# A Monte Carlo estimation of the entropy for Markov chains

Didier Chauveau Pierre Vandekerkhove

Université de Marne-la-Vallée
Laboratoire d'analyse et de mathématiques Appliquées
CNRS UMR 8050
5 Bd Descartes, Champs-sur-Marne
77454 Marne-la-Vallée Cedex 2, France.
chauveau@math.univ-mlv.fr, vandek@math.univ-mlv.fr

March 2004

**Abstract.** We introduce an estimate of the entropy $\mathbb{E}_{p^t}(\log p^t)$ of the marginal density $p^t$ of a (eventually inhomogeneous) Markov chain at time $t \geq 1$. This estimate is based on a double Monte Carlo integration over simulated i.i.d. copies of the Markov chain, whose transition density kernel is supposed to be known. The technique is extended to compute the *external* entropy $\mathbb{E}_{p_1^t}(\log p^t)$, where the $p_1^t$'s are the successive marginal densities of another Markov process at time $t$. We prove, under mild conditions, weak consistency and asymptotic normality of both estimators. The strong consistency is also obtained under stronger assumptions. These estimators can be used to study by simulation the convergence of $p^t$ to its stationary distribution. Potential applications for this work are presented: (i) a diagnostic by simulation of the stability property of a Markovian dynamical system with respect to various initial conditions; (ii) a study of the rate in the Central Limit Theorem for i.i.d. random variables. Simulated examples are provided as illustration.

**keywords:** Entropy, Kullback information, Markov chains, stability, rate in the Central Limit Theorem.
**AMS 1994 subject classification:** 62B10, 60F05, 60J10.

## 1 Introduction

Let $X = (X^t)_{t\geq 0}$ be a inhomogeneous discrete time Markov chain taking values in a measurable state space $(E, \mathcal{E})$, with analytically known transition density kernel $(q^t)$, $t \geq 0$ with respect to a $\sigma$-finite reference measure $\nu$. The successive marginal density functions of the Markov chain $X$ are given by its initial density $p^0$, and the

recurrence formula:

$$p^{t+1}(y) = \int_E p^t(x) q^t(x,y)\, \nu(dx), \quad t \geq 0, \tag{1}$$

The essential idea of this paper is to take advantage of the knowledge of $q^t(x,y)$, and of the integral formula (1), to compute the (relative) entropy of the marginal density at time $t$,

$$\mathcal{H}(p^t) = \mathbb{E}_{p^t}(\log p^t),$$

via simulation and a sort of double Monte Carlo integration.

We assume that we can simulate, for all $t \geq 0$, $2N$ i.i.d. outcomes of the process $X$. For simpler notations we divide these $2N$ copies into two sets: one set of size $N$, denoted at time $t$ by

$$\mathbf{X}^t = (X_1^t, X_2^t, \ldots, X_N^t) \quad \text{i.i.d.} \sim p^t,$$

and the second set (of size $N$) denoted at time $t$ by

$$\tilde{\mathbf{X}}^t = (\tilde{X}_1^t, \tilde{X}_2^t, \ldots, \tilde{X}_N^t) \quad \text{i.i.d.} \sim p^t.$$

For all $y \in E$ and $t \geq 0$, the Strong Law of Large Numbers (SLLN) for the i.i.d. random variables from the first set gives:

$$\frac{1}{N}\sum_{k=1}^{N} q^t(X_k^t, y) \xrightarrow{a.s.} \int_E q^t(x,y) p^t(x)\, \nu(dx) = p^{t+1}(y) \text{ as } N \to \infty. \tag{2}$$

Hence we can expect that the Monte Carlo integration of the logarithm of the left side of (2), using the second set at time $t+1$, which is i.i.d. from $p^{t+1}$, converges to $\mathcal{H}(p^{t+1})$. Thus it is natural to introduce the following "double Monte Carlo" estimator based on the two samples $(\mathbf{X}^t, \tilde{\mathbf{X}}^{t+1})$:

$$\hat{\mathcal{H}}_N(p^{t+1}) = \frac{1}{N}\sum_{\ell=1}^{N} \log\left(\frac{1}{N}\sum_{k=1}^{N} q^t(X_k^t, \tilde{X}_\ell^{t+1})\right). \tag{3}$$

The "technical trick" of simulating two sets of the same Markov process, used at two consecutive times *but independents*, allows for a simpler theoretical study of the asymptotic behavior of $\hat{\mathcal{H}}_N(p^t)$. Precisely, we preserve the marginal distributions but avoid the Markovian dependence between $X^t$ and $X^{t+1}$ by simulating independent copies from $p^t$ and $p^{t+1}$.

Estimation of the relative entropy $\mathbb{E}_{p^t}[\log p^t]$ in a different context (with i.i.d. observations), and using nonparametric techniques (e.g., kernel density estimates), has already been studied by different authors like Levit (1978), Tsybakov and Van Der Meulen (1994), Ahmad and Lin (1989), Eggermont and LaRiccia (1999). These authors prove consistency and asymptotic normality of their estimators in the univariate case, under suitable technical conditions on the densities. Eggermont and LaRiccia (1999) prove consistency in the multivariate case, but do not establish asymptotic normality. One interest of our method is that we can prove consistency and asymptotic normality in the multivariate case.

### 1.1 External entropy between two processes

This intuitive approach can be generalized to compute in the same way the *external* entropy

$$\mathcal{H}(p_1^t, p^t) = \mathbb{E}_{p_1^t}(\log p^t)$$

involving a second Markov process $Y = (Y^t)_{t\geq 0}$ with different transition kernel, and/or initial distribution, and consequently different marginal densities $p_1^t$ at time $t$. The transition density of $Y$ needs not to be analytically known, but we still need to know how to simulate copies of $Y$. If we denote now these simulations at time $t$ by

$$\mathbf{Y}^t = (Y_1^t, Y_2^t, ..., Y_N^t) \quad \text{i.i.d.} \sim p_1^t,$$

then, following the construction given above, it is natural to propose the estimator

$$\hat{\mathcal{H}}_N(p_1^{t+1}, p^{t+1}) = \frac{1}{N}\sum_{\ell=1}^{N} \log\left(\frac{1}{N}\sum_{k=1}^{N} q^t(X_k^t, Y_\ell^{t+1})\right) \tag{4}$$

which should converge to $\mathcal{H}(p_1^{t+1}, p^{t+1})$. Up to our knowledge, this external entropy estimation problem, i.e. estimation of $\mathbb{E}_{p_1^t}[\log(p^t)]$ with $p^t \neq p_1^t$, has never been studied before.

In section 2, we study the asymptotic properties of these entropy estimates and show their weak consistency and asymptotic normality in the multivariate case, under moment conditions on the transition dentities. Their almost sure convergence is also obtained with stronger assumptions.

Then two potential applications where both estimates can be useful are presented: Section 3 introduce a simulation method to diagnose the stability property of a Markov process by studying its sensitivity to different initial conditions. Our method can be used to estimate the Kullback information between the successive densities of different outcomes of the process. This measure should geometrically decrease to zero under good stability properties, as shown in recent theoretical works (see, e.g., Del Moral, Ledoux, and Miclo, 2003). Section 4 gives a simulation method to study the rate in the Central Limit Theorem (CLT) for i.i.d. random variables, by estimating the Kullback information between the pdf of the normalized sums and the gaussian. This question also motivates current theoretical developments, as in, e.g., Ball, Barthe and Naor (2003). The behavior of our estimates are illustrated via simulated examples for both situations.

## 2 A double Monte Carlo entropy estimator

Since we are using the independent samples $(\mathbf{X}^t, \tilde{\mathbf{X}}^{t+1})$ for $\hat{\mathcal{H}}_N(p^{t+1})$, and $(\mathbf{X}^t, \mathbf{Y}^{t+1})$ for $\hat{\mathcal{H}}_N(p_1^{t+1}, p^{t+1})$, the behavior of both estimates are similar; only the successive densities change (as already stated, we avoid the Markovian dependence between $X^t$

and $X^{t+1}$ by simulating different independent copies). This allows us to give a unique proof for their asymptotic behavior. In the framework defined in the introduction we have the following result, stated for $\hat{\mathcal{H}}_N(p_1^{t+1}, p^{t+1})$:

**Theorem 1** *If, for all $t \geq 0$, the "normalized" transition density kernel defined by*

$$r^t(x,y) = \frac{q^t(x,y)}{p^{t+1}(y)} \tag{5}$$

*is non-degenerate and satisfies:*

$$\mathbb{E}_{p^t \otimes p_1^{t+1}} \left[ |r^t(X,Y)|^{2+\gamma} \right] < \infty \quad \textit{for some } \gamma > 0, \tag{6}$$

*and*

$$\mathbb{E}_{p_1^{t+1}}[|\log p^{t+1}(Y)|^2] < \infty, \tag{7}$$

*then:*

$$\begin{aligned} \hat{\mathcal{H}}_N(p_1^{t+1}, p^{t+1}) &\xrightarrow{\mathbb{P}} \mathcal{H}(p_1^{t+1}, p^{t+1}), \quad \textit{as } N \to \infty, \\ \sqrt{N}(\hat{\mathcal{H}}_N(p_1^{t+1}, p^{t+1}) - \mathcal{H}(p_1^{t+1}, p^{t+1})) &\xrightarrow{d} \mathcal{N}(0, \Sigma^t), \quad \textit{as } N \to \infty, \end{aligned}$$

*where $\Sigma^t = \mathrm{var}_{p_1^{t+1}}[\log p^{t+1}] + \mathrm{var}_{p^t}[R(X)]$, and $R(x) = \mathbb{E}_{p_1^{t+1}}[r^t(x,Y)]$.*

*Proof.* We consider the following decomposition:

$$\hat{\mathcal{H}}_N(p_1^{t+1}, p^{t+1}) - \mathcal{H}(p_1^{t+1}, p^{t+1}) = T_{1,N} + T_{2,N},$$

where

$$\begin{aligned} T_{1,N} &= \frac{1}{N}\sum_{\ell=1}^N \log\left(\frac{1}{N}\sum_{k=1}^N q^t(X_k^t, Y_\ell^{t+1})\right) - \frac{1}{N}\sum_{\ell=1}^N \log\left(p^{t+1}(Y_\ell^{t+1})\right) \\ &= \frac{1}{N}\sum_{\ell=1}^N \log\left(\frac{1}{N}\sum_{k=1}^N r^t(X_k^t, Y_\ell^{t+1})\right), \end{aligned}$$

where $r^t$ has been defined in (5) and satisfies $\mathbb{E}_{p^t}\left[r^t(X,y)\right] = 1$ for all $y \in E$, and

$$\begin{aligned} T_{2,N} &= \frac{1}{N}\sum_{\ell=1}^N \log\left(p^{t+1}(Y_\ell^{t+1})\right) - \mathbb{E}_{p_1^{t+1}}[\log(p^{t+1})] \\ &= \frac{1}{N}\sum_{\ell=1}^N w^t(Y_\ell^{t+1}), \end{aligned} \tag{8}$$

where $w^t(Y_\ell^{t+1}) = \log\left(p^{t+1}(Y_\ell^{t+1})\right) - \mathbb{E}_{p_1^{t+1}}[\log(p^{t+1})]$. To simplify notations in the sequel of the proof, we drop from now on the superscripts indicating the fixed times

at which the estimation is performed. Hence we write $X_k$ and $Y_\ell$ for $X_k^t$ and $Y_\ell^{t+1}$, keeping in mind that $X_k \sim p^t$ and $Y_\ell \sim p_1^{t+1}$.

For $T_{1,N}$ we consider (as in Del Moral and Guionnet (1999)) the second-order Taylor representation of the log function:

$$\log x = (x-1) - \frac{(x-1)^2}{2(\theta x + (1-\theta))^2}$$

valid for all $x > 0$ with $\theta = \theta(x)$ such that $\theta(x) \in [0,1]$. We thus can split $T_{1,N}$ into

$$T_{1,N} = J_N - R_N, \tag{9}$$

where

$$J_N = \frac{1}{N^2} \sum_{\ell=1}^{N} \sum_{k=1}^{N} \left( r(X_k, Y_\ell) - 1 \right), \tag{10}$$

and

$$R_N = \frac{1}{2N} \sum_{\ell=1}^{N} \frac{\left( N^{-1} \sum_{k=1}^{N} \left( r(X_k, Y_\ell) - 1 \right) \right)^2}{\left( \theta N^{-1} \sum_{k=1}^{N} \left( r(X_k, Y_\ell) \right) + (1-\theta) \right)^2}.$$

We now prove that $T_{2,N} + J_N$ satisfies the SLLN and the CLT, and that $\sqrt{N} R_N \to 0$ in probability as $N \to \infty$. To handle $T_{2,N}$ and $J_N$, we exhibit a symmetric representation. We have

$$\begin{aligned} T_{2,N} + J_N &= \frac{1}{N^2} \sum_{1 \le \ell = k \le N} \left( w(Y_\ell) + r(X_k, Y_\ell) - 1 \right) \\ &\quad + \frac{1}{N^2} \sum_{1 \le \ell \ne k \le N} \left( w(Y_\ell) + r(X_k, Y_\ell) - 1 \right). \end{aligned} \tag{11}$$

Using standard arguments, the first term of the right-hand side goes to 0 almost surely. Defining $Z_\ell = (X_\ell, Y_\ell)$, and $s_1$ (resp. $s_2$) the mapping from $E^2$ into $E$: $(x,y) \mapsto x$ (resp. $(x,y) \mapsto y$), the second term in the right-hand side of (11) can be written as

$$U(Z_1, Z_2, ..., Z_N) = \frac{1}{N^2} \sum_{1 \le \ell < k \le N} h(Z_k, Z_\ell),$$

with symmetric kernel

$$\begin{aligned} h(z_1, z_2) &= w(s_2(z_2)) + w(s_2(z_1)) \\ &\quad + r(s_1(z_1), s_2(z_2)) + r(s_1(z_2), s_2(z_1)) - 2. \end{aligned}$$

In this way we have written a $U$-statistic depending on the i.i.d. sample $\mathbf{Z} = (Z_1, \ldots, Z_N)$. According to Serfling(1980) p. 190-192, and Lehmann (1975), under the moment condition $\mathbb{E}[|h|^2] < \infty$ (which is a consequence of (6)), $U(\mathbf{Z})$ satisfies the SLLN and the CLT (and the Law of Iterated Logarithm) as $N \to \infty$, i.e.:

$$\begin{aligned} U(\mathbf{Z}) &\longrightarrow 0 \quad \text{a.s.}, \quad \text{as } N \to \infty, \\ \sqrt{N}\, U(\mathbf{Z}) &\xrightarrow{d} \mathcal{N}(0, \Sigma), \quad \text{as } N \to \infty, \end{aligned}$$

with $\Sigma = \mathrm{var}_{p_1^{t+1}}[R_0(Y)] + \mathrm{var}_{p^t}[R_1(X)]$, where $R_0(y) = \mathbb{E}_{p^t}[w^t(y)] = w^t(y)$ and $R_1(x) = \mathbb{E}_{p_1^{t+1}}[r^t(x, Y)] - 1$ (see Lehmann (1975) p. 362–364).

**Remark:** $J_N$ alone can also be expressed as a $U$-statistic that goes to zero a.s., using the same technique. This will be used in the proof of Theorem 2.

We still have to control the remainder term $R_N$. To show weak consistency and asymptotic normality of $T_{1,N}$, we have to prove that $\sqrt{N}R_N \to 0$ in probability when $N \to \infty$. To this end, we use the following decomposition of $R_N$:

$$|\sqrt{N}R_N| \leq \frac{1}{2N^{1/2-\delta}} \sum_{\ell=1}^{N} \left( \frac{1}{N} \sum_{k=1}^{N} r(X_k, Y_\ell) - 1 \right)^2 \frac{1}{N^\delta} \max_{1 \leq \ell \leq N} \xi_\ell \tag{12}$$

for $0 < \delta < 1/2$, where $\xi_\ell^{-1} = \min \left( \left[ \frac{1}{N} \sum_{k=1}^{N} r(X_k, Y_\ell) \right]^2, 1 \right)$.

We consider the first term in the right hand side of (12). In view of Markov's inequality, we have, for any $\varepsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left( \frac{1}{2N^{1/2-\delta}} \sum_{\ell=1}^{N} \left( \frac{1}{N} \sum_{k=1}^{N} r(X_k, Y_\ell) - 1 \right)^2 > \varepsilon \right) \\ &\leq \frac{1}{2\varepsilon N^{5/2-\delta}} \mathbb{E}\left( \sum_{\ell=1}^{N} \left[ \sum_{k,k'=1}^{N} [r(X_k, Y_\ell) - 1][r(X_{k'}, Y_\ell) - 1] \right] \right) \\ &= \frac{1}{2\varepsilon N^{5/2-\delta}} \mathbb{E}\left( \sum_{\ell=1}^{N} \mathbb{E}\left[ \sum_{k,k'=1}^{N} [r(X_k, Y_\ell) - 1][r(X_{k'}, Y_\ell) - 1] \,\middle|\, Y_\ell \right] \right) \\ &= \frac{1}{2\varepsilon N^{5/2-\delta}} \mathbb{E}\left[ \sum_{\ell=1}^{N} \sum_{k=1}^{N} (r(X_k, Y_\ell) - 1)^2 \right] \\ &= \frac{1}{2\varepsilon N^{1/2-\delta}} \mathbb{E}_{p^t \otimes p_1^{t+1}}\left[ (r(X, Y) - 1)^2 \right]. \end{aligned}$$

The main arguments in the previous steps are the conditional independence of $r(X_k, Y_\ell)$ and $r(X_{k'}, Y_\ell)$ with respect to $Y_\ell$, and the fact that $\mathbb{E}[r(X, y)] = 1$ for all $y \in E$.

We now consider the term $\max_{1\leq\ell\leq N}\xi_\ell$. Let $A_N$ and $B_N$ be the events

$$\begin{aligned} A_N &= \left\{ \frac{1}{N^\delta}\left( \min_{1\leq\ell\leq N} \min\left\{ \left(\frac{1}{N}\sum_{k=1}^N r(X_k,Y_\ell)\right)^2, 1\right\}\right)^{-1} > \varepsilon \right\} \\ B_N &= \left\{ \min_{1\leq\ell\leq N}\left(\frac{1}{N}\sum_{k=1}^N r(X_k,Y_\ell)\right)^2 > 1 \right\}. \end{aligned}$$

We have then

$$\begin{aligned} \mathbb{P}(A_N) &= \mathbb{P}(A_N \mid B_N)\mathbb{P}(B_N) + \mathbb{P}(A_N \mid B_N^c)\mathbb{P}(B_N^c) \\ &\leq \mathbb{P}(1 > N^\delta\varepsilon) + \mathbb{P}\left( \left[ \min_{1\leq\ell\leq N}\left(\frac{1}{N}\sum_{k=1}^N r(X_k,Y_\ell)\right)^2\right]^{-1} > N^\delta\varepsilon \right), \end{aligned}$$

where the first term in the last expression vanishes for $N$ large enough. Using Bonferroni's inequality, we have:

$$\begin{aligned} &\mathbb{P}\left( \left[ \min_{1\leq\ell\leq N}\left(\frac{1}{N}\sum_{k=1}^N r(X_k,Y_\ell)\right)^2\right]^{-1} > N^\delta\varepsilon \right) \\ &\leq N\mathbb{P}\left( \left(\frac{1}{N}\sum_{k=1}^N r(X_k,Y)\right)^2 < \frac{1}{N^\delta\varepsilon}\right) \\ &\leq \int_E N\mathbb{P}\left( \left|\frac{1}{N}\sum_{k=1}^N r(X_k,Y)\right| < \frac{1}{\sqrt{N^\delta\varepsilon}} \,\middle|\, Y=y\right) p_1^{t+1}(y)\,\nu(dy), \\ &\leq \int_E N\mathbb{P}\left( |S_N(y)| > N(1-\frac{N^{-\delta/2}}{\sqrt{\varepsilon}})\right) p_1^{t+1}(y)\,\nu(dy), \end{aligned} \tag{13}$$

where $S_N(y) = \sum_{k=1}^N (r(X_k,y)-1)$ is a sum of $N$ i.i.d. and centered random variables. We can then use the moderate deviation type inequality (2.79) in Petrov (1995) (due to Fuk and Nagaev (1971, 1976)), which holds under the moment condition

$$M_m(y) = \mathbb{E}_{p^t}\left[\,|r(X,y)-1|^m\,\right] < \infty, \quad m > 2, \quad y \in E.$$

In our case, this condition is a consequence of condition (6) (moment of order $2+\gamma$) and Fubini's theorem. This inequality, applied here with $m = 2+\gamma/2$ to the conditional probability in the integral term, gives inequality (15) below:

$$N\mathbb{P}\left( |S_N(y)| > N(1-\frac{N^{-\delta/2}}{\sqrt{\varepsilon}})\right) \leq 2N\mathbb{P}(S_N(y) > \alpha N) \tag{14}$$

$$\leq 2N\left[ \left(1+\frac{2}{m}\right)^m \frac{NM_m(y)}{(\alpha N)^m} + \exp\left( -\frac{2(m+2)^{-2}e^{-m}(\alpha N)^2}{N\sigma^2(y)}\right)\right] \tag{15}$$

where we select $\alpha = 1 - \varepsilon^{-1/2}$ in (14), and we define

$$\sigma^2(y) = \mathbb{E}_{p^t}\left[(r(X,y)-1)^2\right] < \infty.$$

Hence we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left[\min_{1\le\ell\le N}\left(\frac{1}{N}\sum_{k=1}^{N} r(X_k,Y_\ell)\right)^2\right]^{-1} > N^\delta\varepsilon\right)\\
&\le \frac{C_1}{N^{\gamma/2}}\int_E M_m(y)p_1^{t+1}(y)\,\nu(dy) + 2N\int_E \exp\left(-\frac{C_2N}{\sigma^2(y)}\right)p_1^{t+1}(y)\,\nu(dy)\\
&\le \frac{C_1}{N^{\gamma/2}}\int_E M_m(y)p_1^{t+1}(y)\,\nu(dy)\\
&\quad + \frac{1}{C_2N^{\gamma/2}}\int_E [\sigma^2(y)]^{1+\gamma/2}p_1^{t+1}(y)\,\nu(dy) \qquad (16)\\
&\le \frac{C3}{N^{\gamma/2}}\left(\mathbb{E}_{p^t\otimes p_1^{t+1}}[|r(X,Y)|^{2+\gamma/2}] + \mathbb{E}_{p^t\otimes p_1^{t+1}}[|r(X,Y)|^{2+\gamma}]\right) \qquad (17)
\end{aligned}$$

where inequality (16) holds since $e^{-x} < x^{-(1+\gamma/2)}$ for $x > 0$, (17) results from Jensen's inequality applied with the convex function $x^{1+\gamma/2}$, and $C_1, C_2$, $C_3$ are some constants depending on $m$ and $\alpha$. This shows that the last term converges to zero as $N \to \infty$. □

Under stronger moment conditions, we can derive the almost sure behavior of our estimator. Here, the technique used to handle $R_N$ is from Del Moral and Guionnet (1999), except that we use a moderate deviation type inequality as above instead of the Markov's inequality used by these authors. This allows us to reduce the moment condition from 6 to $(4+\gamma)$ for some $\gamma > 0$.

**Theorem 2** *Under conditions and assumptions of Theorem 1, if we replace moment condition (6) by:*

$$\mathbb{E}_{p^t\otimes p_1^{t+1}}\left[|r^t(X,Y)|^{4+\gamma}\right] < \infty \quad \textit{for some } \gamma > 0, \qquad (18)$$

*then*

$$\hat{\mathcal{H}}_N(p_1^{t+1}, p^{t+1}) \xrightarrow{a.s.} \mathcal{H}(p_1^{t+1}, p^{t+1}), \quad \textit{as } N \to \infty.$$

*Proof.* Using decomposition (9) and the fact that $J_N$ goes to zero almost surely as $N$ goes to infinity, it suffices to show that $R_N = \mathcal{O}(J_N)$ a.s. For this purpose, we use the decomposition of $|R_N|$ given in Del Moral and Guionnet (1999) p. 290:

$$|R_N| \le \frac{1}{2}|J_N| \max_{1\le\ell\le N}\left|\frac{1}{N}\sum_{k=1}^{N} r(X_k,Y_\ell) - 1\right| \max_{1\le\ell\le N}\xi_\ell. \qquad (19)$$

For any $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{1\leq\ell\leq N}\left|\frac{1}{N}\sum_{k=1}^{N} r(X_k, Y_\ell) - 1\right| > \varepsilon\right) \leq N\mathbb{P}\left(\left|\sum_{k=1}^{N} r(X_k, Y) - 1\right| > N\varepsilon\right)$$

Therefore, by conditioning in $Y$ and using, as in the previous proof, the moderate deviation type inequality with $m = 3 + \gamma/2$, we get:

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq\ell\leq N}\left|\frac{1}{N}\sum_{k=1}^{N} r(X_k, Y_\ell) - 1\right| > \varepsilon\right) \\
&\leq \frac{C_1}{N^{1+\gamma/2}}\int_E M_m(y) p_1^{t+1}(y)\,\nu(dy) \\
&\quad + \frac{1}{C_2 N^{1+\gamma/2}}\int_E [\sigma^2(y)]^{2+\gamma/2} p_1^{t+1}(y)\,\nu(dy) \qquad (20) \\
&\leq \frac{C_3}{N^{1+\gamma/2}}\left(\mathbb{E}_{p^t\otimes p_1^{t+1}}[|r(X,Y)|^{3+\gamma/2}] + \mathbb{E}_{p^t\otimes p_1^{t+1}}[|r(X,Y)|^{4+\gamma}]\right), \qquad (21)
\end{aligned}$$

where (21) comes as before from Jensen's inequality applied with the convex function $x^{2+\gamma/2}$. Using Borel-Cantelli's lemma it comes:

$$\max_{1\leq\ell\leq N}\left|\frac{1}{N}\sum_{k=1}^{N} r(X_k, Y_\ell) - 1\right| \xrightarrow{a.s.} 0 \quad \text{as } N \to \infty, \qquad (22)$$

and from the above it is not difficult to check (as in Del Moral and Guionnet (1999) p. 291) that:

$$\max_{1\leq\ell\leq N} \xi_\ell \xrightarrow{a.s.} 1 \quad \text{as } N \to \infty, \qquad (23)$$

which concludes the proof. □

We conclude in the same way that $\hat{\mathcal{H}}_N(p^t)$ is a consistent estimate of $\mathbb{E}_{p^t}(\log p^t)$. More precisely:

**Corollary 1** *If, for all $t \geq 0$, $r^t(x,y) = q^t(x,y)/p^{t+1}(y)$ is non-degenerate and satisfies $\mathbb{E}_{p^t\otimes p^{t+1}}\left[|r^t(X,Y)|^{2+\gamma}\right] < \infty$, for some $\gamma > 0$, and $p^t$ satisfies $\mathbb{E}_{p^{t+1}}[|\log p^{t+1}(Y)|^2] < \infty$, then:*

$$\begin{aligned}
\hat{\mathcal{H}}_N(p^{t+1}) &\xrightarrow{\mathbb{P}} \mathcal{H}(p^{t+1}), \quad \textit{as } N \to \infty, \\
\sqrt{N}(\hat{\mathcal{H}}_N(p^{t+1}) - \mathcal{H}(p^{t+1})) &\xrightarrow{d} \mathcal{N}(0, \Sigma^t), \quad \textit{as } N \to \infty,
\end{aligned}$$

*where $\Sigma^t = \mathrm{var}_{p^{t+1}}[\log p^{t+1}] + \mathrm{var}_{p^t}[R(X)]$, and $R(x) = \mathbb{E}_{p^{t+1}}[r^t(x,Y)]$.*
*If, in addition, $\mathbb{E}_{p^t\otimes p^{t+1}}\left[|r^t(X,Y)|^{4+\gamma}\right] < \infty$, then*

$$\hat{\mathcal{H}}_N(p^{t+1}) \xrightarrow{a.s.} \mathcal{H}(p^{t+1}), \quad \textit{as } N \to \infty.$$

*Proof.* Since we are using two independent samples $\mathbf{X}^{t+1}$ and $\tilde{\mathbf{X}}^{t+1}$, it suffices to follow step by step the proves of theorem 1 and theorem 2, just replacing $\mathbf{Y}^{t+1}$ by $\tilde{\mathbf{X}}^{t+1}$ and $p_1^{t+1}$ by $p^{t+1}$. □

# 3 Checking stability of Markovian systems

The aim of this section is to propose a simulation procedure to diagnose the stability of Markovian dynamical systems with respect to various initial conditions. When the kernel of such a dynamical system is accessible to some analytic computations, and when some critical constants (e.g., the spectral gap, the Dobrushin's coefficient or a "Foster Lyapunov" drift criteria) can be estimated, then a stability property or a convergence rate may be deduced from these estimations. However, this is not the case in many practical situations, and the exploratory method we propose here may often be the only way to gain information about the stability properties of the system.

We denote by $X = (X^t)_{t\geq 0}$, and $Y = (Y^t)_{t\geq 0}$ the two processes issued from a (eventually) inhomogeneous discrete time Markov chain with transition kernel density $q^t(\cdot,\cdot)$ at time $t$, and two different initial conditions $x_0$ and $x_1$. We assume that the transition densities are analytically known. The successive density functions of the Markov chain $X$ (resp. $Y$) are given by the recurrence formula:

$$p_i^{t+1}(y) = \int_E p_i^t(x) q^t(x,y)\,\nu(dx), \quad i = 0 \quad (\text{resp. } i = 1). \tag{24}$$

We propose to use, for each time $t$, a consistent and asymptotically normal estimator of the Kullback-Leibler information between $p_0^t$ and $p_1^t$. The Kullback-Leibler information between two densities $f$ and $g$ on $E$ is defined by:

$$\mathcal{K}(f,g) = \int_E \log\left(\frac{f(x)}{g(x)}\right) f(x)\,\nu(dx) = \mathbb{E}_f[\log(f)] - \mathbb{E}_f[\log(g)].$$

It is well known that $\mathcal{K}(f,g) \geq 0$ for any $f$ and $g$, and $\mathcal{K}(f,g) = 0$ if and only if $f = g$ almost everywhere. In addition, Del Moral, Ledoux, and Miclo (2003) prove that, under good stability properties, $\mathcal{K}(p_0^t, p_1^t)$ is geometrically decreasing in $t$ (see also Miclo (1997) for entropy considerations and finite Markov chains). Hence the Kullback information is a relevant tool to check that a dynamical system has such stability properties, by comparing densities pairwise for various initial conditions. Practically, we propose a graphical monitoring consisting in plotting our estimate $t \to \mathcal{K}(p_i^t, p_j^t)$, since the stability is related to the decreasing rate (in $t$) of our consistent estimator. This technic may also highlight models for which the initial condition influence the behavior of the densities $p_i^t$ for a long run.

Obviously, using the estimators (3) and (4) theoretically studied in section 2, we are able to estimate $\mathbb{E}_{p_i^t}[\log(p_j^t)]$ and $\mathbb{E}_{p_i^t}[\log(p_i^t)]$ for $i$ and $j = 0, 1$, and $i \neq j$, so that we can estimate $\mathcal{K}(p_i^t, p_j^t)$, for $i \neq j$. For example,

$$\hat{\mathcal{K}}(p_1^t, p_0^t) = \hat{\mathcal{H}}_N(p_1^t) - \hat{\mathcal{H}}_N(p_1^t, p_0^t).$$

### 3.1 An illustrative example

To illustrate the method presented in section 3 on a simple and classical example (where everything is known), we apply it to an AR(1) Gaussian model,

$$X_t = \rho X_{t-1} + \varepsilon_t, \tag{25}$$

where $(\varepsilon_t)_{t\geq 0}$ is a sequence of centered Gaussian noises with variance $\sigma^2$, and $\rho \in (0,1)$. This model with the initial condition $x_0$ can also be written

$$X_t = \rho^t x_0 + \sum_{k=0}^{t-1} \rho^k \varepsilon_{t-k}. \tag{26}$$

The transition kernel is itself Gaussian, with density $q(x,y) = \phi_{\sigma^2}(y-\rho x)$, where $\phi_{\sigma^2}(\cdot)$ is the p.d.f. of $\mathcal{N}(0,\sigma^2)$. It is easy to see from (26) that at time $t$,

$$X_t \sim \mathcal{N}\left(\rho^t x_0\,,\, \frac{1-\rho^{2t}}{1-\rho^2}\sigma^2\right), \tag{27}$$

and $Y_t$ is normally distributed with mean $\rho^t x_1$ and the same variance.

Condition (7) holds and is easy to check, and we show that the moment condition (6) is satisfied for such a process. Here, we consider the estimation of $\mathcal{H}(p_1^t, p_0^t)$, the other case being handled similarly. After calculations we obtain the following decomposition for all $0<\rho<1$ and $\gamma>0$:

$$\mathbb{E}_{p_0^t \otimes p_1^{t+1}}\left[\left|r^t(X,Y)\right|^{2+\gamma}\right] = \int_{\mathbb{R}^2} C\,\exp(L(x,y))\exp(-Q(x,y))\,dx\,dy, \tag{28}$$

where $C$ is a positive constants, $L(x,y)$ is a linear function of $(x,y)$ and $Q(x,y) = a\,x^2 + b\,xy + c\,y^2$, where $a, b$ and $c$ are strictly positive constants depending on $\rho, \gamma$ and $t$. Hence the integral (28) is finite if $4ac - b^2 > 0$. Some analytical computations (with the help of *Mathematica*) show that this condition reduces to (up to a positive constant)

$$\left(ac - \frac{b^2}{4}\right) \propto \frac{(1-\rho^2)\left[1-\rho^2(1+\gamma)^2 + \rho^{2(1+t)}\gamma(2+\gamma)\right]}{(1-\rho^{2t})(1-\rho^{2(1+t)})} > 0.$$

Hence, for all $0<\rho<1$, we can find $\gamma>0$ small enough such that $4ac-b^2>0$, which ensures (6).

Below, we detail the situations to which we applied our stability control method. We first simulated the case where the moment condition (6) for the consistency of our estimator $\hat{\mathcal{H}}_N(p_1^t) - \hat{\mathcal{H}}_N(p_1^t, p_0^t)$ and the stability condition for the AR(1) are both satisfied. We then tried situations where these conditions are not satisfied, to illustrate the fact that our estimator may detect unstable or even explosive phenomena. Hence in that case, our estimator may be used as an empirical indicator,

even when we cannot prove theoretical properties. We also provide in the figures below an approximation of the true Kullback information, which could be computed in this simple situation via numerical integration with *Mathematica*. For all these models, the variance is $\sigma^2 = 4$.

Example 1: homogeneous stable case. This case corresponds to situations where $\rho \in (0, 1)$. Figure 1 compares the behavior of the estimated Kullback-Leibler information when $\rho$ is small enough (strongly mixing, *left*), and when $\rho$ is near 1 (slowly mixing, *right*). It is clear that the decreasing rate of our estimator shows the stability of the process and gives some indication about its mixing behavior. Note also that for these settings, running $N = 30$ chains was enough to obtain such accurate estimations.

Example 2: homogeneous unstable case (Figure 2). This case corresponds to $\rho = 1$. Here, we cannot prove the theoretical properties of our estimator, but it nevertheless indicates clearly the unstable behavior. We choose $x_0 = -20$ and $x_1 = 20$ for the starting locations, since we wanted the true Kullback information to stabilize away from zero for reasonable values of time. In this situation, $p_0$ and $p_1$ are Gaussian distributions with fixed means $x_0$ and $x_1$, and variances increasing with $t$. Surprisingly, good results are already obtained with a limited number of $N = 50$ chains here, even if better results are obtained with more chains, since more particles (chains) are needed to explore the enlarging support of these densities (for example, the mass is located roughly in $[-100; 100]$ at time $t = 200$ here).

Example 3: homogeneous explosive case (Figure 3). This case corresponds to situations where $\rho > 1$. As for the unstable case, we cannot prove theoretical properties for our estimator, but its explosive behavior may be used as an indicator of the unstability of the underlying process. In this situation, $p_0$ and $p_1$ are Gaussian distributions with means going respectively to $+\infty$ and $-\infty$ and increasing variances as $t \to \infty$. Hence we keep the initial conditions at $-5$ and $+5$ to limit the explosive behaviour. As expected, much more parallel chains were needed to obtain a reasonable estimate of the true Kullback information, since with this setting more particles are needed to explore the enlarging support (e.g., the mass is already located roughly in $[-100; 100]$ at time $t = 30$ here).

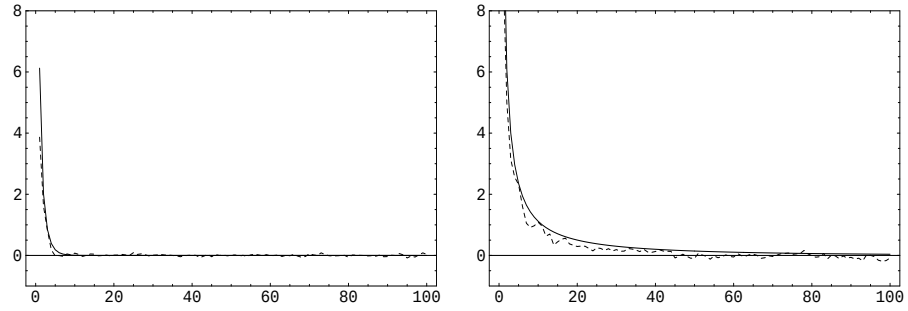


Figure 1: *Plot of the true (solid line) and estimated (dashed line)* $\mathcal{K}(p_1^t, p_0^t)$ *against* $t$. *Left: Stability behavior for* $\rho = 0.7$, $N = 30$. *Right: Stability behavior for* $\rho = 0.99$, $N = 50$. *The initial conditions are* $x_0 = -5$, $x_1 = 5$.

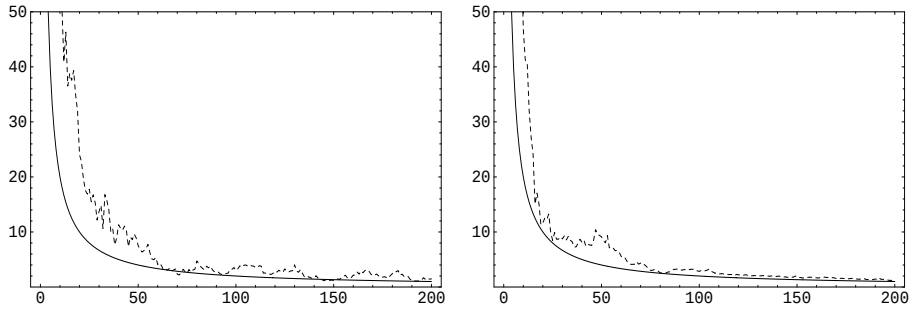


Figure 2: *Plot of the true (solid line) and estimated (dashed line)* $\mathcal{K}(p_1^t, p_0^t)$ *against* $t$. *Unstability behavior for* $\rho = 1$, *and initial conditions* $x_0 = -20$, $x_1 = 20$. *With* $N = 50$ *chains (left), and* $N = 300$ *chains (right).*

## 4 Entropy and the Central Limit Theorem

Another possible application where our estimate can be useful is the estimation of the gap in entropy between a properly normalized sum of i.i.d. random variables and the gaussian limit given by the Central Limit Theorem (CLT). Consider

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i,$$

where $X_1, X_2, \ldots$ are i.i.d. random variables taking values in $\mathbb{R}^d$, with common probability density $f$. We assume for convenience that the $X$'s are centered ($\mathbb{E}_f(X) = 0$). The CLT states that, if $X$ has moment of order two, then

$$Y_n \xrightarrow{d} \mathcal{N}_d(0, \Sigma), \quad \text{when } n \to \infty,$$

where $\Sigma$ is the variance-covariance matrix of $X$ and $\mathcal{N}_d$ is the $d$-dimensional gaussian distribution.

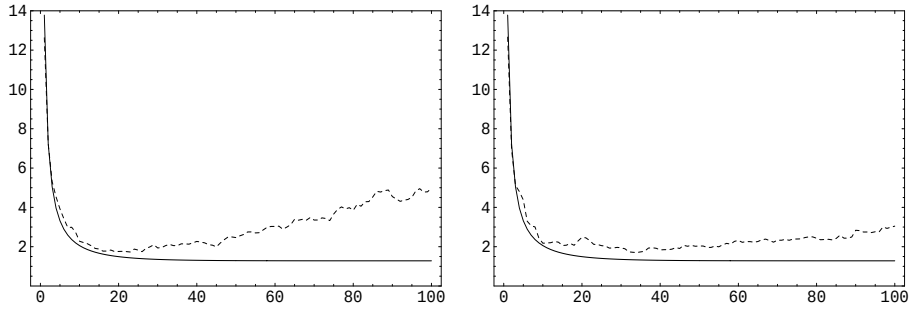


Figure 3: *Plot of the true (solid line) and estimated (dashed line)* $\mathcal{K}(p_1^t, p_0^t)$ *against* $t$*. Unstability behavior for* $\rho = 1.05$*, and initial conditions* $x_0 = -5$*,* $x_1 = 5$*.* $N = 300$ *chains (left) and* $N = 600$ *chains (right).*

In actual applications (e.g., statistical inference), the CLT is used to propose confidence regions for the parameters of interest related to the distribution of $X$, or to approximate a distribution of a test statistic under the null hypothesis, to bound the type I error of the test. The question of the sample size ($n$) needed to apply the CLT accurately, is crucial and not clear for non standard (and multi-dimensional) distributions. Upper bounds for the convergence rate in the TCL are typically given by the Berry-Esséen theorem, but these bounds are known to be highly pessimistic (see, e.g., Seoh and Hallin, 1997).

Our technique can be used to compute by simulation the "distance to normality", in the entropy sense, of the pdf of $Y_n$. Moreover, it is important to point out that this question also motivates current theoretical developments (see, e.g., Ball *et al* (2003) and references therein).

To apply our method, we only need to know how to simulate $X \sim f$, and the analytical expression of $f$. Actually, we can represent the successive densities of $Y_n$, $n \geq 0$, by the marginals of the inhomogeneous Markov chain

$$Y_{n+1} = \sqrt{\frac{n}{n+1}} Y_n + \frac{1}{\sqrt{n+1}} X_{n+1},$$

with transition density kernel at time $n$

$$q^n(x, y) = \sqrt{n+1} f(\sqrt{n+1}\, y - \sqrt{n}\, x). \tag{29}$$

Using the estimate $\hat{\mathcal{H}}_N(p^n)$ given by (3), we can estimate at "time" $n$ $\mathbb{E}_{p^n}(\log p^n)$ by double Monte Carlo integration. In addition, using, e.g., the first set of $N$ i.i.d. copies of $Y_n$, that we denote here by $(Y_{n,1}, \ldots, Y_{n,N})$, we can also estimate $\mathbb{E}_{p^n}(\log \phi_\Sigma)$, where $\phi_\Sigma$ is the multivariate gaussian density with variance-covariance matrix $\Sigma$, by the standard Monte Carlo integration

$$\frac{1}{N} \sum_{\ell=1}^{N} \log\left(\phi_\Sigma(Y_{n,\ell})\right) \overset{a.s.}{\to} \mathbb{E}_{p^n}(\log \phi_\Sigma) \quad \text{as } N \to \infty,$$

so that we can consistently estimate the Kullback distance between the pdf of $Y_n$ and the limiting gaussian:

$$\mathcal{K}(p^n, \phi_\Sigma) = \mathbb{E}_{p^n}(\log p^n) - \mathbb{E}_{p^n}(\log \phi_\Sigma).$$

This distance is also a good measure since, as stated in Ball *et al* (2003), it bounds the $L^1$-norm.

### 4.1 Examples

To illustrate the rate in the CLT in entropy, we have chosen four simple univariate distributions for $f$, to show different levels of difficulties. Intuitively, the more $f$ is "far" from the gaussian, the more time ($n$) it will take for the pdf of $Y_n$ to resemble to a normal distribution. Hence we have chosen centered distributions with non symmetric densities, or heavy tails; namely:

$$X \sim \mathcal{N}(0,1), \quad X \sim \mathcal{U}_{[-20;20]}, \quad X \sim t(3), \quad X \sim \alpha\phi_{\mu_1,\sigma_1^2} + (1-\alpha)\phi_{\mu_2,\sigma_2^2}.$$

The first model, $X \sim \mathcal{N}(0,1)$, is the stationary process, since $Y_n$ is just a linear combination of independent gaussian r.v.'s (no CLT needed in this case), i.e. $Y_n \sim \mathcal{N}(0,1)$ for any $n \geq 1$. The second model is a uniform distribution, which is know to converge quickly to a gaussian-like distribution. The third model is the Student distribution with 3 degrees of freedom, i.e. the heaviest tailed Student distribution satisfying the CLT. The fourth model is a 2-component mixture of gaussian distributions. We have tested two different sets of parameters for the mixture, which differ by their variance and level of skweness:

$$\begin{aligned} M_1 &: \alpha = \frac{1}{2}, \quad \mu_1 = -8, \quad \sigma_1^2 = 1, \quad \mu_2 = 8, \quad \sigma_2^2 = 4, \\ M_2 &: \alpha = \frac{1}{2}, \quad \mu_1 = -20, \quad \sigma_1^2 = 1, \quad \mu_2 = 20, \quad \sigma_2^2 = 16. \end{aligned}$$

In addition to these four models, for which the CLT holds, we have simulated a fifth model, for which the CLT *does not* hold: we choose for $f$ a Cauchy distribution (i.e. $t(1)$), with density

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R},$$

for which $\mathbb{E}(X)$ and $\mathrm{var}(X)$ does not exist.

**Checking moment conditions** To be sure that our estimates are consistent, we have to check that the conditions (6) and (7) of theorem 1, related to moments of $q^n$ and $p^n$, hold. Actually, condition (7) is easy to check in most cases, such as the present models. The difficulty is to check (6). For example, it is easy to see that (6) holds when the common distribution $f$ of the $X$'s is any bounded distribution

with compact support, such as the uniform. When the $n$-fold convolution product of $f$ is accessible to calculation, checking the validity of conditions (6) is feasible. For the stationary model for example ($X \sim \mathcal{N}(0,1)$),the density of $Y_n$, $p^n$, is the pdf of the gaussian $\mathcal{N}(0,1)$, but the kernel $q^n$ still depends on $n$, so that condition (6) is satisfied for $0 < \gamma < 2(\sqrt{(n+1)/n} - 1)$. For the two other models, the direct calculation of $f^{*n}$ is not simple, so that (6) is hard to verify. Here, we have merely cheked that the numerical integration involved in (6), computed with *Mathematica*, converges for small values of $\gamma$ and the first steps (in $n$). Moreover, the simulations show a decreasing to zero behavior for the estimated Kullback distances, as expected.

**Results summary** We have estimate the Kullback distance using various number of i.i.d. copies of the r.v. $Y_n$, from $N = 200$ up to $N = 5000$ to illustrate the asymptotic behavior of our estimates of entropy (the computing time for these simulations (in C) is rather short, at least in these simple cases).

The models resulting in largest $\hat{\mathcal{K}}(p^n, \phi_\Sigma)$ for small values of $n$ (i.e. far from the gaussian) were, in decreasing order, $M_2$, $M_1$, $t(3)$, $\mathcal{U}_{[-20;20]}$ and $\mathcal{N}(0,1)$. The 5 models tested required different time (sample size) for $Y_n$ to get close to the gaussian. If we choose as a criterion the first time at which $\hat{\mathcal{K}}(p^n, \phi_\Sigma)$ crosses the value 0, then we obtain the following results for $N = 5000$:

| Model | Time to CLT |
|---|---|
| $t(3)$ | 88 |
| $M_2$ | 28 |
| $M_1$ | 15 |
| $\mathcal{U}_{[-20;20]}$ | 9 |
| $\mathcal{N}(0,1)$ | 1 |

Due to the different scales of $\hat{\mathcal{K}}(p^n, \phi_\Sigma)$ for small $n$ depending on the different models, we could not represent al the tested models in the same plot. We have chosen to show various figures illustrating the rate of the CLT for each model, and the effect of $N$.

Figure 4 shows the two mixture models, for which the estimated Kullback distance for $Y_2$ are the largest, together with the uniform model which is a typically "easy" model. The effect of high value of $N$ on the quality of the estimates is clear.

Figure 5 illustrates the fact that, for the gaussian model, $\hat{\mathcal{K}}(p^n, \phi_\Sigma)$ is merely random fluctuation around zero. We have also plotted the uniform model for comparison purpose. Even if for the uniform model, the pdf of $Y_n$ goes rather quickly to the gaussian, the difference whith the stationary model is clearly visible. This figure also illustrate the effect of $N$ on the precision of the estimates.

Figure 6 shows the impressive difference between the Student $t(3)$ model and the others: the heavy-tailed Student distribution requires more time than the other to

achieve the CLT. For these plots, we have simulated $Y_n$ up to $n = 100$ to observe the slow decreasing behavior of $\hat{\mathcal{K}}(p^n, \phi_\Sigma)$.

**The Cauchy model** To simulate the model for which the CLT is not valid, we have chosen to normalize the limiting gaussian with the empirical variance of the $N$ simulated values of $Y_n$ at step $n$, since $\Sigma = var(X)$ does not exist in this case. Figure 7 shows that the estimated Kullback distance does not decrease to zero, even during the first $n = 500$ steps. Indeed, despite the fact that the conditions of theorem 1 do not hold here, our estimate behaves properly and can be used to indicate that the CLT is not valid.

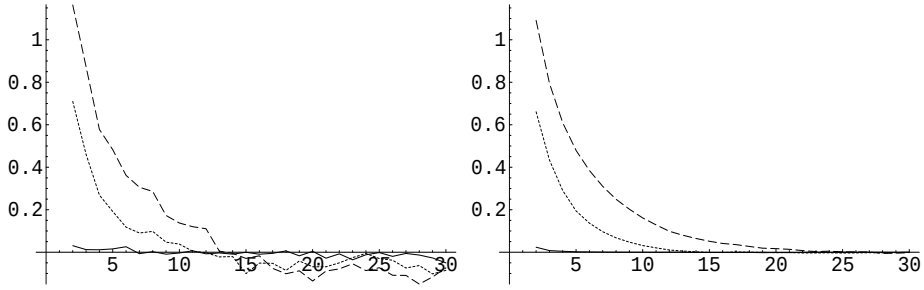


Figure 4: *Models* $\mathcal{U}_{[-20;20]}$ *(solid),* $M_1$ *(dotted) and* $M_2$ *(dashed) for* $N = 200$ *(left), and* $N = 5000$ *(right).*

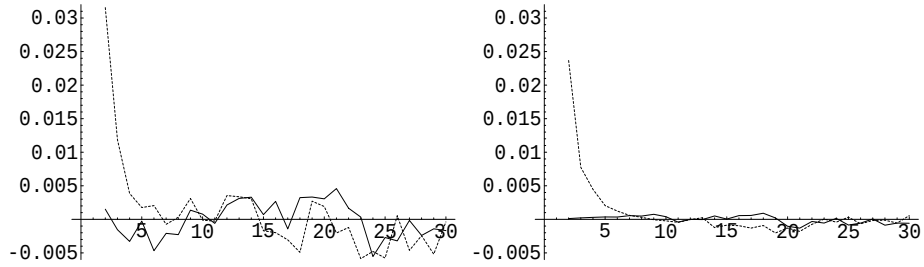


Figure 5: *Models* $\mathcal{N}(0,1)$ *(solid) and* $\mathcal{U}_{[-20;20]}$ *(dotted) for* $N = 1000$ *(left) and* $N = 5000$ *(right).*

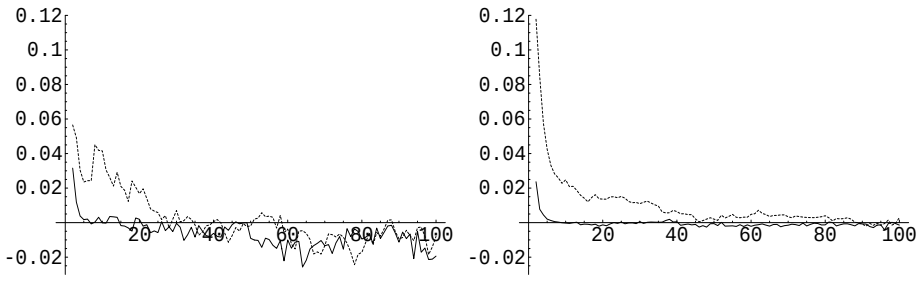


Figure 6: *Models* $\mathcal{U}_{[-20;20]}$ *(solid) and* $t(3)$ *(dotted) for* $N = 1000$ *(left) and* $N = 5000$ *(right).*

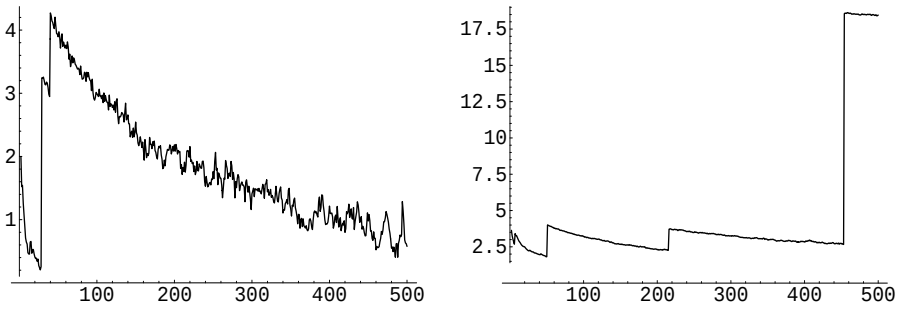


Figure 7: *Cauchy Model for* $N = 100$ *(left) and* $N = 1000$ *(right).*

# References

[1] Ahmad, I. A. and Lin, P. E., A nonparametric estimation of the entropy for absolutely continuous distributions, *IEEE Trans. Inform. Theory*, **36**, 688–692, 1989.

[2] Ball, K., Barthe, F. and Naor, A., Entropy jumps in the presence of a spectral gap, *Duke Mathematical Journal*, **119**, 1, 41–63, 2003.

[3] Del Moral P., Guionnet A., Central Limit Theorem for Nonlinear Filtering and Interacting Particle Systems, *Annals of Applied Probability*, **9**, no 2, 275–297, 1999.

[4] Del Moral P., Ledoux M., Miclo, L., Contraction properties of Markov kernels. *Probab. Theory and Related Fields*, **126**, pp. 395–420, 2003.

[5] Eggermont, P. P. B. and LaRiccia, V. N., Best asymptotic Normality of the Kernel Density Entropy Estimator for Smooth Densities, *IEEE trans. Inform. Theory*, **45**, no. 4, 1321–1326, 1999.

[6] Fuk, D. Kh., and Nagaev, S. V., Probability Inequalities for Sums of Independent Random Variables, *Th. Probab. Appl.* **16**, 643–660, **21**, 875, 1971, 1976.

[7] Lehmann, E. L., *Nonparametrics: Statistical Methods based on Rank*, Holden-Day series in Probability and Statistics. Mc Graw-Hill, 1975.

[8] Levit, B. Y., Asymptotically efficient estimation of nonlinear functionals, *Probl. Inform. Trans.*, **41**, 204–209, 1978.

[9] Miclo, L., Remarques sur l'hypercontractivité et l'évolution de l'entropie des chaînes de Markov finies, Séminaire de Probabilités XXXI, Lecture Notes in Mathematics, Springer, 136–168, 1997.

[10] Petrov, V., *Limit Theorems of Probability Theory*, Oxford Science Pubications, 1995.

[11] Seoh, M. and Hallin, M., *When does Edgeworth beat Berry and Esséen? Numerical evaluations of Edgeworth expansions*, Journal of Stat. Plan. Inference, **63**, 1, 19–38, 1997.

[12] Serfling, R. J., *Approximation theorems of Mathematical statistics*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc, New York, 1980.

[13] Tsybakov, A. B. and Van Der Meulen, E. C., Root $t$ consistent estimators of entropy for densities with unbounded support, *Scand. J. Statist.*, **23**, 75–83, 1994.