Probab. Theory Relat. Fields 110, 427–439 (1998)

Probability Theory and Related Fields

Measure concentration for a class of random processes

Katalin Marton★

Mathematical Institute of the Hungarian Academy of Sciences, P.O. Box 127, H-1364 Budapest, Hungary
e-mail: marton@math-inst.hu

Received: 6 May 1996 / In revised form: 29 September 1997

Summary. Let $X = \{X_i\}_{i=-\infty}^{\infty}$ be a stationary random process with a countable alphabet $\mathscr{X}$ and distribution q. Let $q^{\infty}(\cdot|x_{-k}^{0})$ denote the conditional distribution of $X^{\infty} = (X_1, X_2, \ldots, X_n, \ldots)$ given the k-length past:

$$q^{\infty}\left(\cdot|x_{-k}^{0}\right) = \mathrm{dist}\left(X^{\infty}|X_{-k}^{0} = x_{-k}^{0}\right) \ .$$

Write $d(\hat{x}_1, x_1) = 0$ if $\hat{x}_1 = x_1$, and $d(\hat{x}_1, x_1) = 1$ otherwise. We say that the process X admits a joining with finite distance u if for any two past sequences $\hat{x}_{-k}^{0} = (\hat{x}_{-k+1}, \ldots, \hat{x}_0)$ and $x_{-k}^{0} = (x_{-k+1}, \ldots, x_0)$, there is a joining of $q^{\infty}(\cdot|\hat{x}_{-k}^{0})$ and $q^{\infty}(\cdot|x_{-k}^{0})$, say $\mathrm{dist}(\hat{X}_0^{\infty}, X_0^{\infty}|\hat{x}_{-k}^{0}, x_{-k}^{0})$, such that

$$E\left\{\sum_{i=1}^{\infty} d\left(\hat{X}_i, X_i\right)|\hat{x}_{-k}^{0}, x_{-k}^{0}\right\} \leq u \ .$$

The main result of this paper is the following inequality for processes that admit a joining with finite distance:

Theorem. *Let q^n denote the distribution of $X^n = (X_1, X_2, \ldots, X_n)$. Then for any distribution p^n on $\mathscr{X}^n$*

★This work was supported in part by the grants OTKA 1906 and T 016386 of the Hungarian Academy of Sciences, and by MTA-NSF project 37.

The author wishes to thank R. Burton and P.C. Shields for eliminating a false statement from the first draft of this paper. Thanks are due to P.C. Shields and B. Tóth for useful conversations on the subject.

$$\bar{d}(p^n, q^n) \le (u+1)\sqrt{\tfrac{1}{2n} D(p^n \| q^n)} \ ,$$

where D denotes informational divergence.

The significance of this bound is that it implies a measure concentration inequality. We are able, at least theoretically, to compute u for Markov chains.

We also prove that the existence of finite distance joining is implied by a condition frequently used in the theory of 1-dimensional Gibbs measures.

Mathematics Subject Classification (1991): 60F10, 60G10, 60J10

1. Introduction

Let $X = \{X_i\}_{i=-\infty}^{\infty}$ be a stationary process with a countable alphabet $\mathcal{X}$ and distribution q. If $\{x_j\}_{j \in J}$ is a (possibly infinite) sequence of elements of $\mathcal{X}$, and the interval $(i, m]$ belongs to J then we denote by x_i^m the subsequence $(x_{i+1}, x_{i+2}, \ldots, x_m)$; i or m may be $-\infty$ resp. ∞. If the lower index is missing then 0 is understood.

We also use the notation $\mathcal{X}_i^m$ for the space of sequences x_i^m, where, again, i or m may be infinite. If q is a probability measure on the space of doubly infinite sequences $x_{-\infty}^{\infty}$ then we use the notation q_i^m to denote the induced measure on $\mathcal{X}_i^m$. We denote by $q_{l+1}(\cdot | x_i^l)$ the distribution $\mathrm{dist}(X_{l+1} | X_i^l = x_i^l)$, and by $q_l^{l+m}(\cdot | x_i^l)$ the distribution $\mathrm{dist}(X_l^{l+m} | X_i^l = x_i^l)$.

We denote by $\bar{d}$ the normed Hamming distance on $\mathcal{X}^n \times \mathcal{X}^n$:

$$\bar{d}(x^n, y^n) = \frac{1}{n} d(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i) \ ,$$

$$d(x_i, y_i) = 1 \qquad \text{if} \qquad x_i \ne y_i, \qquad 0 \qquad \text{otherwise} \ .$$

We say that the process X, or the distribution q, has the blowing-up property if for any $\varepsilon > 0$ there are $\delta > 0$ and n_0 such that for $n \ge n_0$ and any $A \subset \mathcal{X}^n$ with $q^n(A) \ge \exp(-n\delta)$, the ε-neighborhood of A has measure almost 1. I.e.,

$$q^n(A) \ge \exp(-n\delta) \Rightarrow q^n([A]_\varepsilon) \ge 1 - \varepsilon \ , \tag{1.1}$$

where $[A]_\varepsilon$ is the ε-neighborhood of A:

$$[A]_\varepsilon = \left\{ y^n \in \mathcal{X}^n : \bar{d}(x^n, y^n) \le \varepsilon \text{ for some } x^n \in A \right\} \ .$$

Note that (1.1) can be replaced by the seemingly stronger implication

$$q^n(A) \geq \exp(-n\delta) \Rightarrow q^n([A]_\varepsilon) \geq 1 - \exp(-n\delta) \ , \tag{1.1'}$$

which can be seen by applying (1.1) to both A and the complement of $[A]_\varepsilon$. (The δ of (1.1′) is not the same as that of (1.1).) The implication (1.1′) can be written in the following symmetric form:

$$q^n(A) \geq \exp(-n\delta), \quad q^n(B) \geq \exp(-n\delta) \Rightarrow \bar{d}(A,B) \leq \varepsilon \ .$$

Equivalently, the blowing-up property for q means that there exists a function $\varphi(\delta)$ with $\lim_{\delta\to 0} \varphi(\delta) = 0$ such that

$$\bar{d}(A,B) \leq \varphi\left(\frac{1}{2n}\log\frac{1}{q^n(A)}\right) + \varphi\left(\frac{1}{2n}\log\frac{1}{q^n(B)}\right) \ .$$

Definition. *We say that X, or q, has the measure concentration property, if for any $A,B \subset \mathcal{X}^n$ we have*

$$\bar{d}(A,B) \leq c \cdot \left(\sqrt{\frac{1}{2n}\log\frac{1}{q^n(A)}} + \sqrt{\frac{1}{2n}\log\frac{1}{q^n(B)}}\right)$$

for some constant c.

Thus measure concentration is much stronger than blowing-up. In this paper we focus on measure concentration.

Ahlswede et al. [1] proved that if q is i.i.d. (independent identically distributed) then it does have the blowing-up property. In fact, the proof given in [1] yielded also the measure concentration property for the i.i.d. case. Later the measure concentration phenomenon was extensively studied for i.i.d. processes. C.f. [2] and McDiarmid [3], where the best constant for the i.i.d. case ($c = 1$) was first obtained. See also Talagrand's survey papers [4] and [5] where new proofs, lots of applications and a large bibliography are given. – Proofs of measure concentration, based on the use of informational divergence, were given in the author's papers [6] and [7]. In [7] also some processes with memory were considered.

There is a simple but powerful inequality by Pinsker between variational distance and informational divergence. (See later.) The extension of this inequality to one between $\bar{d}$-distance and informational divergence was the basis of the proofs of measure concentration given in [6] and [7].

If p and r are probability distributions on $\mathcal{X}$ then $|p - r|$ will denote their variational distance (divided by 2).

Let p^n and q^n be two distributions on $\mathcal{X}^n$; their $\bar{d}$-distance is

$$\bar{d}(p^n, q^n) = \min E\bar{d}\left(\hat{X}^n, X^n\right) \ ,$$

where the min is taken over all joint distributions with marginals $q^n = \operatorname{dist} X^n$ and $p^n = \operatorname{dist} \hat{X}^n$. The distance $\bar{d}(p^n, q^n)$ is a natural generalization of $|p - r|$, since

$$|p - r| = \min \Pr\{\hat{X} \neq X\} \ ,$$

where the min is taken over all joint distributions $\operatorname{dist}(\hat{X}, X)$ having marginals $p = \operatorname{dist} \hat{X}$ and $r = \operatorname{dist} X$.

If p and r are two probability distributions on $\mathcal{X}$ then the informational divergence of p with respect to q is

$$D(p\|r) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{r(x)} \ .$$

Thus the informational divergence of p^n with respect to q^n is

$$D(p^n\|q^n) = \sum_{x^n \in \mathcal{X}^n} p^n(x^n) \log \frac{p^n(x^n)}{q^n(x^n)} \ .$$

Now we recall

Pinsker's inequality, (C.f. [8], [9].). *Let p and r be two distributions on* $\mathcal{X}$; *then p and r admit a joining* $\operatorname{dist}(U, V)$ *satisfying*

$$\Pr\{U \neq V\} = |p - r| \leq \sqrt{\tfrac{1}{2} D(p\|r)} \ .$$

In [6], [7] a similar inequality was proved between $\bar{d}(p^n, q^n)$ and $\frac{1}{n} D(p^n\|q^n)$ for the case when q is i.i.d. In [7], this inequality was generalized to the case of mixing Markov chains, and also for a class of processes q with very fast and uniform decay of dependence. Namely, for a class of processes q, [7] established

$$\bar{d}(p^n, q^n) \leq c \cdot \sqrt{\tfrac{1}{2n} D(p^n\|q^n)} \ , \tag{1.2}$$

for any probability measure p^n on $\mathcal{X}^n$, where the constant c depends on the behavior of the transition probability function $q^\infty(\cdot|x^0_{-\infty}) = \operatorname{dist}\left(X^\infty_0 | X^0_{-\infty} = x^0_{-\infty}\right)$. E.g., if q is a Markov measure with a transition matrix satisfying

$$|\operatorname{dist}(X_1|X_0 = x) - \operatorname{dist}(X_1|X_0 = y)| \leq 1 - a \ , \tag{1.3}$$

then the constant can be taken $1/a$. If q is i.i.d. then $c = 1$ is good, but a smaller c can be given if c is allowed to depend on the distribution q.

By the following lemma, (1.2) implies measure concentration for q.

Lemma 1. *If there is a constant c such that for any distribution* p^n *on* $\mathcal{X}^n$, *the inequality* (1.2) *holds, then q has the measure concentration property with the same constant c.*

(Bobkov and Götze [16] proved recently that (1.2) is also necessary for measure concentration, although possibly with a value of c different from the one we used in the definition of measure concentration.)

Proof of Lemma 1. Assume (1.2). Consider two sets $A, B \subset \mathcal{X}^n$. Define a distribution p^n, associated with the set A as follows:

$$p^n(x^n) = \begin{cases} q^n(x^n)/q^n(A) & x^n \in A \\ 0, & \text{otherwise} \end{cases} ,$$

i.e., p^n is q^n conditioned on A. Define similarly the distribution r^n associated with B.

Then

$$\frac{1}{n} D(p^n \| q^n) = \frac{1}{n} \log \frac{1}{q^n(A)} .$$

By our assumption, this implies

$$\bar{d}(p^n, q^n) \le c \cdot \sqrt{\frac{1}{2n} \log \frac{1}{q^n(A)}} .$$

Similarly,

$$\bar{d}(r^n, q^n) \le c \cdot \sqrt{\frac{1}{2n} \log \frac{1}{q^n(B)}} .$$

Since p^n and r^n are concentrated on A and B, respectively, it follows that

$$\bar{d}(A, B) \le \bar{d}(p^n, r^n) \le c \cdot \left(\sqrt{\frac{1}{2n} \log \frac{1}{q^n(A)}} + \sqrt{\frac{1}{2n} \log \frac{1}{q^n(B)}} \right) . \qquad \square$$

The aim of this paper is to give a sufficient condition for (1.2), and thereby for measure concentration. The condition we give both generalizes and improves the main theorem of [7]. The improvement concerns improving the constant in (1.2). Even for Markov chains satisfying (1.3) the constant can be improved. The process X is always assumed to be stationary.

We shall use the following concept introduced by Eberlein [10].

Definition. *We say that the process X, or the measure q, admits a joining of finite distance u if for any k and any two past sequences $\hat{x}_{-k}^0$ and x_{-k}^0 of positive probability there is a joining of $q^\infty(\cdot|\hat{x}_{-k}^0)$ and $q^\infty(\cdot|x_{-k}^0)$, say* $\mathrm{dist}(\hat{X}_0^\infty, X_0^\infty | \hat{x}_{-k}^0, x_{-k}^0)$, *such that*

$$E\left\{\sum_{i=1}^{\infty} d\left(\hat{X}_i, X_i\right) | \hat{x}_{-k}^0, x_{-k}^0\right\} \leq u \ . \tag{1.4}$$

(Eberlein called such processes Very Weak Bernoulli of order $1/n$*.)*

Eberlein proved that if the process X admits a joining of finite distance, and f is a real valued function on $\mathscr{X}$ then the process $\{f(X_i)\}$ satisfies the central limit theorem (under some quite natural additional conditions).

Our main result is the following.

Theorem 2. *If the process* X *admits a joining of finite distance* u *then*

$$\bar{d}(p^n, q^n) \leq (u+1)\sqrt{\tfrac{1}{2n} D(p^n \| q^n)} \tag{1.5}$$

for any distribution p^n *on* $\mathscr{X}^n$.

We can give sufficient conditions for the existence of finite-distance joining in terms of ergodic properties of q.

The following theorem asserts that a condition frequently used in the theory of 1-dimensional Gibbs measures implies the existence of finite-distance joining. We need the following notation:

$$\gamma_k = \sup_N \sup_{x_{-N}^0, y_{-N}^0 : y_{-k}^0 = x_{-k}^0} \left| q\left(\cdot | x_{-N}^0\right) - q\left(\cdot | y_{-N}^0\right) \right| \ .$$

Theorem 3. *Assume that* $q(x_1 | x_{-\infty}^0)$ *is bounded from below, and* $\sum_{k=1}^{\infty} \gamma_k < \infty$. *Then* q *admits a joining of finite distance, and, consequently, has the measure concentration property.*

Finally, the following result of Goldstein [11] on maximal coupling can be used to prove the existence of a finite distance joining. We use this result as formulated in Lindvall's book [12, formula (14.1), p. 99].

Let $Y^\infty = (Y_1, Y_2, \ldots)$ and $Z^\infty = (Z_1, Z_2, \ldots)$ be (non-stationary) random processes with values in $\mathscr{X}$, and distribution p^∞ and r^∞, respectively. Write

$$p_n^\infty = \mathrm{dist}\left(Y_n^\infty\right), \quad r_n^\infty = \mathrm{dist}\left(Z_n^\infty\right) \ .$$

It is a trivial consequence of the definition of variational distance that for any joining $\mathrm{dist}(Y^\infty, Z^\infty)$ of p^∞ and r^∞, and any $n \geq 0$

$$\Pr\left\{Y_n^\infty \neq Z_n^\infty\right\} \geq \left|p_n^\infty - r_n^\infty\right|, \quad \text{all} \quad n \geq 0 \ .$$

Goldstein's Theorem. *There exists a joining* $\mathrm{dist}(Y^\infty, Z^\infty)$ *of* p^∞ *and* r^∞ *such that*

$$\Pr\left\{Y_n^\infty \neq Z_n^\infty\right\} = \left|p_n^\infty - r_n^\infty\right|, \quad \text{all} \quad n \geq 0 \ . \tag{1.6}$$

A joining that satisfies (1.6) is called maximal. It is clear that for a maximal joining of $\mathrm{dist}(Y^\infty)$ and $\mathrm{dist}(Z^\infty)$

$$\sum_{i=1}^{\infty} Ed(Y_i, Z_i) \le \sum_{n=0}^{\infty} \left| p_n^\infty - r_n^\infty \right| \ .$$

Let us apply Goldstein's theorem to the distributions $q^\infty(\cdot|\hat{x}_{-k}^0)$ and $q^\infty(\cdot|x_{-k}^0)$, where $\hat{x}_{-k}^0$ and x_{-k}^0 are two fixed past sequences.

Proposition 4. *Assume that there is a constant u such that for any k and any two past sequences* $\hat{x}_{-k}^0$ *and* x_{-k}^0

$$\sum_{n=0}^{\infty} \left| q_n^\infty(\cdot|\hat{x}_{-k}^0) - q_n^\infty(\cdot|x_{-k}^0) \right| \le u \ .$$

Then q admits a joining of finite distance u.

Proposition 4 specializes to Markov chains as follows. For Markov chains the existence of maximal coupling was proved by Griffeath [12]. For a stationary Markov chain $\{X_i\}$ and fixed $j, k \in \mathscr{X}$, consider the distributions $q^\infty(\cdot|j) = \mathrm{dist}(X_0^\infty|X_0 = j)$ and $q^\infty(\cdot|k) = \mathrm{dist}(X_0^\infty|X_0 = k)$.

We have in this case

$$\left| q_n^\infty(\cdot|j) - q_n^\infty(\cdot|k) \right| = \left| q_{(n+1)}(\cdot|j) - q_{(n+1)}(\cdot|k) \right| \ ,$$

where $q_{(n+1)}(\cdot|j) = \mathrm{dist}(X_{n+1}|X_0 = j)$.

Proposition 4′. *If q is the distribution of a stationary Markov chain then q admits a joining of finite distance u with*

$$u = \sup_{j,k} \sum_{n=1}^{\infty} \left| q_{(n)}(\cdot|j) - q_{(n)}(\cdot|k) \right| \ .$$

It is clear that if a Markov chain satisfies (1.3) then $u + 1 \le 1/a$. But $u + 1$ can be substantially smaller than $1/a$. In the case of finite-state time-reversible Markov chains one can bound $u + 1$, using spectral theory of stochastic matrices. Define

$$\lambda = \max |\lambda_i| \ ,$$

where λ_i ranges over the eigenvalues of the transition matrix corresponding to non-constant eigenfunctions. It is well known that a Markov chain is mixing if and only if $\lambda < 1$. Let $s = \{s(x)\}$ denote the stationary distribution of the Markov chain. Then (c.f. [14], Proposition 3)

$$\left|q_{(n)}(\cdot|j) - q_n\right| \le \frac{1}{\sqrt{s(j)}} \cdot \lambda^n \quad .$$

(We used a weaker but simpler bound than that in [14].) It follows that

$$u + 1 \le \frac{2}{\min_j \sqrt{s(j)}} \cdot \frac{1}{1-\lambda} \quad .$$

(Similar bounds also exist in the non-reversible case [15].)

Obviously there exist time-reversible mixing Markov chains, say with uniform stationary distribution, whose transition matrix does not satisfy (1.3). For such Markov chains Theorem 4 can be applied. It is also clear that, by a small perturbation of the transition matrix of such a Markov chain we can get a transition matrix satisfying (1.3) with an arbitrarily small $a > 0$, but with second largest eigenvalue still bounded away from 1. In this case $u + 1$ will be much smaller than $1/a$.

The proof of Theorems 2 and 3 is given in Section 2.

2. Proof of the theorems

We shall prove Theorem 2 in the following stronger form.

Theorem 2′. *If X admits a joining of finite distance u then for any $k \ge 0$, any fixed past sequence x_{-k}^0, and any distribution p^n on $\mathscr{X}^n$*

$$\bar{d}\left(p^n, q^n\left(\cdot|x_{-k}^0\right)\right) \le (u+1)\sqrt{\tfrac{1}{2n} D\left(p^n \| q^n\left(\cdot|x_{-k}^0\right)\right)} \quad . \tag{2.1}$$

To get Theorem 2 from Theorem 2′, we apply it for $k = 0$; then x_{-k}^0 is the empty sequence, and so (1.5) is a special case of (2.1).

Remark. The inequality

$$\bar{d}\left(p^n, q^n\left(\cdot|x_{-\infty}^0\right)\right) \le (u+1)\sqrt{\tfrac{1}{2n} D\left(p^n \| q^n\left(\cdot|x_{-\infty}^0\right)\right)}$$

(for almost all $x_{-\infty}^0$) would not be enough to get Theorem 2, since the integral of the right-hand-side with respect to $q_{-\infty}^0$ may be larger than $\sqrt{\frac{1}{2n} D(p^n \| q^n)}$.

We introduce the following notation. Let us fix a past sequence x_{-k}^0, and let $\hat{X}^n$ and X^n denote random sequences distributed according to p^n and $q^n\left(\cdot|x_{-k}^0\right)$, respectively.

We have then

$$\Pr\{X_1^n|X_1=x_1\} = \frac{q^n\left(x^n|x_{-k}^0\right)}{q_1\left(x_1|x_{-k}^0\right)} = \frac{q_{-k}^n\left(x_{-k}^n\right)}{q_{-k}^1\left(x_{-k}^1\right)} = q\left(x_1^k|x_{-k}^1\right) \ ,$$

i.e.,

$$\mathrm{dist}\left(X_1^n|X_1=x_1\right) = q_1^n\left(\cdot|x_{-k}^1\right) \ .$$

Let us put $p_1 = \mathrm{dist}\hat{X}_1$. Moreover, for a fixed $\hat{x}_1 \in \mathscr{X}$ write

$$p_1^n(\cdot|\hat{x}_1) = \mathrm{dist}\left(\hat{X}_1^n|\hat{X}_1 = \hat{x}_1\right) \ .$$

We shall use the following important identity for expansion of divergence.

$$D\left(p^n\|q^n\left(\cdot|x_{-k}^0\right)\right) = D\left(p_1\|q_1\left(\cdot|x_{-k}^0\right)\right) + \sum_{\hat{x}_1} p_1(\hat{x}_1) D\left(p_1^n(\cdot|\hat{x}_1)\|q_1^n\left(\cdot|x_{-k}^0\hat{x}_1\right)\right) \ , \tag{2.2}$$

where $x_{-k}^0\hat{x}_1$ is the sequence obtained by appending $\hat{x}_1$ after x_{-k}^0.

Proof of Theorem 2′. We prove (2.1) by induction on n. For $n = 1$ it follows from Pinsker's inequality. (For *any* k!)

Assume that (2.1) holds for $n-1$ and *any* k. Fix a k and a sequence x_{-k}^0. Let $\hat{X}^n$ and X^n denote random sequences distributed according to p^n and $q^n(\cdot|x_{-k}^0)$, respectively. Our goal is to define a joining

$$\mathrm{dist}\left(\hat{X}^n, X^n\right)$$

of the distributions p^n and $q^n(\cdot|x_{-k}^0)$ so that $E\bar{d}\left(\hat{X}^n, X^n\right)$ be possibly small.

First we define a joint distribution $\mathrm{dist}(\hat{X}^n, Y_1{}^n)$, where Y_1^n is a random sequence $(Y_2, \ldots, Y_n)$ of length $n-1$. For a fixed value $\hat{x}_1$ of $\hat{X}_1$, define

$$\mathrm{dist}\left(Y_1^n|\hat{X}_1 = \hat{x}_1\right) = q_1^n\left(\cdot|x_{-k}^0\hat{x}_1\right) \ .$$

Since q is stationary, we can use the induction hypothesis for the sequence $x_{-k}^0\hat{x}_1$ instead of x_{-k}^0, to get a joining

$$\mathrm{dist}\left(\hat{X}^n, Y_1^n|\hat{X}_1 = \hat{x}_1\right)$$

that achieves

$$E\left\{\bar{d}\left(\hat{X}_1^n, Y_1^n\right)|\hat{X}_1 = \hat{x}_1\right\} \le (u+1)\sqrt{\frac{1}{2(n-1)} D\left(p_1^n(\cdot|\hat{x}_1)\|q_1^n\left(\cdot|x_{-k}^0\hat{x}_1\right)\right)} \ .$$

This implies, by the concavity of the square root function,

$$E\bar{d}\left(\hat{X}_1^n, Y_1^n\right) \le (u+1)\sqrt{\frac{1}{2(n-1)}\sum_{\hat{x}_1} p_1(\hat{x}_1)D\left(p_1^n(\cdot|\hat{x}_1)\|q_1^n\left(\cdot|x_{-k}^0\hat{x}_1\right)\right)} \ . \tag{2.3}$$

Now we join the distributions $\mathrm{dist}(\hat{X}^n, Y_1^n)$ and $\mathrm{dist}X^n = q^n(\cdot|x_{-k}^0)$. Define $\mathrm{dist}(\hat{X}_1, X_1)$ so as to achieve

$$\Pr\left\{\hat{X}_1 \ne X_1\right\} = \left|p_1 - q_1\left(\cdot|x_{-k}^0\right)\right| \ .$$

Further, if $\hat{x}_1$ and x_1 are possible values of $\hat{X}_1$ and X_1, then we can take a joining

$$\mathrm{dist}\left(Y_1^n, X_1^n | \hat{X}_1 = \hat{x}_1, X_1 = x_1\right)$$

satisfying

$$E\left\{\sum_{i=2}^n d(Y_i, X_i)|\hat{X}_1 = \hat{x}_1, X_1 = x_1\right\} \le u \cdot d(\hat{x}_1, x_1) \ . \tag{2.4}$$

Indeed, if $\hat{x}_1 = x_1$ then

$$\mathrm{dist}\left(Y_1^n|\hat{X}_1 = \hat{x}_1\right) = q_1^n\left(\cdot|x_{-k}^0\hat{x}_1\right) = q_1^n\left(\cdot|x_{-k}^1\right) = \mathrm{dist}\left(X_1^n|X_1 = x_1\right) \ ,$$

and so we get 0 for the expected distance; if $\hat{x}_1 \ne x_1$ then the minimum expected distance is $\le u$, since q admits a joining of finite distance u.

Now take any joint distribution

$$\mathrm{dist}\left(\hat{X}^n, Y_1^n, X^n\right)$$

for which $\mathrm{dist}(\hat{X}^n, Y_1^n)$, $\mathrm{dist}(\hat{X}_1, X_1)$ and $\mathrm{dist}(Y_1^n, X_1^n|\hat{X}_1, X_1)$ are as described above. Then we have, using (2.3) and (2.4),

$$\begin{aligned}
&E\bar{d}\left(\hat{X}^n, X^n\right) \\
&\quad \le \frac{1}{n}\Pr\{\hat{X}_1 \ne X_1\} + \frac{n-1}{n}E\bar{d}(\hat{X}_1^n, Y_1^n) + \frac{1}{n}E\sum_{i=2}^n d(Y_i, X_i) \\
&\quad \le \frac{1}{n}\Pr\left\{\hat{X}_1 \ne X_1\right\} + \frac{n-1}{n}(u+1) \\
&\qquad \times \sqrt{\frac{1}{2(n-1)}\sum_{\hat{x}_1} p_1(\hat{x}_1)D\left(p_1^n\left(\ \cdot\ |\hat{x}_1\right)\|q_1^n\left(\ \cdot\ |x_{-k}^0\hat{x}_1\right)\right)} \\
&\qquad + \frac{u}{n}\Pr\{\hat{X}_1 \ne X_1\} = (u+1)\cdot\left\{\frac{1}{n}\Pr\left\{\hat{X}_1 \ne X_1\right\} + \frac{n-1}{n}\right. \\
&\qquad \left. \times \sqrt{\frac{1}{2(n-1)}\sum_{\hat{x}_1} p_1(x_1)D\left(p_1^n\left(\ \cdot\ |\hat{x}_1\right).\|q_1^n\left(\ \cdot\ |x_{-k}^0\hat{x}_1\right)\right.}\right\} \ .
\end{aligned} \tag{2.5}$$

Now we use Pinsker's inequality to get

$$\Pr\{\hat{X}_1 \neq X_1\} \leq \sqrt{\tfrac{1}{2} D\big(p_1 \| q_1\big(\cdot|x_{-k}^0\big)} \ .$$

Substituting this into (2.5), we get the bound

$$E\bar{d}\Big(\hat{X}^n, X^n\Big) \leq (u+1) \cdot \Bigg[\frac{1}{n}\sqrt{\tfrac{1}{2} D\big(p_1 \| q_1\big(\cdot|x_{-k}^0\big)}$$
$$+ \frac{n-1}{n}\sqrt{\frac{1}{2(n-1)} \sum_{\hat{x}_1} p_1(\hat{x}_1) D\big(p_1^n(\cdot|\hat{x}_1) \| q_1^n\big(\cdot|x_{-k}^0 \hat{x}_1\big)\big)}\Bigg] \ .$$

By the concavity of the square root function, and using the expansion of divergence (formula (2.2)), the right-hand side of the last formula can be continued to

$$\leq (u+1) \cdot \sqrt{\tfrac{1}{2n} D\big(p^n \| q^n\big(\cdot|x_{-k}^0\big)} \ .$$

We have proved (2.1) for n, so the induction step is completed, and the proof also. □

Discussion. In the proof of Theorem 2 we only used the following consequence of the existence of finite-distance joining:

$$E\left\{\sum_{i=1}^{\infty} d\big(\hat{X}_i, X_i\big) \,|\hat{x}_{-k}^0, x_{-k}^0\right\} \leq u \ ,$$

provided $\hat{x}_{-k}^0$ and x_{-k}^0 differ only in the last (i.e., 0'th) symbol. But we do not know whether this assumption is indeed weaker than the existence of finite-distance joining. If we only had assumed

$$E\left\{\sum_{i=1}^{\infty} d\big(\hat{X}_i, X_i\big) \,|\hat{x}_{-\infty}^0, x_{-\infty}^0\right\} \leq u$$

for $\hat{x}_{-\infty}^{-1} = x_{-\infty}^{-1}$, *that* condition would not have been enough to prove Theorem 2. (It is enough to prove the inequality

$$\bar{d}\big(p^n, q^n\big(\cdot|x_{-\infty}^0\big)\big) \leq (u+1)\sqrt{\tfrac{1}{2n} D\big(p^n \| q^n\big(\cdot|x_{-\infty}^0\big)\big)}$$

with probability 1.)

Proof of Theorem 3. Since $q(x_1|x_{-\infty}^0)$ is bounded from below, we have $\gamma_1 < 1$, and

$$\prod_{i=1}^{\infty}(1-\gamma_i) > 0 \ .$$

Write

$$w = 1 - \prod_{i=1}^{\infty}(1 - \gamma_i) \ .$$

We have $w < 1$.

Let us fix two past sequences $\hat{x}_{-\infty}^0, x_{-\infty}^0 \in \mathscr{X}_{-\infty}^0$, and define a joining

$$\mathrm{dist}\left(\hat{X}^\infty, X^\infty | \hat{x}_{-\infty}^0, x_{-\infty}^0\right) \tag{2.6}$$

of $q^\infty(\cdot|\hat{x}_{-\infty}^0)$ and $q^\infty(\cdot|x_{-\infty}^0)$ as follows.

Let $\mathrm{dist}(\hat{X}_1, X_1|\hat{x}_{-\infty}^0, x_{-\infty}^0)$ achieve

$$\Pr\{\hat{X}_1 \neq X_1|\hat{x}_{-\infty}^0, x_{-\infty}^0\} = \left|q_1\left(\cdot|\hat{x}_{-\infty}^0\right) - q_1\left(\cdot|x_{-\infty}^0\right)\right| \leq \gamma_1 \ .$$

Assume that $\mathrm{dist}(\hat{X}^i, X^i|\hat{x}_{-\infty}^0, x_{-\infty}^0)$ is already defined. Fix sequences $\hat{x}^i, x^i \in \mathscr{X}^i$. Let us append the sequences $\hat{x}^i, x^i$ to $\hat{x}_{-\infty}^0$ and $x_{-\infty}^0$, respectively, and denote the resulting sequences by $\hat{x}_{-\infty}^i$ and $x_{-\infty}^i$. Now define

$$\mathrm{dist}\left(\hat{X}_{i+1}, X_{i+1}|\hat{x}_{-\infty}^i, x_{-\infty}^i\right)$$

so as to achieve

$$\Pr\{\hat{X}_{i+1} \neq X_{i+1}\big|\hat{x}_{-\infty}^i, x_{-\infty}^i\} = \left|q_{i+1}\left(\cdot|\hat{x}_{-\infty}^i\right) - q_{i+1}\left(\cdot|x_{-\infty}^i\right)\right| \leq \gamma_j \ ,$$

where $j \leq i$ is the largest integer such that

$$\hat{x}_{i-j}^i = x_{i-j}^i \ .$$

Thus we have defined the joining (2.6).

Define

$$d^\infty(\hat{x}^\infty, x^\infty) = \sum_{i=1}^{\infty} d(\hat{x}_i, x_i) \ .$$

Let us estimate $E\{d^\infty(\hat{X}^\infty, X^\infty)|\hat{x}_{-\infty}^0, x_{-\infty}^0\}$. We have

$$\Pr\left\{\hat{X}^\infty = X^\infty\big|\hat{x}_{-\infty}^0, x_{-\infty}^0\right\} \geq \prod_{i=1}^{\infty}(1 - \gamma_i) \ ,$$

i.e.,

$$\Pr\left\{d^\infty\left(\hat{X}^\infty, X^\infty\right) \geq 1\big|\hat{x}_{-\infty}^0, x_{-\infty}^0\right\} \leq w \ .$$

Consider two sequences $\hat{x}^\infty$ and x^∞ such that $\hat{x}^\infty \neq x^\infty$, and let k be the first index for which $\hat{x}_k \neq x_k$. Then we have

$$\Pr\left\{\hat{X}_k^\infty = X_k^\infty\Big|\hat{x}_{-\infty}^0, x_{-\infty}^0, \hat{X}^k = \hat{x}^k, X^k = x^k\right\} \geq \prod_{i=1}^{\infty}(1 - \gamma_i) \ .$$

This implies that

$$\Pr\left\{d^\infty\left(\hat{X}^\infty, X^\infty\right) \geq 2 \middle| \hat{x}^0_{-\infty}, x^0_{-\infty}, d^\infty\left(\hat{X}^\infty, X^\infty\right) \geq 1\right\} \leq w \ .$$

It can be proved similarly that for any $l \geq 1$

$$\Pr\left\{d^\infty\left(\hat{X}^\infty, X^\infty\right) \geq l+1 \middle| \hat{x}^0_{-\infty}, x^0_{-\infty}, d^\infty\left(\hat{X}^\infty, X^\infty\right) \geq l\right\} \leq w \ ,$$

i.e.,

$$Pr\left\{d^\infty\left(\hat{X}^\infty, X^\infty\right) \geq l \middle| \hat{x}^0_{-\infty}, x^0_{-\infty}\right\} \leq w^l \ .$$

Since $w < 1$, this implies that $E\{d^\infty(\hat{X}^\infty, X^\infty)|\hat{x}^0_{-\infty}, x^0_{-\infty}\}$ is bounded. □

References

[1] Ahlswede, R., Gács, P., Körner, J.: Bounds on conditional probabilities with applications in multi-user communication, Zeitschrift f. Wahrscheinlichkeitstheorie u. Verw. Geb. **34**, 157–177 (1976)

[2] Unconditional symmetric sets in n-dimensional normed spaces, Israel J. of Math. **37**, 3–20 (1980)

[3] McDiarmid, C.: On the method of bounded differences, in Surveys in Combinatorics London Mathematical Society Lecture Notes, Vol 141 (J. Simons, ed.), Cambridge University Press, London-New York, pp. 148–188 (1989)

[4] Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces, Publications of IHES (1995)

[5] Talagrand, M.: A new look at independence, Annals of Prob. **24**, 1–34 (1996)

[6] Marton, K.: A simple proof of the blowing-up lemma, IEEE Trans. on Information Theory **32**, 445–446 (1986)

[7] Marton, K.: Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration, Annals of Probability **24**, 857–866 (1996)

[8] Pinsker, M.S.: Information and information stability of random variables and processes, Holden-Day, San Francisco, 1964

[9] Csiszár, I., Körner, J.: Information theory: Coding theorems for discrete memoriless systems, Academic Press, Inc., New York, London etc., 1981

[10] Eberlein, E.: Strong approximation of Very Weak Bernoulli processes, Zeitschrift für Wahrscheinlichkeitstheorie u. verw. Geb. **62**, 17–37 (1983)

[11] Goldstein, S.: Maximal coupling, Zeitschrift für Wahrscheinlichkeitstheorie u. verw. Geb. **46**, 193–204 (1979)

[12] Lindvall, T.: Lectures on the coupling method, Wiley & Sons, Inc., New York, 1992

[13] Griffeath, D.: A maximal coupling for Markov chains, Zeitschrift für wahrscheinlichkeitstheorie u. verw. Geb. **31**, 36–61 (1975)

[14] Diaconis, P., Stroock, D.: Geometric bounds for eigen values of Markov chains, Annals of Applied Probability, **1**, 36–61 (1991)

[15] Fill, J.A.: Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, Annals of Applied Probability, **1**, 62–87 (1991)

[16] Bobkov, S., Götze, F.: Exponential integrability and transportation cost related to logarithmic Sobolev inequalities, preprint (1997)