Bernoulli **20**(4), 2014, 1698–1716
DOI: 10.3150/13-BEJ537

Convergence rate and concentration inequalities for Gibbs sampling in high dimension

NENG-YI WANG[1] and LIMING WU[2]

[1]*Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100190, Beijing, China. E-mail: wangnengyi@amss.ac.cn*
[2]*Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100190, Beijing, China and Laboratoire de Math. CNRS-UMR 6620, Université Blaise Pascal, 63177 Aubière, France. E-mail: Li-Ming.Wu@math.univ-bpclermont.fr*

The objective of this paper is to study the Gibbs sampling for computing the mean of observable in very high dimension – a powerful Markov chain Monte Carlo method. Under the Dobrushin's uniqueness condition, we establish some explicit and sharp estimate of the exponential convergence rate and prove some Gaussian concentration inequalities for the empirical mean.

Keywords: concentration inequality; coupling method; Dobrushin's uniqueness condition; Gibbs measure; Markov chain Monte Carlo

1. Introduction

Let μ be a Gibbs probability measure on E^N with dimension N very big, that is,

$$\mu\big(\mathrm{d}x^1,\ldots,\mathrm{d}x^N\big)=\frac{\mathrm{e}^{-V(x^1,\ldots,x^N)}}{\int\cdots\int_{E^N}\mathrm{e}^{-V(x^1,\ldots,x^N)}\pi(\mathrm{d}x^1)\cdots\pi(\mathrm{d}x^N)}\pi\big(\mathrm{d}x^1\big)\cdots\pi\big(\mathrm{d}x^N\big),$$

where π is some σ-finite reference measure on E. Our purpose is to study the Gibbs sampling – a Markov chain Monte Carlo method (MCMC in short) for approximating μ. In fact, even for the simplest case where $E=\{+,-\}$, as the denominator contains an exponential number of terms and each of them may be very big or small for high dimension, it is very difficult to model μ.

Let $\mu_i(\cdot|x)$ $(x=(x^1,\ldots,x^N)\in E^N)$ be the regular conditional distribution of x^i knowing $(x^j, j\neq i)$ under μ; and $\bar{\mu}_i(\mathrm{d}y|x)=(\prod_{j\neq i}\delta_{x^j}(\mathrm{d}y^j))\otimes\mu_i(\mathrm{d}y^i|x)$ (product measure), where $\delta_{\cdot}$ is the Dirac measure at the point $\cdot$. We see that

$$\mu_i\big(\mathrm{d}x^i|x\big)=\frac{\mathrm{e}^{-V(x^1,\ldots,x^N)}}{\int_E\mathrm{e}^{-V(x^1,\ldots,x^N)}\pi(\mathrm{d}x^i)}\pi\big(\mathrm{d}x^i\big),$$

which is a one-dimensional measure, easy to be realized in practice.

The idea of the Gibbs sampling consists in approximating μ via iterations of the one-dimensional conditional distributions $\mu_i, i=1,\ldots,N$. It is described as follows. Given a starting configuration $x_0=(x_0^1,\ldots,x_0^N)\in E^N$, let $(X_n; n\geq 0)$ be a non-homogeneous Markov chain

defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P}_{x_0})$, such that

$$X_0 = x_0 \qquad \text{a.s.},$$

and given

$$X_{kN+i-1} = x = \left(x^1, \ldots, x^N\right) \in E^N \qquad (k \in \mathbb{N}, 1 \le i \le N),$$

then $X^j_{kN+i} = x^j$ for $j \neq i$ and the conditional law of X^i_{kN+i} is $\mu_i(\mathrm{d}y^i|x)$. In other words, the transition probability at step $kN + i$ is:

(1) $\mathbb{P}(X_{kN+i} \in \mathrm{d}y | X_{kN+i-1} = x) = \bar{\mu}_i(\mathrm{d}y|x)$.

Therefore,

(2) $\mathbb{P}(X_{kN} \in \mathrm{d}y | X_{(k-1)N} = x) = (\bar{\mu}_1 \cdots \bar{\mu}_N)(x, \mathrm{d}y) =: P(x, \mathrm{d}y)$.

Finally, the Gibbs sampling is the time-homogeneous Markov chain $(Z_k = X_{kN}, k = 0, 1, \ldots)$, whose transition probability is P.

This MCMC algorithm is known sometimes as *Gibbs sampler* in the literature (see Winkler [31], Chapters 5 and 6). It is actively used in statistical physics, chemistry, biology and throughout the Bayesian statistics (a sentence taken from [3]). It was used by Zegarlinski [34] as a tool for proving the logarithmic Sobolev inequality for Gibbs measures, see also the second named author [33] for a continuous time MCMC.

Our purpose is two-fold:

(1) the convergence rate of P^k to μ;
(2) the concentration inequality for $\mathbb{P}(\frac{1}{n}\sum_{k=1}^n f(Z_k) - \mu(f) \ge t), t > 0$.

Question (1) is a classic subject. Earlier works by Meyn and Tweedie [21] and Rosenthal [25,26] are based on the Harris ergodicity theorem (minorization condition together with the drift condition in the non-compact case). Quantitative estimates in the Harris ergodic theorem are obtained more recently by Rosenthal [27] and Hairer and Mattingly [11]. But as indicated by Diaconis, Khare and Saloff-Coste [2,3], theoretical results obtained from the Harris theorem are very far (even too far) from the convergence rate of numerical simulations in high dimension (e.g., $N = 100$). That is why Diaconis, Khare and Saloff-Coste [2,3] use new methods and tools (orthogonal polynomials, stochastic monotonicity and coupling) for obtaining sharp estimates of $\|\nu P^k - \mu\|_{\mathrm{TV}}$ (total variation norm) for several special models in Bayesian statistics, with E^N replaced by $E \times \Theta$, a space of two different components.

For the question (1), our tool will be the Dobrushin interdependence coefficients (very natural and widely used in statistical physics), instead of the minorization condition in the Harris theorem or the special tools in [2,3]. Our main idea consists in constructing an appropriate coupling well adapted to the Dobrushin interdependence coefficients, close to that of Marton [20].

To the second question, we will apply the recent theory on transport inequalities (see Marton [17], Ledoux [13,14], Villani [30], Gozlan and Léonard [10] and references therein), and our approach is inspired from Marton [18,20] and Djellout, Guillin and Wu [4] for dependent tensorization of transport inequalities.

See [8,24,31] for Monte Carlo algorithms and diverse applications, and [12] for concentration inequalities of general MCMC under the positive curvature condition.

This paper is organized as follows. The main results are stated in the next section, and we prove them in Section 3.

2. Main results

Throughout the paper, E is a Polish space with the Borel σ-field $\mathscr{B}$, and d is a metric on E such that $d(x, y)$ is lower semi-continuous on E^2 (so d does not necessarily generate the topology of E). On the product space we consider the L^1-metric

$$d_{L^1}(x, y) := \sum_{i=1}^{N} d\big(x^i, y^i\big), \qquad x, y \in E^N.$$

If $d(x^i, y^i) = 1_{x^i \neq y^i}$ is the discrete metric on E, d_{L^1} becomes the Hamming distance on E^N, a good metric for concentration in high dimension as shown by Marton [17,18].

2.1. Dobrushin's interdependence coefficient

Let $\mathcal{M}_1(E)$ be the space of probability measures on E and

$$\mathcal{M}_1^d(E) := \left\{ \nu \in \mathcal{M}_1(E); \int_E d(x_0, x)\nu(\mathrm{d}x) < \infty \right\}$$

($x_0 \in E$ is some fixed point). Given $\nu_1, \nu_2 \in \mathcal{M}_1^d(E)$, the L^1-Wasserstein distance between ν_1, ν_2 is given by

$$W_{1,d}(\nu_1, \nu_2) := \inf_{\pi} \int\!\!\int_{E\times E} d(x, y)\pi(\mathrm{d}x, \mathrm{d}y), \tag{2.1}$$

where the infimum is taken over all probability measures π on $E \times E$ such that its marginal distributions are, respectively, ν_1 and ν_2 (*coupling of ν_1 and ν_2, say*). When $d(x, y) = \mathbb{1}_{x\neq y}$ (*the discrete metric*), it is well known that

$$W_{1,d}(\nu_1, \nu_2) = \sup_{A\in\mathscr{B}} \big|\nu_1(A) - \nu_2(A)\big| = \frac{1}{2}\|\nu_1 - \nu_2\|_{\mathrm{TV}} \qquad \text{(total variation)}.$$

Recall the Kantorovich–Rubinstein duality relation [30]

$$W_{1,d}(\nu_1, \nu_2) = \sup_{\|f\|_{\mathrm{Lip}}\le 1} \int_E f d(\nu_1 - \nu_2), \qquad \|f\|_{\mathrm{Lip}} := \sup_{x\neq y} \frac{|f(x) - f(y)|}{d(x, y)}.$$

Let $\mu_i(\mathrm{d}x^i|x)$ be the given regular conditional distribution of x^i knowing $(x^j, j \neq i)$.

Throughout the paper, we assume that $\int_{E^N} d(y^i, x_0^i)\,\mathrm{d}\mu(y) < \infty$, $\mu_i(\cdot|x) \in \mathcal{M}_1^d(E)$ for all $i = 1, \ldots, N$ and $x \in E^N$, where x_0 is some fixed point of E^N, and $x \to \mu_i(\cdot|x)$ is Lipschitzian from (E^N, d_{L^1}) to $(\mathcal{M}_1^d(E), W_{1,d})$.

Define the matrix of the d-Dobrushin interdependence coefficients $C := (c_{ij})_{i,j=1,\dots,N}$

$$c_{ij} := \sup_{x=y \text{ off } j} \frac{W_{1,d}(\mu_i(\cdot|x), \mu_i(\cdot|y))}{d(x^j, y^j)}, \qquad i, j = 1, \dots, N. \tag{2.2}$$

Obviously $c_{ii} = 0$. Then the well-known Dobrushin uniqueness condition (see [5,6]) is read as

$$\text{(H1)} \quad r := \|C\|_\infty := \max_{1 \le i \le N} \sum_{j=1}^{N} c_{ij} < 1$$

or

$$\text{(H2)} \quad r_1 := \|C\|_1 := \max_{1 \le j \le N} \sum_{i=1}^{N} c_{ij} < 1.$$

By the triangular inequality for the metric $W_{1,d}$,

$$W_{1,d}\big(\mu_i(\cdot|x), \mu_i(\cdot|y)\big) \le \sum_{j=1}^{N} c_{ij} d\big(x^j, y^j\big), \qquad 1 \le i \le N. \tag{2.3}$$

2.2. Transport inequality and Bobkov–Götze's criterion

When μ, ν are probability measures, the Kullback information (or relative entropy) of ν with respect to μ is defined as

$$H(\nu|\mu) = \begin{cases} \displaystyle\int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\, \mathrm{d}\nu, & \text{if } \nu \ll \mu, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.4}$$

We say that the probability measure μ satisfies the L^1-*transport-entropy inequality* on (E, d) with some constant $C > 0$, if

$$W_{1,d}(\mu, \nu) \le \sqrt{2CH(\nu|\mu)}, \qquad \nu \in \mathcal{M}_1(E). \tag{2.5}$$

To be short, we write $\mu \in T_1(C)$ for this relation. This inequality, related to the phenomenon of measure concentration, was introduced and studied by Marton [17,18], developed subsequently by Talagrand [29], Bobkov and Götze [1], Djellout, Guillin and Wu [4] and amply explored by Ledoux [13,14], Villani [30] and Gozlan-Léonard [10]. Let us mention the following Bobkov–Götze's criterion.

Lemma 2.1 ([1]). *A probability measure μ satisfies the L^1-transport-entropy inequality on (E, d) with constant $C > 0$, that is, $\mu \in T_1(C)$, if and only if for any Lipschitzian function $F : (E, d) \to \mathbb{R}$, F is μ-integrable and*

$$\int_E \mathrm{e}^{\lambda(F - \langle F \rangle_\mu)}\, \mathrm{d}\mu \le \exp\left\{ \frac{\lambda^2}{2} C \|F\|_{\mathrm{Lip}}^2 \right\} \qquad \forall \lambda \in \mathbb{R},$$

where $\langle F\rangle_\mu = \mu(F) = \int_E F\,\mathrm{d}\mu$. *In that case*,

$$\mu\big(F - \langle F\rangle_\mu \geq t\big) \leq \exp\left\{-\frac{t^2}{2C\|F\|^2_{\mathrm{Lip}}}\right\}, \qquad t > 0.$$

Another necessary and sufficient condition for $T_1(C)$ is the Gaussian integrability of μ, see Djellout, Guillin and Wu [4]. For further results and recent progresses see Gozlan and Léonard [9,10].

Remark 2.2. Recall also that w.r.t. the discrete metric $d(x, y) = 1_{x\neq y}$, any probability measure μ satisfies $T_1(C)$ with the sharp constant $C = 1/4$ (the well known CKP inequality).

2.3. Main results

For any function $f : E^N \to \mathbb{R}$, let

$$\delta_i(f) := \sup_{x=y \text{ off } i} \frac{|f(x) - f(y)|}{d(x^i, y^i)}$$

be the Lipschitzian coefficient w.r.t. the ith coordinate x^i. It is easy to see that

$$\|f\|_{\mathrm{Lip}(d_{L^1})} = \max_{1\leq i\leq N} \delta_i(f).$$

Theorem 2.3 (Convergence rate). *Under the Dobrushin uniqueness condition* (H1), *we have*:

(a) *For any Lipschitzian function* f *on* E^N *and two initial distributions* ν_1, ν_2 *on* E^N,

$$\big|\nu_1 P^k f - \nu_2 P^k f\big| \leq r^k \max_{1\leq i\leq N} \mathbb{E}d\big(Z_0^i(1), Z_0^i(2)\big) \sum_{i=1}^N \delta_i(f), \qquad k \geq 1, \tag{2.6}$$

where $(Z_0(1), Z_0(2))$ *is a coupling of* (ν_1, ν_2), *that is, the law of* $Z_0(j)$ *is* ν_j *for* $j = 1, 2$.

(b) *In particular for any initial distribution* ν *on* E^N,

$$W_{1,d_{L^1}}\big(\nu P^k, \mu\big) \leq N r^k \max_{1\leq i\leq N} \mathbb{E}d\big(Z_0^i(1), Z_0^i(2)\big),$$

where $(Z_0(1), Z_0(2))$ *is a coupling of* (ν, μ).

By part (b) above μ is the unique invariant measure of P under the Dobrushin uniqueness condition, and $P^k(x, \cdot)$ converges exponentially rapidly to μ in the metric $W_{1,d_{L^1}}$, showing theoretically why the numerical simulations by the Gibbs sampling are very rapid.

Remark 2.4. Let us compare Theorem 2.3 with the known results in [2,3,21,25,26] on the convergence rate of the Gibbs sampling.

At first the convergence rate in those known works is in the total variation norm, not in the metric $W_{1,d_{L^1}}$. When d is the discrete metric, we have by part (b) of Theorem 2.3

$$\tfrac{1}{2}\left\|\nu P^k - \mu\right\|_{\mathrm{TV}} \le W_{1,d_{L^1}}\left(\nu P^k, \mu\right) \le N r^k \qquad \forall \nu, \forall k \ge 1.$$

Next, let us explain once again why the minorization condition in the Harris theorem does not yield accurate estimates in high dimension (see Diaconis *et al.* [2,3] for similar discussions based on concrete examples). Indeed assume that E is finite, then under reasonable assumption on $V(x^1, \ldots, x^N)$, there are constant $c > 0$ and a probability measure $\nu(\mathrm{d}y)$ such that

$$P(x, \mathrm{d}y) \ge \mathrm{e}^{-cN}\nu(\mathrm{d}y)$$

(i.e., almost the best minorization that one can obtain in the dependent case). Hence by the Doeblin theorem (the ancestor of the Harris theorem),

$$\frac{1}{2}\sup_{x\in E^N}\left\|P^k(x,\cdot) - \mu\right\|_{\mathrm{TV}} \le \left(1 - \mathrm{e}^{-cN}\right)^k.$$

So one requires at least an exponential number e^{cN} of steps for the right-hand side becoming small. Our estimate of the convergence rate is much better in high dimension, that is the good point of Theorem 2.3.

The weak point of Theorem 2.3 is that our result depends on the Dobrushin uniqueness condition, even in low dimension. If N is small, the results in [21,25,26] are already good enough. Particularly the estimates of Diaconis, Khare and Saloff-Coste [2,3] for the special space of two different components in Bayesian statistics are sharp.

We should indicate that the Dobrushin uniqueness condition is quite natural for the exponential convergence of P^k to μ with the rate r independent of N as in this theorem, since the Dobrushin uniqueness condition is well known to be sharp for the phase transition of mean field models [5–7].

Finally, our tool (Dobrushin's interdependence coefficients) is completely different from those in the known works.

Remark 2.5. As indicated by a referee, it would be very interesting to investigate the convergence rate problem under the more flexible Dobrushin–Shlosman analyticity condition (i.e., box version of Dobrushin uniqueness condition, reference [7]), but in that case we feel that we should change the algorithm: instead of μ_i, one uses the conditional distribution $\mu_I(\mathrm{d}x_I|x)$ of x_I knowing $(x_j, j \notin I)$ where I is a box containing i.

Remark 2.6. A much more classical topic is Glauber dynamics associated with the Gibbs measures in finite or infinite volume. We are content here to mention only Zegarlinski [34], Martinelli and Olivieri [16], and the lecture notes of Martinelli [15] for a great number of references.

The convergence rate estimate above will be our starting point for computing the mean $\mu(f)$, that is, to approximate $\mu(f)$ by the empirical mean $\frac{1}{n}\sum_{k=1}^n f(Z_k)$.

Theorem 2.7. *Assume*

$$r_1 := \|C\|_1 = \max_{1\le j\le N} \sum_{i=1}^{N} c_{ij} < \frac{1}{2}$$

and for some constant $C_1 > 0$,

$$\text{(H3)} \quad \forall i = 1, \dots, N, \forall x \in E^N \qquad \mu_i(\cdot|x) \in T_1(C_1).$$

(*Recall that* $C_1 = 1/4$ *for the discrete metric* $d(x, y) = 1_{x\neq y}$.) *Then for any Lipschitzian function* f *on* E^N *with* $\|f\|_{\mathrm{Lip}(d_{L^1})} \le \alpha$, *we have*:

(a)

$$\mathbb{P}_x\left(\frac{1}{n}\sum_{k=1}^{n} f(Z_k) - \frac{1}{n}\sum_{k=1}^{n} P^k f(x) \ge t\right) \le \exp\left\{-\frac{t^2(1-2r_1)^2 n}{2C_1\alpha^2 N}\right\} \qquad \forall t > 0, n \ge 1; \tag{2.7}$$

(b) *furthermore if* (H1) *holds*,

$$\begin{aligned}&\mathbb{P}_x\left(\frac{1}{n}\sum_{k=1}^{n} f(Z_k) - \mu(f) \ge \frac{M}{n} + t\right)\\ &\quad\le \exp\left\{-\frac{t^2(1-2r_1)^2 n}{2C_1\alpha^2 N}\right\} \qquad \forall t > 0, n \ge 1,\end{aligned} \tag{2.8}$$

where

$$M = \frac{r}{1-r} \max_{1\le i\le N} \int_{E^N} d\big(x^i, y^i\big)\, \mathrm{d}\mu(y) \cdot \sum_{i=1}^{N} \delta_i(f).$$

In conclusion under the conditions of this theorem, when $n \gg N$, the empirical means $\frac{1}{n}\sum_{k=1}^{n} f(Z_k)$ will approximate to $\mu(f)$ exponentially rapidly in probability with the speed n/N, with the bias not greater than M/n. The speed n/N is the correct one, as will be shown in the remark below.

We do not know whether the concentration inequality with the speed n/N still holds under the more natural Dobrushin's uniqueness condition $r_1 = \|C\|_1 < 1$. We know only that $r_1 < 1$ does not imply that P is contracting in the metric $W_{1,d_{L^1}}$, see the example in Remark 3.3.

Remark 2.8. Consider $f(x) = \frac{1}{N}\sum_{i=1}^{N} g(x^i)$ where $g : E \to \mathbb{R}$ is d-Lipschitzian with $\|g\|_{\mathrm{Lip}} = \alpha$ (the observable of this type is often used in statistical mechanics). Since $\|f\|_{\mathrm{Lip}(d_{L^1})} = \frac{\alpha}{N}$, the inequality (2.7) implies for all $t > 0, n \ge 1$,

$$\mathbb{P}_x\left(\frac{1}{nN}\sum_{k=1}^{n}\sum_{i=1}^{N} g\big(Z_k^i\big) - \mathbb{E}_x\frac{1}{nN}\sum_{k=1}^{n}\sum_{i=1}^{N} g\big(Z_k^i\big) \ge t\right) \le \exp\left\{-\frac{t^2(1-2r_1)^2 nN}{2C_1\alpha^2}\right\}, \tag{2.9}$$

which is of speed nN.

Let us show that the concentration inequality (2.7) is sharp. In fact in the free case, that is, $\mu_i(\mathrm{d}x_i|x) = \beta(\mathrm{d}x_i)$ does not depend upon $(x_j)_{j\neq i}$ and i, and μ is the product measure $\beta^{\otimes N}$. In this case $P(x,\mathrm{d}y) = \mu(\mathrm{d}y)$, in other words $(Z_k)_{k\geq 1}$ is a sequence of independent and identically distributed (i.i.d. in short) random variables valued in E^N, of common law μ. Since $r_1 = 0$ in the free case, the concentration inequality (2.9) is equivalent to the transport inequality (H3) for $\mu_i = \beta$, by Gozlan-Léonard [9]. That shows also the speed n/N in Theorem 2.7 is the correct one.

Remark 2.9. We explain now why we do not apply directly the nice concentration results of Joulin and Ollivier [12] for general MCMC. In fact under the condition that $r_1 < 1/2$, we can prove that

$$W_{1,d_{L^1}}\big(P(x,\cdot),P(y,\cdot)\big) \leq \frac{r_1}{1-r_1} d_{L^1}(x,y)$$

(by Lemma 3.2). In other words the Ricci curvature in [12] is bounded from below by

$$\kappa := 1 - \sup_{x\neq y} \frac{W_{1,d_{L^1}}(P(x,\cdot),P(y,\cdot))}{d_{L^1}(x,y)} \geq 1 - \frac{r_1}{1-r_1} = \frac{1-2r_1}{1-r_1}.$$

Unfortunately we cannot show that the Ricci curvature κ is positive in the case where $r_1 \in [1/2,1)$.

If (E,d) is unbounded, the results of [12], Theorems 4 and 5, do not apply here, because their granularity constant

$$\sigma_\infty = \frac{1}{2} \sup_{x\in E^N} \mathrm{Diam}\big(\mathrm{supp}\big(P(x,\cdot)\big)\big)$$

explodes.

Assume now that (E,d) is bounded. If we apply the results ([12], Theorems 4 and 5) and their notations, their coarse diffusion constant

$$\sigma(x)^2 = \frac{1}{2}\iint d_{L^1}(y,z)^2 P(x,\mathrm{d}y)P(x,\mathrm{d}z)$$

is of order N^2; and their local dimension

$$n_x = \inf_{f:\|f\|_{\mathrm{Lip}(d_{L^1})}=1} \frac{\sigma(x)^2}{\mathrm{Var}_{P(x,\cdot)}(f)}$$

is of order N (by Lemma 3.4 below), and their granularity constant σ_∞ is of order N. Setting

$$V^2 = \frac{1}{\kappa n} \sup_{x\in E^N} \frac{\sigma(x)^2}{n_x \kappa}; \qquad r_{\max} = \frac{4V^2\kappa n}{3\sigma_\infty}.$$

Theorem 4 in [12] says that if $\|f\|_{\mathrm{Lip}(d_{L^1})} = 1$,

$$\mathbb{P}_x\left(\left|\frac{1}{n}\sum_{k=1}^{n} f(Z_k) - \frac{1}{n}\sum_{k=1}^{n} P^k f(x)\right| \geq t\right) \leq \begin{cases} 2\exp\left(-\dfrac{t^2}{16V^2}\right), & t \in (0, r_{\max}), \\ 2\exp\left(-\dfrac{\kappa n t}{12\sigma_\infty}\right), & t \geq r_{\max}, \end{cases}$$

for all $n \geq 1$ and $t > 0$. So for small deviation t, their result yields the same order Gaussian concentration inequality, but for large deviation t, their estimate is only exponential, not Gaussian as one may expect in this bounded case. In [12], Theorem 5, they get a same type Gaussian-exponential concentration inequality with $V^2, r_{\max}$ depending upon the starting point x.

Anyway the key lemmas in this paper are necessary for applying the results of [12] to this particular model.

Remark 2.10. For the Gibbs measure μ on $\mathbb{R}^N$, Marton [20] established the Talagrand transport inequality T_2 on $\mathbb{R}^N$ equipped with the Euclidean metric, under the Dobrushin–Shlosman analyticity type condition. The second named author [33] proved $T_1(C)$ for μ on E^N equipped with the metric d_{L^1}, under (H1). But those transport inequalities are for the equilibrium distribution μ, not for the Gibbs sampling which is a Markov chain with μ as invariant measure. However our coupling is very close to that of K. Marton.

Remark 2.11. For ϕ-mixing sequence of dependent random variables, Rio [23] and Samson [28] established accurate concentration inequalities, see also Djellout, Guillin and Wu [4] and the recent works by Paulin [22] and Wintenberger [32] for generalizations and improvements. In the Markov chain case ϕ-mixing means the Doeblin uniform ergodicity. If one applies the results in [23,28] to the Gibbs sampling, one obtains the concentration inequalities with the speed $n/(NS^2)$, where

$$S = \sum_{k=0}^{\infty} \sup_{x \in E^N} \left\|\nu P^k - \mu\right\|_{\mathrm{TV}}.$$

When (H1) holds with the discrete metric d, S is actually finite but it is of order N by Theorem 2.3 (and its remarks). The concentration inequalities so obtained from [23,28] are of speed n/N^3, very far from the correct speed n/N.

Remark 2.12. When f depends on a very small number of variables, since $\|f\|_{\mathrm{Lip}(d_{L^1})} = \max_i \delta_i(f)$ does not reflect the nature of such observable, one can imagine that our concentration inequalities do not yield the correct speed. In fact in the free case and for $f(x) = g(x^1)$, the correct speed must be n, not n/N. For this type of observable, one may use the metric d_{L^2} which reflects much better the number of variables in such observable. The ideas in Marton [19, 20] should be helpful. That will be another history.

3. Proofs of the main results

3.1. The construction of the coupling

Given any two initial distributions ν_1 and ν_2 on E^N, we begin by constructing our coupled non-homogeneous Markov chain $(X_i, Y_i)_{i\geq 0}$, which is quite close to the coupling by Marton [20].

Let (X_0, Y_0) be a coupling of (ν_1, ν_2). And given

$$(X_{kN+i-1}, Y_{kN+i-1}) = (x, y) \in E^N \times E^N, \qquad k \in \mathbb{N}, 1 \leq i \leq N,$$

then

$$X^j_{kN+i} = x^j, \qquad Y^j_{kN+i} = y^j, \qquad j \neq i,$$

and

$$\mathbb{P}\big(\big(X^i_{kN+i}, Y^i_{kN+i}\big) \in \cdot | (X_{kn+i-1}, Y_{kn+i-1}) = (x, y)\big) = \pi(\cdot|x, y),$$

where $\pi(\cdot|x, y)$ is an optimal coupling of $\mu_i(\cdot|x)$ and $\mu_i(\cdot|y)$ such that

$$\int\!\!\int_{E^2} d(\tilde{x}, \tilde{y})\pi(\mathrm{d}\tilde{x}, \mathrm{d}\tilde{y}|x, y) = W_{1,d}\big(\mu_i(\cdot|x), \mu_i(\cdot|y)\big).$$

Define the partial order on $\mathbb{R}^N$ by $a \leq b$ if and only if $a^i \leq b^i, i = 1, \ldots, N$. Then, by (2.3), we have for $\forall k \in \mathbb{N}, 1 \leq i \leq N$,

$$\begin{pmatrix} \mathbb{E}\big[d\big(X^1_{kN+i}, Y^1_{kN+i}\big)|X_{kN+i-1}, Y_{kN+i-1}\big] \\ \vdots \\ \vdots \\ \mathbb{E}\big[d\big(X^N_{kN+i}, Y^N_{kN+i}\big)|X_{kN+i-1}, Y_{kN+i-1}\big] \end{pmatrix} \leq B_i \begin{pmatrix} d\big(X^1_{kN+i-1}, Y^1_{kN+i-1}\big) \\ \vdots \\ \vdots \\ d\big(X^N_{kN+i-1}, Y^N_{kN+i-1}\big) \end{pmatrix},$$

where

$$B_i = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ c_{i1} & c_{i2} & \cdots & \cdots & \cdots & c_{iN} \\ & & & 1 & & & \\ & & & & \ddots & & \\ & & & & & 1 \end{pmatrix} \qquad \text{(the blank in the matrix means 0).}$$

Therefore by iterations, we have

$$\begin{pmatrix} \mathbb{E}d\big(X^1_N, Y^1_N\big) \\ \vdots \\ \vdots \\ \mathbb{E}d\big(X^N_N, Y^N_N\big) \end{pmatrix} \leq B_N B_{N-1} \cdots B_1 \begin{pmatrix} \mathbb{E}d\big(X^1_0, Y^1_0\big) \\ \vdots \\ \vdots \\ \mathbb{E}d\big(X^N_0, Y^N_0\big) \end{pmatrix}. \tag{3.1}$$

Let

$$Q := B_N B_{N-1} \cdots B_1. \tag{3.2}$$

Then we have the following lemma.

Lemma 3.1. *Under* (H1), $\|Q\|_\infty := \max_{1\le i\le N} \sum_{j=1}^{N} Q_{ij} \le r$.

Proof. We use the probabilistic method. Under (H1) we can construct Markov chain $\{\xi_0, \ldots, \xi_N\}$, taking values in $\{1, \ldots, N\} \sqcup \Delta$ where Δ is an extra point representing the cemetery, and write as follows:

$$\xi_0 \overset{B_N}{\to} \xi_1 \overset{B_{N-1}}{\to} \cdots \overset{B_1}{\to} \xi_N,$$

where the transition matrix from ξ_i to ξ_{i+1} is B_{N-i}, more precisely for $\forall k = 1, \ldots, N$,

$$\begin{aligned}
\mathbb{P}\big(\xi_k = j | \xi_{k-1} = N - (k-1)\big) &= c_{N-(k-1),j}, \\
\mathbb{P}\big(\xi_k = \Delta | \xi_{k-1} = N - (k-1)\big) &= 1 - \sum_{j=1}^{N} c_{N-(k-1),j}, \\
\mathbb{P}(\xi_k = j | \xi_{k-1} = i) &= \delta_{ij}, \qquad i \ne N - (k-1), \\
\mathbb{P}(\xi_k = \Delta | \xi_{k-1} = \Delta) &= 1.
\end{aligned}$$

Here $\delta_{ij} = 1$ if $i = j$ and 0 otherwise (Kronecker's symbol). Then

$$\|Q\|_\infty = \max_{1\le i\le N} \mathbb{P}(\xi_N \ne \Delta | \xi_0 = i).$$

For any $i = 1, \ldots, N$, when $\xi_0 = i$, we have $\xi_0 = \xi_1 = \cdots = \xi_{N-i} = i$. Therefore,

$$\mathbb{P}(\xi_{N-i+1} \ne \Delta | \xi_0 = i) = \sum_{j=1}^{N} c_{ij} \le r$$

and thus

$$\mathbb{P}(\xi_N \ne \Delta | \xi_0 = i) \le \mathbb{P}(\xi_{N-1} \ne \Delta | \xi_0 = i) \le \cdots \le \mathbb{P}(\xi_{N-i+1} \ne \Delta | \xi_0 = i) \le r.$$

So $\|Q\|_\infty \le r$. □

3.2. Proof of Theorem 2.3

By (3.1) above, Markov property and iterations,

$$\begin{pmatrix} \mathbb{E}d\big(X_{kN}^1, Y_{kN}^1\big) \\ \vdots \\ \vdots \\ \mathbb{E}d\big(X_{kN}^N, Y_{kN}^N\big) \end{pmatrix} \le Q^k \begin{pmatrix} \mathbb{E}d\big(X_0^1, Y_0^1\big) \\ \vdots \\ \vdots \\ \mathbb{E}d\big(X_0^N, Y_0^N\big) \end{pmatrix}. \tag{3.3}$$

Let $Z_k(1) = X_{kN}$, $Z_k(2) = Y_{kN}$, $k \ge 0$, then by Lemma 3.1

$$\max_{1\le i\le N} \mathbb{E}d\big(Z_k^i(1), Z_k^i(2)\big) \le r^k \max_{1\le i\le N} \mathbb{E}d\big(Z_0^i(1), Z_0^i(2)\big). \tag{3.4}$$

Now the results of this theorem follow quite easily from this inequality. In fact,

(a) For any Lipschitzian function $f : E^N \to \mathbb{R}$,

$$\begin{aligned} \big|\nu_1 P^k f - \nu_2 P^k f\big| &= \big|\mathbb{E}f\big(Z_k(1)\big) - \mathbb{E}f\big(Z_k(2)\big)\big| \\ &\le \sum_{i=1}^N \delta_i(f)\mathbb{E}d\big(Z_k^i(1), Z_k^i(2)\big) \\ &\le r^k \max_{1\le i\le N} \mathbb{E}d\big(Z_0^i(1), Z_0^i(2)\big) \sum_{i=1}^N \delta_i(f), \end{aligned}$$

where the last inequality follows by (3.4). That is (2.6).

(b) Now for $\nu_1 = \nu$, $\nu_2 = \mu$, as $\mu P = \mu$, we have

$$\begin{aligned} W_{1,d_{L^1}}\big(\nu P^k, \mu\big) = W_{1,d_{L^1}}\big(\nu P^k, \mu P^k\big) &\le \mathbb{E}\sum_{i=1}^N d\big(Z_k^i(1), Z_k^i(2)\big) \\ &\le N \max_{1\le i\le N} \mathbb{E}d\big(Z_k^i(1), Z_k^i(2)\big) \\ &\le N r^k \max_{1\le i\le N} \mathbb{E}d\big(Z_0^i(1), Z_0^i(2)\big), \end{aligned}$$

the desired result.

3.3. Proof of Theorem 2.7

We begin with

Lemma 3.2. *If $r_1 := \|C\|_1 < 1$ (i.e., (H2)), then for the matrix Q given in (3.2),*

$$\|Q\|_1 := \max_{1\le j\le N} \sum_{k=1}^N Q_{kj} \le \frac{r_1}{1-r_1}. \tag{3.5}$$

In particular

$$W_{1,d_{L^1}}\big(P(x,\cdot),P(y,\cdot)\big)\le\frac{r_1}{1-r_1}d_{L^1}(x,y)\qquad\forall x,y\in E^N. \tag{3.6}$$

Proof. The last conclusion (3.6) follows from (3.5) and (3.1). We show now (3.5).

By the definition of $Q=B_N\cdots B_1$, it is not difficult to verify for $1\le k\le N$,

$$Q_{kj}=\begin{cases}0, & \text{if } j=1,\\ \displaystyle\sum_{h=1}^{j-1}\Bigg(\sum_{l=1}^{k-1}\sum_{k>i_l>\cdots>i_2>i_1=h}c_{k,i_l}c_{i_l,i_{l-1}}\cdots c_{i_2,i_1=h}c_{h,j}+c_{h,j}1_{h=k}\Bigg), & \text{if } 2\le j\le N.\end{cases} \tag{3.7}$$

Here we make the convention $\sum_{\varnothing}\cdot=0$. This can be obtained again by the Markov chain $(\xi_0,\xi_1,\dots,\xi_N)$ valued in $\{1,\dots,N\}\cup\{\Delta\}$ constructed in Lemma 3.1. Since $\mathbb{P}(\xi_N=1|\xi_{N-1}=i)=0$ for all i, $Q_{k1}=\mathbb{P}(\xi_N=1|\xi_0=k)=0$: that is the first line in the expression of Q. Now for $j=2,\dots,N$, as

$$Q_{kj}=\mathbb{P}(\xi_N=j|\xi_0=k)=\mathbb{P}(\xi_N=j|\xi_{N-k}=k)$$

and if $\xi_{N-k}=k$ and $\xi_{N-k+1}>k$, then $\xi_{N-k+1}=\cdots=\xi_N$ and so

$$\mathbb{P}(\xi_N=j|\xi_{N-k}=k)=c_{kj}1_{k<j}+\sum_{h=1}^{j-1}c_{kh}\mathbb{P}(\xi_N=j|\xi_{N-k+1}=h).$$

This implies the expression of Q above by induction.

Thus for $2\le j\le N$,

$$\begin{aligned}\sum_{k=1}^{N}Q_{kj}&=\sum_{k=1}^{N}\sum_{h=1}^{j-1}\Bigg(\sum_{l=1}^{k-1}\sum_{k>i_l>\cdots>i_2>i_1=h}c_{k,i_l}c_{i_l,i_{l-1}}\cdots c_{i_2,i_1=h}c_{h,j}+c_{h,j}1_{h=k}\Bigg)\\&=\sum_{k=1}^{N}\sum_{h=1}^{j-1}c_{hj}1_{h=k}+\sum_{k=1}^{N}\sum_{h=1}^{j-1}\sum_{l=1}^{k-1}\sum_{k>i_l>\cdots>i_2>i_1=h}c_{k,i_l}c_{i_l,i_{l-1}}\cdots c_{i_2,i_1=h}c_{h,j}\\&=\sum_{h=1}^{j-1}c_{hj}+\sum_{h=1}^{j-1}\sum_{k=1}^{N}\sum_{l=1}^{k-1}\sum_{k>i_l>\cdots>i_2>i_1=h}c_{k,i_l}c_{i_l,i_{l-1}}\cdots c_{i_2,i_1=h}c_{h,j}\\&\le r_1+\sum_{l=1}^{N-1}\sum_{h=1}^{j-1}\sum_{k=l+1}^{N}\sum_{k>i_l>\cdots>i_2>i_1=h}c_{k,i_l}c_{i_l,i_{l-1}}\cdots c_{i_2,i_1=h}c_{h,j}\\&\le r_1+r_1^2+\cdots+r_1^N,\end{aligned}$$

where the last inequality holds because for fixed l: $1 \le l \le N-1$ and h: $1 \le h \le j-1$,

$$\sum_{k=1}^{N} \sum_{k>i_l>\cdots>i_2>i_1=h} c_{k,i_l} c_{i_l,i_{l-1}} \cdots c_{i_2,i_1=h} \le \sum_{k=1}^{N} \bigl(C^l\bigr)_{kh} \le \bigl\|C^l\bigr\|_1 \le r_1^l.$$

So the proof of (3.5) is completed. □

Remark 3.3. Let μ be the Gaussian distribution on $\mathbb{R}^2$ with mean 0 and the covariance matrix $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$ where $r \in (0,1)$. We have $c_{12} = c_{21} = r < 1$ (i.e., (H1) and (H2) both hold); and under $P((x_1,x_2),(\mathrm{d}y_1,\mathrm{d}y_2))$, $z_1 = y_1 - rx_2$ and $z_2 = y_2 - ry_1$ are i.i.d. Gaussian random variables with mean 0 and variance $1-r^2$. Hence,

$$Q = B_2 B_1 = \begin{pmatrix} 1 & 0 \\ r & 0 \end{pmatrix} \begin{pmatrix} 0 & r \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & r \\ 0 & r \end{pmatrix}$$

and since $y_1 = rx_2 + z_1$, $y_2 = z_2 + rz_1 + r^2 x_2$,

$$W_{1,d_{L^1}}\bigl[P\bigl((x_1,x_2),\cdot\bigr), P\bigl(\bigl(x_1',x_2'\bigr),\cdot\bigr)\bigr] = \bigl(r+r^2\bigr)\bigl|x_2 - x_2'\bigr|.$$

Thus, $\|Q\|_1 = 2r$ and the Ricci curvature κ is positive if and only if $r + r^2 < 1$. In other words, though we have missed many terms in the proof above, the estimate of $\|Q\|_1$ cannot be qualitatively improved.

Lemma 3.4. *Assume* (H2) *and* (H3), *then*

$$P(x_0,\cdot) \in T_1\left(\frac{NC_1}{(1-r_1)^2}\right) \qquad \forall x_0 = \bigl(x_0^1,\ldots,x_0^N\bigr) \in E^N.$$

Proof. The proof is similar to the one used by Djellout, Guillin and Wu [4], Theorem 2.5. First for simplicity denote $P(x_0,\cdot)$ by P and note that for $1 \le i \le N$,

$$X_N^1 = X_1^1, \ldots, X_N^i = X_i^i,$$
$$P\bigl(X_N^i \in \cdot | X_N^1, \ldots, X_N^{i-1}\bigr) = \mu_i\bigl(\cdot | X_N^1, \ldots, X_N^{i-1}, x_0^{i+1}, \ldots, x_0^N\bigr)$$

and thus

$$W_{1,d}\bigl(P\bigl(X_N^i \in \cdot | X_N^1 = x^1, \ldots, X_N^{i-1} = x^{i-1}\bigr), P\bigl(X_N^i \in \cdot | X_N^1 = y^1, \ldots, X_N^{i-1} = y^{i-1}\bigr)\bigr)$$
$$\le \sum_{j=1}^{i-1} c_{ij} d\bigl(x^j, y^j\bigr).$$

For any probability measure Q on E^N such that $H(Q|P) < \infty$, let $Q_i(\cdot|x^{[1,i-1]})$ be the regular conditional law of x^i knowing $x^{[1,i-1]}$, where $i \ge 2$, $x^{[1,i-1]} = (x^1, \ldots, x^{i-1})$, and $Q_i(\cdot|x^{[1,i-1]})$

the law of x^1 for $i = 1$, all under law Q. Define $P_i(\cdot|x^{[1,i-1]})$ similarly but under P. We shall use the Kullback information between conditional distributions,

$$H_i\big(y^{[1,i-1]}\big) = H\big(Q_i\big(\cdot|y^{[1,i-1]}\big)|P_i\big(\cdot|y^{[1,i-1]}\big)\big)$$

and exploit the following important identity:

$$H(Q|P) = \sum_{i=1}^{N} \int_{E^N} H_i\big(y^{[1,i-1]}\big)\,\mathrm{d}Q(y).$$

The key is to construct an appropriate coupling of Q and P, that is, two random sequences $Y^{[1,N]}$ and $X^{[1,N]}$ taking values on E^N distributed according to Q and P, respectively, on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We define a joint distribution $\mathscr{L}(Y^{[1,N]}, X^{[1,N]})$ by induction as follows (the Marton coupling).

At first the law of (Y^1, X^1) is the optimal coupling of $Q(x^1 \in \cdot)$ and $P(x^1 \in \cdot)$ $(= \mu_1(\cdot|x_0))$. Assume that for some $i, 2 \le i \le N$, $(Y^{[1,i-1]}, X^{[1,i-1]}) = (y^{[1,i-1]}, x^{[1,i-1]})$ is given. Then the joint conditional distribution $\mathscr{L}(Y^i, X^i|Y^{[1,i-1]} = y^{[1,i-1]}, X^{[1,i-1]} = x^{[1,i-1]})$ is the optimal coupling of $Q_i(\cdot|y^{[1,i-1]})$ and $P_i(\cdot|x^{[1,i-1]})$, that is,

$$\mathbb{E}\big(d\big(Y^i, X^i\big)|Y^{[1,i-1]} = y^{[1,i-1]}, X^{[1,i-1]} = x^{[1,i-1]}\big) = W_{1,d}\big(Q_i\big(\cdot|y^{[1,i-1]}\big), P_i\big(\cdot|x^{[1,i-1]}\big)\big).$$

Obviously, $Y^{[1,N]}$, $X^{[1,N]}$ are of law Q, P, respectively. By the triangle inequality for the $W_{1,d}$ distance,

$$\begin{aligned}
&\mathbb{E}\big(d\big(Y^i, X^i\big)|Y^{[1,i-1]} = y^{[1,i-1]}, X^{[1,i-1]} = x^{[1,i-1]}\big)\\
&\quad\le W_{1,d}\big(Q_i\big(\cdot|y^{[1,i-1]}\big), P_i\big(\cdot|y^{[1,i-1]}\big)\big) + W_{1,d}\big(P_i\big(\cdot|y^{[1,i-1]}\big), P_i\big(\cdot|x^{[1,i-1]}\big)\big)\\
&\quad\le \sqrt{2C_1 H_i\big(y^{[1,i-1]}\big)} + \sum_{j=1}^{i-1} c_{ij} d\big(x^j, y^j\big).
\end{aligned}$$

By recurrence on i, this entails that $\mathbb{E}d(Y^i, X^i) < \infty$ for all $i = 1, \dots, N$. Taking the average with respect to $\mathscr{L}(Y^{[1,i-1]}, X^{[1,i-1]})$, summing on i and using Jessen's inequality, we have

$$\begin{aligned}
\frac{\sum_{i=1}^N \mathbb{E}d(Y^i, X^i)}{N} &\le \sqrt{\frac{2C_1 \sum_{i=1}^N \mathbb{E}H_i(Y^{[1,i-1]})}{N}} + \frac{\sum_{i=1}^N \sum_{j=1}^{i-1} c_{ij}\mathbb{E}d(Y^j, X^j)}{N}\\
&= \sqrt{\frac{2C_1 H(Q|P)}{N}} + \frac{\sum_{j=1}^{N-1} \mathbb{E}d(Y^j, X^j) \sum_{i=j+1}^N c_{ij}}{N}\\
&\le \sqrt{\frac{2C_1 H(Q|P)}{N}} + \frac{r_1 \sum_{j=1}^N \mathbb{E}d(Y^j, X^j)}{N}
\end{aligned}$$

the above inequality gives us

$$W_{1,d_{L^1}}(Q,P) \le \sum_{i=1}^{N} \mathbb{E} d\big(Y^i, X^i\big) \le \sqrt{2\frac{NC_1}{(1-r_1)^2} H(Q|P)},$$

that is, $P = P(x_0,\cdot) \in T_1(\frac{NC_1}{(1-r_1)^2})$. □

Theorem 2.7 is based on the following dependent tensorization result of Djellout, Guillin and Wu [4].

Lemma 3.5 ([4], Theorem 2.11). *Let $\mathbb{P}$ be a probability measure on the product space $(E^n, \mathscr{B}^n), n \ge 2$. For any $x = (x_1, \dots, x_n) \in E^n, x_{[1,k]} := (x_1, \dots, x_k)$. Let $\mathbb{P}_k(\cdot|x_{[1,k-1]})$ denote the regular conditional law of x_k given $x_{[1,k-1]}$ under $\mathbb{P}$ for $2 \le k \le n$, and $\mathbb{P}_k(\cdot|x_{[1,k-1]})$ be the distribution of x_1 for $k = 1$.*

Assume that:

(1) *For some metric d on E, $\mathbb{P}_k(\cdot|x_{[1,k-1]}) \in T_1(C)$ on (E,d) for all $k \ge 1, x_{[1,k-1]} \in E^{k-1}$;*
(2) *there is some constant $S > 0$ such that for all real bounded Lipschitzian function $f(x_{k+1}, \dots, x_n)$ with $\|f\|_{\mathrm{Lip}(d_{L^1})} \le 1$, for all $x \in E^n, y_k \in E$,*

$$\big|\mathbb{E}_{\mathbb{P}}\big(f(X_{k+1}, \dots, X_n)|X_{[1,k]} = x_{[1,k]}\big) - \mathbb{E}_{\mathbb{P}}\big(f(X_{k+1}, \dots, X_n)|X_{[1,k]} = (x_{[1,k-1]}, y_k)\big)\big| \le S d(x_k, y_k).$$

Then for all function F on E^n satisfying $\|F\|_{\mathrm{Lip}(d_{L^1})} \le \alpha$, we have

$$\mathbb{E}_{\mathbb{P}} e^{\lambda(F - \mathbb{E}_{\mathbb{P}}F)} \le \exp\big(C\lambda^2(1+S)^2\alpha^2 n/2\big) \qquad \forall \lambda \in \mathbb{R}.$$

Equivalently, $\mathbb{P} \in T_1(C_n)$ on (E^n, d_{L^1}) with

$$C_n = nC(1+S)^2.$$

We are now ready to prove Theorem 2.7.

Proof of Theorem 2.7. We will apply Lemma 3.5 with (E,d) being (E^N, d_{L^1}), and $\mathbb{P}$ be the law of $(Z_1, \dots, Z_n)$ on $(E^N)^n$.

By (3.3), Lemma 3.2 and the condition that $r_1 < 1/2$, the constant S in Lemma 3.5 is bounded from above by

$$\sup_{x,y \in E^N} \frac{1}{d_{L^1}(x,y)} \mathbb{E}^{x,y} \sum_{k=1}^{\infty} d_{L^1}(X_{kN}, Y_{kN}) \le \sum_{k=1}^{\infty} \left(\frac{r_1}{1-r_1}\right)^k = \frac{r_1}{1-2r_1}.$$

Take $S = \frac{r_1}{1-2r_1}$, $F(Z_1, \dots, Z_n) = \frac{1}{n}\sum_{k=1}^n f(Z_k)$, then the Lipschitzian norm $\|F\|_{\mathrm{Lip}}$ of F w.r.t. the $d_{L^1}(x,y) = \sum_{k=1}^n d_{L^1}(x_k, y_k)$ (for $x, y \in (E^N)^n$) is not greater than $\|f\|_{\mathrm{Lip}(d_{L^1})}/n \le$

α/n. Thus by Lemmas 3.5 and 3.4,

$$\mathbb{E}^x \exp\left(\lambda\left(\frac{1}{n}\sum_{k=1}^{n} f(Z_k) - \frac{1}{n}\sum_{k=1}^{n} P^k f(x)\right)\right)$$
$$\leq \exp\left\{\frac{NC_1}{(1-r_1)^2}\lambda^2(1+S)^2\left(\frac{\alpha}{n}\right)^2 n/2\right\} \qquad \forall \lambda \in \mathbb{R}.$$

So, by the classic approach, firstly using Chebyshev's inequality, and then optimizing over $\lambda \geq 0$, we obtain the desired part (a) in Theorem 2.7.

Furthermore by Theorem 2.3, we have

$$\left|\frac{1}{n}\sum_{k=1}^{n} P^k f(x) - \mu(f)\right| \leq \frac{1}{n}\sum_{k=1}^{n}\left|P^k f(x) - \mu(f)\right|$$
$$\leq \frac{1}{n}\sum_{k=1}^{n} r^k \max_{1\leq i\leq N}\int_{E^N} d\left(x^i, y^i\right) \mathrm{d}\mu(y) \cdot \sum_{i=1}^{N}\delta_i(f)$$
$$\leq \frac{1}{n}\frac{r}{1-r}\max_{1\leq i\leq N}\int_{E^N} d\left(x^i, y^i\right) \mathrm{d}\mu(y) \cdot \sum_{i=1}^{N}\delta_i(f).$$

Thus, we obtain part (b) in Theorem 2.7 from its part (a). □

Acknowledgements

Supported in part by Thousand Talents Program of the Chinese Academy of Sciences and le projet ANR EVOL. We are grateful to the two referees for their suggestions and references, which improve sensitively the presentation of the paper.

References

[1] Bobkov, S.G. and Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.* **163** 1–28. MR1682772

[2] Diaconis, P., Khare, K. and Saloff-Coste, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials. *Statist. Sci.* **23** 151–178. MR2446500

[3] Diaconis, P., Khare, K. and Saloff-Coste, L. (2010). Gibbs sampling, conjugate priors and coupling. *Sankhya A* **72** 136–169. MR2658168

[4] Djellout, H., Guillin, A. and Wu, L. (2004). Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab.* **32** 2702–2732. MR2078555

[5] Dobrushin, R.L. (1968). The description of a random field by means of conditional probabilities and condition of its regularity. *Theory Probab. Appl.* **13** 197–224.

[6] Dobrushin, R.L. (1970). Prescribing a system of random variables by conditional distributions. *Theory Probab. Appl.* **15** 458–486.

[7] Dobrushin, R.L. and Shlosman, S.B. (1985). Completely analytical Gibbs fields. In *Statistical Physics and Dynamical Systems* (*Köszeg*, 1984). *Progress in Probability* **10** 371–403. Boston, MA: Birkhäuser. MR0821307
[8] Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. *Statistics for Engineering and Information Science*. New York: Springer. MR1847783
[9] Gozlan, N. and Léonard, C. (2007). A large deviation approach to some transportation cost inequalities. *Probab. Theory Related Fields* **139** 235–283. MR2322697
[10] Gozlan, N. and Léonard, C. (2010). Transport inequalities. A survey. *Markov Process. Related Fields* **16** 635–736. MR2895086
[11] Hairer, M. and Mattingly, J.C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis*, *Random Fields and Applications VI*. *Progress in Probability* **63** 109–117. Basel: Birkhäuser. MR2857021
[12] Joulin, A. and Ollivier, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38** 2418–2442. MR2683634
[13] Ledoux, M. (1999). Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités*, *XXXIII*. *Lecture Notes in Math*. **1709** 120–216. Berlin: Springer. MR1767995
[14] Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Providence, RI: Amer. Math. Soc. MR1849347
[15] Martinelli, F. (1999). Lectures on Glauber dynamics for discrete spin models. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1997). *Lecture Notes in Math*. **1717** 93–191. Berlin: Springer. MR1746301
[16] Martinelli, F. and Olivieri, E. (1994). Approach to equilibrium of Glauber dynamics in the one phase region. I. The attractive case. *Comm. Math. Phys*. **161** 447–486. MR1269387
[17] Marton, K. (1996). Bounding $\overline{d}$-distance by informational divergence: A method to prove measure concentration. *Ann. Probab.* **24** 857–866. MR1404531
[18] Marton, K. (1996). A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal*. **6** 556–571. MR1392329
[19] Marton, K. (2003). Measure concentration and strong mixing. *Studia Sci. Math. Hungar.* **40** 95–113. MR2002993
[20] Marton, K. (2004). Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.* **32** 2526–2544. MR2078549
[21] Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. *Communications and Control Engineering Series*. London: Springer. MR1287609
[22] Paulin, D. (2012). Concentration inequalities for Markov chains by Marton coupling. Preprint.
[23] Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math*. **330** 905–908. MR1771956
[24] Roberts, G.O. and Rosenthal, J.S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71. MR2095565
[25] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc*. **90** 558–566. MR1340509
[26] Rosenthal, J.S. (1996). Analysis of the Gibbs sampler for a model related to James–Stein estimations. *Statist. Comput*. **6** 269–275.
[27] Rosenthal, J.S. (2002). Quantitative convergence rates of Markov chains: A simple account. *Electron. Commun. Probab.* **7** 123–128 (electronic). MR1917546
[28] Samson, P.M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28** 416–461. MR1756011
[29] Talagrand, M. (1996). Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal*. **6** 587–600. MR1392331

[30] Villani, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Providence, RI: Amer. Math. Soc. MR1964483

[31] Winkler, G. (1995). *Image Analysis*, *Random Fields and Dynamic Monte Carlo Methods*: *A Mathematical Introduction*. *Applications of Mathematics* (*New York*) **27**. Berlin: Springer. MR1316400

[32] Wintenberger, O. (2012). Weak transport inequalities and applications to exponential inequalities and oracle inequalities. Preprint.

[33] Wu, L. (2006). Poincaré and transportation inequalities for Gibbs measures under the Dobrushin uniqueness condition. *Ann*. *Probab*. **34** 1960–1989. MR2271488

[34] Zegarliński, B. (1992). Dobrushin uniqueness theorem and logarithmic Sobolev inequalities. *J*. *Funct*. *Anal*. **105** 77–111. MR1156671

Received February 2013 and revised April 2013