-10} 10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}
ϵ, starting from 0.5, confidence level $= 1 - 2\epsilon$

Calcul des estimateurs

Schémas itératifs :

$$\theta_0 = \widehat{m}, \qquad \theta_{k+1} = \theta_k + \frac{1}{n\alpha}\sum_{i=1}^{n} \psi\big[\alpha(Y_i - \theta_k)\big],$$

$$\beta_0 = \frac{\delta - y}{\widehat{V}}, \qquad \beta_{k+1} = \frac{(\delta - y)\beta_k}{\delta + Q(\beta_k)}.$$


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.