Nonasymptotic bounds on the mean square error for MCMC estimates via renewal techniques[1]

Wojciech Niemiro

Nicolaus Copernicus University, Toruń and University of Warsaw
Poland

MCQMC2010
Warszawa, August 2010

[1] Joint work with Krzysztof Łatuszyński, University of Warwick, U.K. and Błażej Miasojedow, University of Warsaw, Poland.
Work partially supported by Polish Ministry of Science and Higher Education Grant No. N N201387234.

Outline

Outline

1. Introduction
   - Computing integrals via MCMC
   - Accuracy bounds
2. MSE bounds
   - Small set and regeneration
   - Mean Square Error
   - Drift condition
   - Explicit bounds
3. Confidence estimation
   - Median of Averages
   - Complexity comparison

Outline

Computing integrals via MCMC

We are to compute

$$\theta = \pi(f) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x),$$

where

- $\mathcal{X}$ – state space,
- π – probability distribution on $\mathcal{X}$,

Markov chain

$$X_0, X_1, \ldots, X_n, \ldots \qquad \mathbb{P}(X_n \in \cdot) \to \pi(\cdot) \quad (n \to \infty).$$

MCMC estimator

$$\hat{\theta}_n = \hat{\pi}_n(f) = \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \to \theta \qquad (n \to \infty).$$

Computing integrals via MCMC

We are to compute

$$\theta = \pi(f) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x),$$

where

- $\mathcal{X}$ – state space,
- π – probability distribution on $\mathcal{X}$,

Markov chain

$$X_0, X_1, \ldots, X_n, \ldots \qquad \mathbb{P}(X_n \in \cdot) \to \pi(\cdot) \quad (n \to \infty).$$

MCMC estimator

$$\hat{\theta}_n = \hat{\pi}_n(f) = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \to \theta \qquad (n \to \infty).$$

Computing integrals via MCMC

We are to compute

$$\theta = \pi(f) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x),$$

where

- $\mathcal{X}$ – state space,
- π – probability distribution on $\mathcal{X}$,

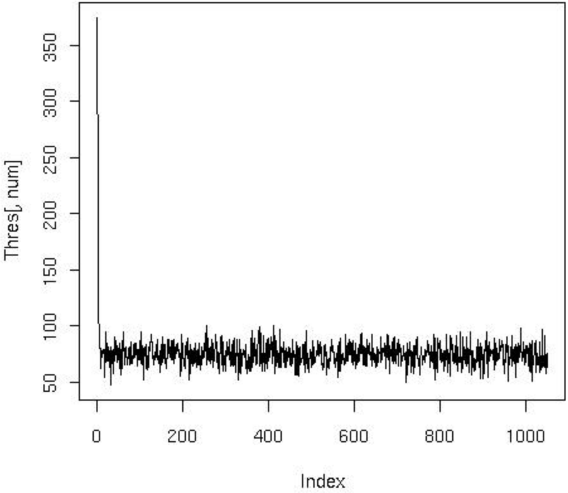
Markov chain

$$X_0, X_1, \ldots, X_n, \ldots \qquad \mathbb{P}(X_n \in \cdot) \to \pi(\cdot) \quad (n \to \infty).$$

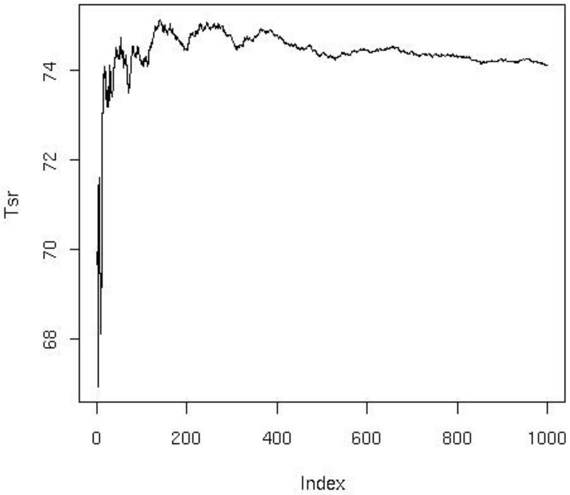
MCMC estimator

$$\hat{\theta}_n = \hat{\pi}_n(f) = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \to \theta \qquad (n \to \infty).$$

Convergence of probability distributions:



Convergence of sample averages:



Accuracy bounds

Rate of convergence of probability distributions:

$$\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\| \leq ?$$

– not considered here.

Rate of convergence of sample averages.

Mean square error:

$$\sqrt{\text{MSE}} = \sqrt{\mathbb{E}\,(\hat{\theta}_n - \theta)^2} \leq ?$$

Confidence bounds:

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq ?$$

Accuracy bounds

Rate of convergence of probability distributions:

$$\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\| \leq ?$$

– not considered here.

Rate of convergence of sample averages.

Mean square error:

$$\sqrt{\mathrm{MSE}} \;=\; \sqrt{\mathbb{E}\,(\hat{\theta}_n - \theta)^2} \leq ?$$

Confidence bounds:

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq ?$$

Accuracy bounds

Rate of convergence of probability distributions:

$$\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\| \leq ?$$

– not considered here.

Rate of convergence of sample averages.

Mean square error:

$$\sqrt{\mathrm{MSE}} \;=\; \sqrt{\mathbb{E}\,(\hat{\theta}_n - \theta)^2} \leq ?$$

Confidence bounds:

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq ?$$

Small set

Transition kernel:

$$P(x, A) = \mathbb{P}(X_n \in A | X_{n-1} = x).$$

ASSUMPTION

- ***Stationary distribution.*** *There exists a probability distribution π on $\mathcal{X}$ such that $\pi P = \pi$ and P is π-irreducible.*
- ***Small set.*** *There exist $J \subseteq \mathcal{X}$ with $\pi(J) > 0$, a probability measure ν and $\beta > 0$ such that*

$$P(x, \cdot) \geq \beta \mathbb{I}(x \in J) \nu(\cdot).$$

Small set

Transition kernel:

$$P(x, A) = \mathbb{P}(X_n \in A | X_{n-1} = x).$$

ASSUMPTION

- ***Stationary distribution.*** *There exists a probability distribution π on $\mathcal{X}$ such that $\pi P = \pi$ and P is π-irreducible.*
- ***Small set.*** *There exist $J \subseteq \mathcal{X}$ with $\pi(J) > 0$, a probability measure ν and $\beta > 0$ such that*

$$P(x, \cdot) \geq \beta \mathbb{I}(x \in J) \nu(\cdot).$$

Regeneration

(Nummelin 1978, Athreya and Ney 1978.)

The "residual" stochastic kernel:

$$Q(x, \cdot) = \frac{P(x, \cdot) - \beta \mathbb{I}(x \in J)\nu(\cdot)}{1 - \beta \mathbb{I}(x \in J)}.$$

- If $X_{n-1} \notin J$ then draw $X_n \sim P(X_{n-1}, \cdot)$, **no regeneration**;
- If $X_{n-1} \in J$ then
  - with probability $1 - \beta$ draw $X_n \sim Q(X_{n-1}, \cdot)$, **no regeneration**;
  - with probability β draw $X_n \sim \nu(\cdot)$, **Regeneration**.

Regeneration

Times of regeneration partition a trajectory into **iid blocks**:

$$\underbrace{X_0,\dots,X_{T_1-1}}_{T_1},\quad \underbrace{X_{T_1},\dots,X_{T_2-1}}_{T_2-T_1},\quad \underbrace{X_{T_2},\dots,X_{T_3-1}}_{T_3-T_2},\quad \dots$$

$\uparrow$ R $\qquad\qquad$ $\uparrow$ R

R = Regeneration

Block sums:

$$B_k(f) = \sum_{i=T_{k-1}}^{T_k-1} f(X_i).$$

$T = T_1 =_d T_k - T_{k-1}$ under $\mathbb{P}_\nu$,

$B(f) = B_1(f) =_d B_k(f)$ under $\mathbb{P}_\nu$,

(ν is the regeneration distribution).

Regeneration

Times of regeneration partition a trajectory into **iid blocks**:

$$\underbrace{X_0, \ldots, X_{T_1-1}}_{T_1}, \quad \underbrace{X_{T_1}, \ldots, X_{T_2-1}}_{T_2-T_1}, \quad \underbrace{X_{T_2}, \ldots, X_{T_3-1}}_{T_3-T_2}, \quad \ldots$$

$$\begin{array}{cc} \uparrow & \uparrow \\ \text{R} & \text{R} \end{array}$$

R = Regeneration

Block sums:

$$B_k(f) = \sum_{i=T_{k-1}}^{T_k-1} f(X_i).$$

$T = T_1 =_d T_k - T_{k-1}$ under $\mathbb{P}_\nu$,

$B(f) = B_1(f) =_d B_k(f)$ under $\mathbb{P}_\nu$,

(ν is the regeneration distribution).

Regeneration

Times of regeneration partition a trajectory into **iid blocks**:

$$\underbrace{X_0, \dots, X_{T_1-1},}_{T_1} \quad \underbrace{X_{T_1}, \dots, X_{T_2-1},}_{T_2-T_1} \quad \underbrace{X_{T_2}, \dots, X_{T_3-1},}_{T_3-T_2} \quad \dots$$

$$\qquad\qquad\quad \uparrow \qquad\qquad\qquad\qquad \uparrow$$

$$\qquad\qquad\quad \mathrm{R} \qquad\qquad\qquad\qquad \mathrm{R}$$

R = Regeneration

Block sums:

$$B_k(f) = \sum_{i=T_{k-1}}^{T_k-1} f(X_i).$$

$T = T_1 =_d T_k - T_{k-1}$ under $\mathbb{P}_\nu$,

$B(f) = B_1(f) =_d B_k(f)$ under $\mathbb{P}_\nu$,

(ν is the regeneration distribution).

Mean Square Error

The integral of interest and its MCMC estimator:

$$\theta = \pi(f) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x), \qquad \hat{\theta}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i).$$

THEOREM

Under Assumptions **Stationary distribution** *and* **Small set**,

$$\sqrt{\mathbb{E}_\xi\left(\hat{\theta}_n - \theta\right)^2} \le \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}}\left(1 + \frac{C_0}{n}\right) + \frac{C_1(f)}{n} + \frac{C_2(f)}{n},$$

where $\bar{f} = f - \theta$,

$$\sigma^2_{\mathrm{as}}(f) := \frac{\mathbb{E}_\nu B(\bar{f})^2}{\mathbb{E}_\nu T}, \qquad C_0 := \mathbb{E}_\pi T = \frac{\mathbb{E}_\nu T^2}{2\mathbb{E}_\nu T} + \frac{1}{2},$$

$$C_1(f) := \sqrt{\mathbb{E}_\xi B(\bar{f})^2},$$

$$C_2(f) := \sqrt{\mathbb{E}_{\xi P^n} B(\bar{f})^2}.$$

Mean Square Error

The integral of interest and its MCMC estimator:

$$\theta = \pi(f) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x), \qquad \hat{\theta}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i).$$

THEOREM

Under Assumptions **Stationary distribution** *and* **Small set**,

$$\sqrt{\mathbb{E}_\xi\left(\hat{\theta}_n - \theta\right)^2} \le \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}}\left(1 + \frac{C_0}{n}\right) + \frac{C_1(f)}{n} + \frac{C_2(f)}{n},$$

where $\bar{f} = f - \theta$,

$$\sigma^2_{\mathrm{as}}(f) := \frac{\mathbb{E}_\nu B(\bar{f})^2}{\mathbb{E}_\nu T}, \qquad C_0 := \mathbb{E}_\pi T = \frac{\mathbb{E}_\nu T^2}{2\mathbb{E}_\nu T} + \frac{1}{2},$$

$$C_1(f) := \sqrt{\mathbb{E}_\xi B(\bar{f})^2},$$

$$C_2(f) := \sqrt{\mathbb{E}_{\xi P^n} B(\bar{f})^2}.$$

Mean Square Error

Remark

In our theorem we have

- *Possibly unbounded function f;*
- *Nonasymptotic bound with the "correct leading term":*

$$\sqrt{\mathbb{E}_\xi\,(\hat{\theta}_n - \theta)^2} \le \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}}\left(1 + \frac{C_0}{n}\right) + \frac{C_1(f)}{n} + \frac{C_2(f)}{n},$$

while

$$\sqrt{\mathbb{E}_\xi\,(\hat{\theta}_n - \theta)^2} \sim \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}} \qquad (n \to \infty).$$

and $\sigma^2_{\mathrm{as}}(f)$ is the "asymptotic variance" in the CLT:

$$\frac{\sqrt{n}}{\sigma_{\mathrm{as}}(f)}(\hat{\theta}_n - \theta) \to_d \mathcal{N}(0,1) \qquad (n \to \infty).$$

Mean Square Error

Remark

In our theorem we have

- *Possibly unbounded function f;*
- *Nonasymptotic bound with the "correct leading term":*

$$\sqrt{\mathbb{E}_\xi\,(\hat\theta_n - \theta)^2} \le \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}}\left(1 + \frac{C_0}{n}\right) + \frac{C_1(f)}{n} + \frac{C_2(f)}{n},$$

while

$$\sqrt{\mathbb{E}_\xi\,(\hat\theta_n - \theta)^2} \sim \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}} \qquad (n \to \infty).$$

and $\sigma^2_{\mathrm{as}}(f)$ is the "asymptotic variance" in the CLT:

$$\frac{\sqrt{n}}{\sigma_{\mathrm{as}}(f)}(\hat\theta_n - \theta) \to_d \mathcal{N}(0,1) \qquad (n \to \infty).$$

Mean Square Error

Remark

In our theorem we have

- *Possibly unbounded function f;*
- *Nonasymptotic bound with the “correct leading term”:*

$$\sqrt{\mathbb{E}_\xi\,(\hat{\theta}_n-\theta)^2} \le \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}}\left(1+\frac{C_0}{n}\right)+\frac{C_1(f)}{n}+\frac{C_2(f)}{n},$$

while

$$\sqrt{\mathbb{E}_\xi\,(\hat{\theta}_n-\theta)^2} \sim \frac{\sigma_{\mathrm{as}}(f)}{\sqrt{n}} \qquad (n\to\infty).$$

and $\sigma^2_{\mathrm{as}}(f)$ is the “asymptotic variance” in the CLT:

$$\frac{\sqrt{n}}{\sigma_{\mathrm{as}}(f)}(\hat{\theta}_n-\theta)\to_d \mathcal{N}(0,1) \qquad (n\to\infty).$$

Mean Square Error

Relation to previous work (nonasymptotic bounds on MSE of MCMC):

WN, P.Pokarowski, *Fixed precision MCMC Estimation by Median of Products of Averages*, *J. Appl. Probab.* 2009. Discrete space.

D.Rudolf, *Explicit error bounds for lazy reversible Markov chain Monte Carlo*, *J. Complex.* 2009. Bounds in terms of spectral gap.

K.Łatuszyński, WN, *Rigorous confidence bounds for MCMC under a geometric drift condition*, to appear in *J. Complex.* Same problem, different method.

K.Łatuszyński, B.Miasojedow, WN, *Nonasymptotic bounds on the estimation error for regenerative MCMC algorithms*, submitted. A result on *sequential* estimator, which requires **identification of regeneration times**.

Mean Square Error

Remark

- *Our present result concerns the **standard** scheme with deterministic length n of simulation.*
- *Our bound would **not** improve if we added a burn-in time at the beginning of simulation. (Perhaps this is a disadvantage?)*

Proof

Look for the first regeneration past n:

$$R(n) = \min\{r : T_r > n\}.$$

$$0, \ldots, T_1 - 1, \quad \underset{\substack{\uparrow \\ \color{red}{\mathrm{R}}}}{T_1}, \ldots\ldots, \quad \underset{\substack{\uparrow \\ \color{red}{\mathrm{R}}}}{T_{R(n)-1}}, \ldots, \underset{\substack{\uparrow \\ n}}{n}, \ldots, T_{R(n)} - 1, \quad \underset{\substack{\uparrow \\ \color{red}{\mathrm{R}}}}{T_{R(n)}}, \ldots$$

Then express the error as follows:

$$\hat{\theta}_n - \theta = \frac{1}{n}\sum_{i=0}^{n-1} \bar{f}(X_i) = \frac{1}{n}(\mathcal{O}_1 + \mathcal{Z} - \mathcal{O}_2)$$

$$= \frac{1}{n}\left(\sum_{i=0}^{T_1-1} \bar{f}(X_i) + \sum_{i=T_1}^{T_{R(n)-1}} \bar{f}(X_i) - \sum_{i=n}^{T_{R(n)}-1} \bar{f}(X_i)\right),$$

$$\overbrace{0, \ldots, T_1 - 1}^{\mathcal{O}_1}, \quad \color{blue}{T_1, \ldots\ldots, \quad T_{R(n)-1}, \ldots, \overbrace{n, \ldots, T_{R(n)} - 1}^{\mathcal{O}_2},} \quad T_{R(n)}, \ldots$$

Proof continued

The **main term**

$$\mathcal{Z} = \sum_{i=T_1}^{T_{R(n)-1}} \bar{f}(X_i) = \sum_{k=2}^{R(n)} B_k(\bar{f})$$

is marked in blue:

$$0, \dots, T_1 - 1, \quad \overbrace{T_1, \dots, T_{R(n)-1}, \dots, n, \dots, T_{R(n)} - 1,}^{\mathcal{Z}} \quad T_{R(n)}, \dots$$

$\mathcal{Z}$ is a sum of a **random number** of **iid** summands and $R(n)$ is a **stopping time.** Therefore tools of *sequential analysis* apply:

- Two identities of Abraham Wald.
- Lorden's theorem (1970, renewal theory).

The other two terms, $\mathcal{O}_1$ and $\mathcal{O}_1$ have to be bounded separately.

Drift condition – geometrically ergodic chains

Drift towards **small set** J.

ASSUMPTION

- ***Drift.*** *There exist a function* $V : \mathcal{X} \to [1, \infty[$, *constants* $\lambda < 1$ *and* $K < \infty$ *such that*

$$PV^2(x) := \int_{\mathcal{X}} P(x, dy) V^2(y) \leq \begin{cases} \lambda^2 V^2(x) & \text{for } x \notin J, \\ K^2 & \text{for } x \in J. \end{cases}$$

Remark

Notation V^2, λ^2, K^2 *– to simplify further statements.*

Explicit bounds under drift condition

Bounds on constants $\sigma^2_{\mathrm{as}}(f)$, C_0, $C_1(f)$, $C_2(f)$.

Under Assumptions **SmallSet** and **Drift**, if f is such that $\|\bar{f}\|_V := \sup_x |\bar{f}(x)|/V(x) < \infty$ then

$$\sigma^2_{\mathrm{as}}(f) \leq \|\bar{f}\|^2_V \left(\frac{1+\lambda}{1-\lambda}\pi(V^2) + \frac{2(K-\lambda-\beta)}{\beta(1-\lambda)}\pi(V) \right),$$

$$C_0 \leq \frac{\lambda}{1-\lambda}\pi(V) + \frac{K-\lambda(1+\beta)}{\beta(1-\lambda)},$$

$$C_1(f)^2 \leq \frac{1}{(1-\lambda)^2}\xi(V^2) + \frac{2(K-\lambda-\beta)}{\beta(1-\lambda)^2}\xi(V) + \frac{\beta(3+\lambda)(K^2-\lambda^2-\beta)+2(1+\lambda)(K-\lambda-\beta)^2}{\beta^2(1-\lambda)^2(1+\lambda)},$$

$C_2(f)^2 \leq$ analogous expression with ξ replaced by ξP^n.

Explicit bounds under drift condition

Further bounds on quantities $\pi(V)$, $\pi(V^2)$, $\xi P^n(V)$, $\xi P^n(V^2)$, $\|\bar{f}\|_V$.

Under Assumptions **SmallSet** and **Drift**,

$$\pi(V) \leq \pi(J)\frac{K-\lambda}{1-\lambda} \leq \frac{K-\lambda}{1-\lambda},$$

$$\pi(V^2) \leq \pi(J)\frac{K^2-\lambda^2}{1-\lambda^2} \leq \frac{K^2-\lambda^2}{1-\lambda^2},$$

$$\text{if } \xi(V) \leq \frac{K}{1-\lambda} \text{ then } \xi P^n(V) \leq \frac{K}{1-\lambda},$$

$$\text{if } \xi(V^2) \leq \frac{K^2}{1-\lambda^2} \text{ then } \xi P^n(V^2) \leq \frac{K^2}{1-\lambda^2},$$

$$\|\bar{f}\|_V \leq \|f\|_V + \frac{\pi(J)(K-\lambda)}{(1-\lambda)\inf_{x\in\mathcal{X}} V(x)} \leq \|f\|_V + \frac{K-\lambda}{1-\lambda}.$$

Confidence estimation via Median of Averages

Goal:

$$\mathbb{P}(|\hat{\theta} - \theta| \leq \varepsilon) \geq 1 - \alpha$$

(given precision ε at a given level of confidence $1 - \alpha$).

Median of Averages (MA) (Jerrum, Valiant and Vazirani, 1986):

Generate m independent copies of the Markov chain and compute averages:

$$X_0^{(1)}, X_1^{(1)}, \dots, X_{n-1}^{(1)} \quad \longmapsto \quad \hat{\theta}_n^{(1)} = \sum_{i=0}^{n-1} f(X_i^{(1)}),$$

$$\dots$$

$$X_0^{(m)}, X_1^{(m)}, \dots, X_{n-1}^{(m)} \quad \longmapsto \quad \hat{\theta}_n^{(m)} = \sum_{i=0}^{n-1} f(X_i^{(m)}).$$

Estimator MA:

$$\hat{\theta}_{m,n} = \mathrm{med}\left(\hat{\theta}_n^{(1)}, \dots, \hat{\theta}_n^{(m)}\right).$$

Complexity comparison

Asymptotic level of confidence, based on CLT:

$$\lim_{\varepsilon \to 0} \mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) = \alpha,$$

for $\hat{\theta}_n$ – a simple average over n samples.

Nonasymptotic level of confidence:

$$\mathbb{P}(|\hat{\theta}_{m,n} - \theta| > \varepsilon) \leq \alpha,$$

for $\hat{\theta}_{m,n}$ – the MA estimator.

Complexity comparison

Asymptotic level of confidence, based on CLT:

$$\lim_{\varepsilon \to 0} \mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) = \alpha,$$

for $\hat{\theta}_n$ – a simple average over n samples.

Nonasymptotic level of confidence:

$$\mathbb{P}(|\hat{\theta}_{m,n} - \theta| > \varepsilon) \leq \alpha,$$

for $\hat{\theta}_{m,n}$ – the MA estimator.

Complexity comparison

How many samples do we need to achieve accuracy ε at the confidence level $1-\alpha$?

Asymptotic level for $\hat{\theta}_n$:

$$n \sim \frac{\sigma^2_{\text{as}}(f)}{\varepsilon^2}\left[\Phi^{-1}(1-\alpha/2)\right]^2,$$

Nonasymptotic level for $\hat{\theta}_{m,n}$:

$$mn \sim \frac{\sigma^2_{\text{as}}(f)}{\varepsilon^2} C \log(2\alpha)^{-1},$$

where symbol $\sim$ refers to $\alpha, \varepsilon \to 0$ and $C \approx 19.34$ (Niemiro and Pokarowski, 2009).

Since $\left[\Phi^{-1}(1-\alpha/2)\right]^2 \sim 2\log(2\alpha)^{-1}$ for $\alpha \to 0$, nonasymptotic confidence is about 10 times more expensive than asymptotic one.

Complexity comparison

How many samples do we need to achieve accuracy ε at the confidence level $1 - \alpha$?

Asymptotic level for $\hat{\theta}_n$:

$$n \sim \frac{\sigma^2_{\mathrm{as}}(f)}{\varepsilon^2}\left[\Phi^{-1}(1-\alpha/2)\right]^2,$$

Nonasymptotic level for $\hat{\theta}_{m,n}$:

$$mn \sim \frac{\sigma^2_{\mathrm{as}}(f)}{\varepsilon^2} C \log(2\alpha)^{-1},$$

where symbol $\sim$ refers to $\alpha, \varepsilon \to 0$ and $C \approx 19.34$ (Niemiro and Pokarowski, 2009).

Since $\left[\Phi^{-1}(1-\alpha/2)\right]^2 \sim 2\log(2\alpha)^{-1}$ for $\alpha \to 0$, nonasymptotic confidence is about 10 times more expensive than asymptotic one.

The end

Thank you.