Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2014, Vol. 50, No. 2, 539–563
DOI: 10.1214/12-AIHP517

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

On the mean speed of convergence of empirical and occupation measures in Wasserstein distance

Emmanuel Boissard and Thibaut Le Gouic

Institut de Mathématiques de Toulouse (CNRS UMR 5219). Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France.
E-mail: emmanuel.boissard@math.univ-toulouse.fr; thibaut.le-gouic@math.univ-toulouse.fr

Received 22 July 2011; revised 16 February 2012; accepted 15 July 2012

Abstract. In this work, we provide non-asymptotic bounds for the average speed of convergence of the empirical measure in the law of large numbers, in Wasserstein distance. We also consider occupation measures of ergodic Markov chains. One motivation is the approximation of a probability measure by finitely supported measures (the quantization problem). It is found that rates for empirical or occupation measures match or are close to previously known optimal quantization rates in several cases. This is notably highlighted in the example of infinite-dimensional Gaussian measures.

Résumé. Dans ce travail, on exhibe des bornes non asymptotiques pour la vitesse de convergence en moyenne de la mesure empirique dans la loi des grands nombres, en distance de Wasserstein. On considère également la mesure d'occupation d'une chaîne de Markov ergodique. L'une des motivations est l'approximation d'une mesure de probabilité par des mesures à support fini (le problème de la quantification). On détermine que les taux de convergence des mesures empiriques ou des mesures d'occupation correspondent dans plusieurs cas aux taux de quantification optimale déjà établis par ailleurs. Ce fait est notamment établi pour des mesures gaussiennes dans des espaces de dimension infinie.

MSC: 60B10; 65C50; 60J05

Keywords: Wasserstein metrics; Optimal transportation; Functional quantization; Transportation inequalities; Markov chains; Measure theory

1. Introduction

This paper is concerned with the rate of convergence in Wasserstein distance for the so-called *empirical law of large numbers*: let (E, d, μ) denote a measured Polish space, and let

$$L_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{1}$$

denote the empirical measure associated with the i.i.d. sample $(X_i)_{1\le i\le n}$ of law μ, then with probability 1, $L_n \rightharpoonup \mu$ as $n \to +\infty$ (convergence is understood in the sense of the weak topology of measures). This theorem is also known as Glivenko–Cantelli theorem and is due in this form to Varadarajan [31].

For $1 \le p < +\infty$, the p-*Wasserstein distance* is defined on the set $\mathcal{P}_p(E)^2$ of couples of measures with a finite pth moment by

$$W_p^p(\mu, \nu) = \inf_{\pi\in\mathcal{P}(\mu,\nu)} \int d^p(x, y)\pi(\mathrm{d}x, \mathrm{d}y),$$

where the infimum is taken over the set $\mathcal{P}(\mu, \nu)$ of probability measures with first, resp. second, marginal μ, resp. ν. This defines a metric on $\mathcal{P}_p$, and convergence in this metric is equivalent to weak convergence plus convergence of the moment of order p. These metrics, and more generally the Monge transportation problem from which they originate, have played a prominent role in several areas of probability, statistics and the analysis of P.D.E.s: for a rich account, see C. Villani's St-Flour course [32].

Our purpose in this paper is to give bounds on the speed of convergence in W_p distance for the Glivenko–Cantelli theorem, i.e. bounds for the convergence $\mathbb{E}(W_p(L_n, \mu)) \to 0$. Such results are desirable notably in view of numerical and statistical applications: indeed, the approximation of a given probability measure by a measure with finite support in Wasserstein distance is a topic that appears in various guises in the literature, see for example [16]. The first motivation for this work was to extend the results obtained by F. Bolley, A. Guillin and C. Villani [5] in the case of variables with support in $\mathbb{R}^d$. As in this paper, we aim to produce bounds that are non-asymptotic and effective (that is with explicit constants), in order to achieve practical relevance.

We also extend the investigation to the convergence of occupation measure for suitably ergodic Markov chains: again, we have practical applications in mind, as this allows to use Metropolis–Hastings-type algorithms to approximate an unknown measure (see Section 1.3 for a discussion of this).

There are many works in statistics devoted to convergence rates in some metric associated with the weak convergence of measures, see e.g. the book of A. Van der Vaart and J. Wellner [30]. Of particular interest to us is R. M. Dudley's article [12], see Remark 1.1.

Other works have been devoted to convergence of empirical measures in Wasserstein distance, we quote some of them. Horowitz and Karandikar [19] gave a bound for the rate of convergence of $\mathbb{E}[W_2^2(L_n, \mu)]$ to 0 for general measures supported in $\mathbb{R}^d$ under a moment condition. M. Ajtai, J. Komlos and G. Tusnády [1] and M. Talagrand [29] studied the related problem of the average cost of matching two i.i.d. samples from the uniform law on the unit cube in dimension $d \geq 2$. This line of research was pushed further, among others, by V. Dobrić and J. E. Yukich [11] or F. Barthe and C. Bordenave [2] (the reader may refer to this last paper for an up-to-date account of the Euclidean matching problem). These papers give a sharp result for measures in $\mathbb{R}^d$, with an improvement both over [19] and [5]. In the case $\mu \in \mathcal{P}(\mathbb{R})$, E. del Barrio, E. Giné and C. Matrán [8] obtain a central limit theorem for $W_1(L_n, \mu)$ under the condition that $\int_{-\infty}^{+\infty} \sqrt{F(t)(1-F(t))}\,\mathrm{d}t < +\infty$ where F is the cumulative distribution function (c.d.f.) of μ. In the companion paper [4], we investigate the case of the W_1 distance by using the dual expression of the W_1 transportation cost by Kantorovich and Rubinstein, see therein for more references.

We will also discuss problems related to the optimal quantization of probability measures, that is the approximation of probability distributions by distributions with finite support. A general reference on this topic is the book by S. Graf and H. Luschgy [16]. Let us also mention the paper [17] that lays out the connections between optimal quantization and empirical processes.

Before moving on to our results, we make a remark on the scope of this work. Generally speaking, the problem of convergence of $W_p(L_n, \mu)$ to 0 can be divided in two separate questions:

- the first one is to estimate the *mean rate of convergence*, that is the convergence rate of $\mathbb{E}[W_p(L_n, \mu)]$,
- while the second one is to study the concentration properties of $W_p(L_n, \mu)$ around its mean, that is to find bounds on the quantities

$$\mathbb{P}\big(W_p(L_n, \mu) - \mathbb{E}\big[W_p(L_n, \mu)\big] \geq t\big).$$

Our main concern here is the first point. The second one can be dealt with by techniques of measure concentration. We will elaborate on this in Appendices A and B.

1.1. *Main result and first consequences*

Definition 1.1. *For $S \subset E$, the covering number of order δ for S, denoted by $N(S, \delta)$, is defined as the minimal $n \in \mathbb{N}$ such that there exist $x_1, \ldots, x_n$ in S with*

$$S \subset \bigcup_{j=1}^{n} B(x_i, \delta).$$

Our main statement is summed up in the following result.

Theorem 1.1. *Choose $t > 0$. Let $\mu \in \mathcal{P}(E)$ with support included in $S \subset E$ with finite diameter Δ_S such that $N(S,t) < +\infty$. We have the bound*:

$$\mathbb{E}\big(W_p(L_n,\mu)\big) \le c\left(t + n^{-1/2p}\int_t^{\Delta_S/4} N(S,\delta)^{1/2p}\,\mathrm{d}\delta\right)$$

with $c \le 64/3$.

Remark. *Theorem 1.1 is related in spirit and proof to the results of R. M. Dudley* [12] *in the case of the bounded Lipschitz metric*

$$d_{BL}(\mu,\nu) = \inf_{f\,1-\mathrm{Lip},\|f\|_\infty\le 1}\int f\,\mathrm{d}(\mu-\nu).$$

The analogy is not at all fortuitous: *indeed*, *the bounded Lipschitz metric is linked to the* 1-*Wasserstein distance via the well-known Kantorovich–Rubinstein dual definition of W_1*:

$$W_1(\mu,\nu) = \sup_{f\,1-\mathrm{Lip}}\int f\,\mathrm{d}(\mu-\nu).$$

The analogy stops at $p = 1$ since there is no representation of W_p as an empirical process for $p > 1$ (there is, *however*, *a general dual expression of the transport cost). In spite of this*, *the technique of proof in* [12] *proves useful in our case*, *and the technique of using a sequence of coarser and coarser partitions is at the heart of many later results*, *notably in the literature concerned with the problem of matching two independent samples in Euclidean space*, *see e.g.* [29] *or the recent paper* [2].

We now give a first example of application, under an assumption that the underlying metric space is of finite-dimensional type in some sense. More precisely, we assume that there exist $k_E > 0$, $\alpha > 0$ such that

$$N(E,\delta) \le k_E(\mathrm{Diam}\,E/\delta)^\alpha. \tag{2}$$

Here, the parameter α plays the role of a dimension.

Corollary 1.2. *Assume that E satisfies* (2), *and that $\alpha > 2p$. With notations as earlier, the following holds*:

$$\mathbb{E}\big[W_p(L_n,\mu)\big] \le c\left(\frac{2p}{\alpha-2p}\right)^{2p/\alpha}\mathrm{Diam}\,E k_E^{1/\alpha} n^{-1/\alpha}$$

with $c \le 64/3$.

Remark. *In the case of measures supported in $\mathbb{R}^d$*, *this result is neither new nor fully optimal. For a sharp statement in this case*, *the reader may refer to* [2] *and references therein. However, we recover at least the exponent of $n^{-1/d}$ which is sharp for $d \ge 3$*, *see* [2] *for a discussion. And on the other hand*, *Corollary 1.2 extends to more general metric spaces of finite-dimensional type*, *for example manifolds.*

Remark. *Corollary 1.2 and other results throughout this work require a lower bound on the dimension parameter α. Here for instance we impose $\alpha > 2p$, which implies that $\alpha > 2$ since $p \ge 1$. This high-dimensional hypothesis is also commonplace in matching problems*: *for instance in* [1] (*where matchings over the cube with $p = 1$ are studied*), *the convergence rates for dimension* 3 *and above differ from those in dimension* 1 *and* 2. *In low dimensions*, *further difficulties arise and Theorem 1.1 will overestimate the convergence rate. For instance on the* 2-*dimensional cube we would get a convergence rate of $\log n/\sqrt{n}$ instead of a correct $\sqrt{\log n/n}$ as established in* [1].

It is possible to remove the assumption of boundedness of the metric space and replace it with the following: we assume that there exist $k_E > 0$, $\alpha > 0$ such that for all bounded $S \subset E$,

$$N(S, \delta) \le k_E (\operatorname{Diam} S/\delta)^\alpha. \tag{3}$$

In this context, the following result holds.

Corollary 1.3. *Assume that (E, d) satisfies (3), that that $\mu \in \mathcal{P}_p(E)$ has some finite moment of order $q > 2p \vee \alpha p/(\alpha - p)$, meaning that $M_q = \int d^q(x_0, x)\,\mathrm{d}\mu < +\infty$ for some $x_0 \in E$. Also assume that $\alpha > 2p$.*

Then there exists $C > 0$, depending on E, p, α and M_q, such that

$$\mathbb{E}\big(W_p(L_n, \mu)\big) \le C n^{-1/\alpha}.$$

Remark. *A more precise statement is the following: with the same notations as above, for all $\xi > 1$ and $\zeta > 1$, we have*

$$\mathbb{E}\big(W_p(L_n, \mu)\big) \le C(\xi) n^{-1/2p} + C'(\zeta) n^{-1/\alpha},$$

where there exist constants c, c' depending on p, α, E, but not on μ (c is universal, c' is the constant that appears in Corollary 1.2), such that

$$C(\xi) = c\left[1 + M_p^{1/p} + M_{2\xi p}^{1/2p}\left(M_p^{1/p}\frac{2^{-\xi}}{1 - 2^{-\xi}} + \frac{2^{1-\xi}}{1 - 2^{1-\xi}}\right)\right],$$

$$C'(\zeta) = c'\left[1 + M_{\zeta\alpha p/(\alpha - p)}^{(\alpha - p)/\alpha p}\frac{4^{1-\zeta}}{1 - 2^{1-\zeta}}\right].$$

As opposed to Corollaries 1.2 and 1.3, our next result is set in an infinite-dimensional framework.

1.2. *An application to Gaussian r.v.s in Banach spaces*

We apply the results above to the case where E is a separable Banach space with norm $\|\cdot\|$, and μ is a centered Gaussian random variable with values in E, meaning that the image of μ by every continuous linear functional $f \in E^*$ is a centered Gaussian variable in $\mathbb{R}$. The couple (E, μ) is called a (separable) Gaussian–Banach space.

Let X be an E-valued r.v. with law μ, and define the weak variance of μ as

$$\sigma = \sup_{f \in E^*, |f| \le 1} \big(\mathbb{E} f^2(X)\big)^{1/2}.$$

The small ball function of a Gaussian–Banach space (E, μ) is the function

$$\psi(t) = -\log \mu\big(B(0, t)\big).$$

We can associate to the couple (E, μ) their Cameron–Martin Hilbert space $H \subset E$, see e.g. [22] for a reference. It is known that the small ball function has deep links with the covering numbers of the unit ball of H, see e.g. the papers by Kuelbs and Li [21] and Li and Linde [25], as well as with the approximation of μ by measures with finite support in Wasserstein distance (the quantization or optimal quantization problem), see Fehringer's Ph.D. thesis [13], Dereich, Fehringer, Matoussi and Scheutzow [9], Graf, Luschgy and Pagès [18].

We make the following assumptions on the small ball function:

(1) there exists $\kappa > 1$ such that $\psi(t) \le \kappa\psi(2t)$ for $0 < t \le t_0$,
(2) for all $\varepsilon > 0$, $n^{-\varepsilon} = \mathrm{o}(\psi^{-1}(\log n))$.

Assumption (2) implies that the Gaussian measure is genuinely infinite dimensional: indeed, in the case when $\dim K < +\infty$, the measure is supported in a finite-dimensional Banach space, and in this case the small ball function behaves as $\log t$.

Theorem 1.4. *Let (E,μ) be a Gaussian–Banach space with weak variance σ and small ball function ψ. Assume that assumptions (1) and (2) hold.*

Then there exists a universal constant c such that for all integers $n \geq 1$ such that

$$\log n \geq (6+\kappa)\big(\log 2 \vee \psi(1) \vee \psi(2t_0) \vee 1/\sigma^2\big),$$

the following holds:

$$\mathbb{E}\big(W_2(L_n,\mu)\big) \leq c\left[\psi^{-1}\left(\frac{1}{6+\kappa}\log n\right) + \sigma n^{-1/[4(6+\kappa)]}\right]. \tag{4}$$

In particular, there is a $C = C(\mu)$ such that

$$\mathbb{E}\big(W_2(L_n,\mu)\big) \leq C\psi^{-1}(\log n). \tag{5}$$

Moreover, for $\lambda > 0$,

$$W_2(L_n,\mu) \leq (C+\lambda)\psi^{-1}(\log n) \quad \textit{with probability } 1 - \exp\left[-n\big(\psi^{-1}(\log n)\big)^2 \frac{\lambda^2}{2\sigma^2}\right]. \tag{6}$$

Remark. *Note that the choice of $6+\kappa$ is not particularly sharp and may likely be improved.*

In order to underline the interest of the result above, we introduce some definitions from optimal quantization. For $n \geq 1$ and $1 \leq r < +\infty$, define the optimal quantization error at rate n as

$$\delta_{n,r}(\mu) = \inf_{\nu \in \Theta_n} W_r(\mu,\nu),$$

where the infimum runs over the set Θ_n of probability measures with finite support of cardinal bounded by n.

Precise connections have been made in the literature between the rate of optimal quantization (i.e. the speed of convergence of $\delta_{n,r}$ to 0) and the behaviour of the small ball function near 0. Two rather complete works on this topic are [9] and [18]. Assume that $t \mapsto \psi(1/t)$ is *regularly varying at infinity*, i.e. that there exists a function L and $a > 0$ such that

$$\psi(\varepsilon) = \varepsilon^{-a} L\left(\frac{1}{\varepsilon}\right),$$

$$\frac{L(st)}{L(t)} \to_{t\to+\infty} 1 \quad \text{for all } s > 0.$$

Roughly speaking, Theorem 4.1 in [9] and Theorem 2 in [18] imply that there exist $c, c' > 0$ such that

$$c\psi^{-1}(\log n) \lesssim \delta_{n,r} \lesssim c'\psi^{-1}(\log n)$$

(where $a_n \lesssim b_n$ means $\liminf b_n/a_n \geq 1$).

Please note that the regular variation condition is not the sharpest condition stated in either paper; however it is satisfied by usual Gaussian processes. In the terminology of quantization, the quantization rate is given by the sequence $\psi^{-1}(\log n)$. We can restate Theorem 1.4 by saying that the empirical measure is a rate-optimal quantizer with high probability (under some assumptions on the small ball function and when the distortion index is $r = 2$). This is of practical interest, since obtaining the empirical measure is only as difficult as simulating an instance of the Gaussian vector, and one avoids dealing with computation of appropriate weights in the approximating discrete measure.

We now quote some results on the asymptotic behaviour of the quantization rate and the small ball function for classic Gaussian processes, to illustrate the result above. In all of these examples, the assumptions of Theorem 1.4 on the small ball function are satisfied, so that the convergence rate is also the proper one for empirical measures.

- $E=(L^2([0,1]),\|\cdot\|_2)$ and μ is the Wiener measure. In this case, we quote [9] to get

  $$\psi(t)\sim_{t\to 0}\frac{1}{8t^2}.$$

  Thus,

  $$\frac{1}{\sqrt{8\log n}}\lesssim\delta_{n,r}\lesssim\frac{1}{\sqrt{\log n}}.$$

  Actually, a sharper result is $\delta_{n,r}\sim\sqrt{2}/\pi\sqrt{\log n}$, c.f. [26]. In our case, we get the bound

  $$\mathbb{E}W_2(L_n,\mu)=\mathrm{O}\left(\frac{1}{\sqrt{\log n}}\right).$$

- $E=(C([0,1]),\|\cdot\|_\infty)$ (the space of continuous functions endowed with the sup-norm) and μ is the Wiener measure. Quoting again from [9], we have

  $$\psi(t)\sim\frac{\pi^2}{8t^2}$$

  from which we deduce again that there exist $c,c'>0$ with

  $$\frac{c}{\sqrt{\log n}}\lesssim\delta_{n,r}\lesssim\frac{c'}{\sqrt{\log n}}$$

  and $\mathbb{E}W_2(L_n,\mu)=\mathrm{O}(1/\sqrt{\log n})$.
- $E=(C([0,1]),\|\cdot\|_\infty)$, and this time μ is the law of a fractional Brownian motion with Hurst exponent $\rho\in(0,1)$. As stated in [18,25,26], we have

  $$\psi(t)\sim\frac{C_\rho}{t^{1/\rho}}$$

  which entails a bound $\mathbb{E}W_2(L_n,\mu)=\mathrm{O}((\log n)^{-\rho})$ (with a matching lower bound on $\delta_{n,2}$).
- Finally, consider $E=(C([0,1]^2),\|\cdot\|_\infty)$ and the law of the Brownian sheet on $I=[0,1]^2$, i.e. the centered continuous Gaussian process $(X_t)_{t\in I}$ such that

  $$\mathbb{E}X_sX_t=(s_1\wedge t_1)(s_2\wedge t_2)$$

  if $s=(s_1,s_2)$, $t=(t_1,t_2)$. Quoting again from [26], we get

  $$ct^{-2}\log(1/t)^3\lesssim\psi(t)\lesssim c't^{-2}\log(1/t)^3.$$

  We obtain $\mathbb{E}W_2(L_n,\mu)=\mathrm{O}((\log n)^{-1/2}(\log\log n)^{3/2})$ (and a matching lower bound on $\delta_{n,2}$).

Many more results may be readily obtained from the literature. References are given e.g. in the two articles we quoted. One may also take advantage directly of some estimates of the Kolmogorov entropy of Gaussian processes (i.e. covering numbers of the Cameron–Martin ball), bypassing small ball estimates. Such direct estimates are provided for example in [27].

We leave aside the question of determining the sharp asymptotics for the average error $\mathbb{E}(W_2(L_n,\mu))$, that is of finding whether there exists $c>0$ such that

$$\mathbb{E}\big(W_2(L_n,\mu)\big)\sim c\psi^{-1}(\log n).$$

Let us underline that the corresponding question for the quantization rate is tackled for example in [26]. In this paper, instead of connecting the quantization rate to small deviation asymptotics, a truncation of the Karhunen–Loève expansion of the Gaussian vector is used. Proving the existence of c as above and computing its value on classical Gaussian processes would allow a sharp comparison of the relative asymptotic performance of empirical measures and optimal quantizers.

1.3. *The case of Markov chains*

We wish to extend the control of the speed of convergence to weakly dependent sequences, such as rapidly-mixing Markov chains. There is a natural incentive to consider this question: there are cases when one does not know how to sample from a given measure π, but a Markov chain with stationary measure π is nevertheless available for simulation. This is the basic set-up of the Markov Chain Monte Carlo framework, and a very frequent situation, even in finite dimension.

When looking at the proof of Theorem 1.1, it is apparent that the main ingredient missing in the dependent case is the argument following (19), i.e. that whenever $A \subset X$ is measurable, $nL_n(A)$ follows a binomial law with parameters n and $\mu(A)$, and this must be remedied in some way. It is natural to look for some type of quantitative ergodicity property of the chain, expressing almost-independence of X_i and X_j in the long range ($|i-j|$ large).

We will consider decay-of-variance inequalities of the following form:

$$\mathrm{Var}_\pi P^n f \leq C\lambda^n \mathrm{Var}_\pi f. \tag{7}$$

In the reversible case, a bound of the type of (7) is ensured by Poincaré or spectral gap inequalities. We recall one possible definition in the discrete-time Markov chain setting.

Definition 1.2. *Let P be a Markov kernel with reversible measure $\pi \in \mathcal{P}(E)$. We say that a Poincaré inequality with constant $C_P > 0$ holds if*

$$\mathrm{Var}_\pi f \leq C_P \int f\left(I - P^2\right) f \,\mathrm{d}\pi \tag{8}$$

for all $f \in L^2(\pi)$.

If (8) holds, we have

$$\mathrm{Var}_\pi P^n f \leq \lambda^n \mathrm{Var}_\pi f$$

with $\lambda = (C_P - 1)/C_P$.

The choice of assumption (7) is fairly standard. More generally, one may assume that we have a control of the decay of the variance in the following form:

$$\mathrm{Var}_\pi P^n f \leq C\lambda^n \left\| f - \int f \,\mathrm{d}\pi \right\|_{L^p}. \tag{9}$$

As soon as $p > 2$, these inequalities are weaker than (7). Our proof would be easily adaptable to this weaker decay-of-variance setting. We do not provide a complete statement of this claim.

For a discussion of the links between Poincaré inequality and other notions of weak dependence (e.g. mixing coefficients), see the recent paper [7].

For the next two theorems, we make the following dimension assumption on E: there exists $k_E > 0$ and $\alpha > 0$ such that for all $S \subset E$ with finite diameter,

$$N(S, \delta) \leq k_E (\mathrm{Diam}\, S/\delta)^\alpha. \tag{10}$$

The following theorem is the analogue of Corollary 1.2 under the assumption that the Markov chain satisfies a decay-of-variance inequality.

Theorem 1.5. *Assume that E has finite diameter $\Delta > 0$ and (10) holds. Let $\pi \in \mathcal{P}(E)$, and let $(X_i)_{i\geq 0}$ be an E-valued Markov chain with initial law ν such that π is its unique invariant probability. Assume also that (7) holds for some $C > 0$ and $\lambda < 1$.*

Then if $2p < \alpha(1+1/r)$ *and* L_n *denotes the occupation measure* $1/n \sum_{i=1}^n \delta_{X_i}$, *the following holds*:

$$\mathbb{E}_\nu\big[W_p(L_n,\pi)\big] \le C_1 k_E^{(1/\alpha(1+1/r))} \Delta\left(\frac{C\|\mathrm{d}\nu/\mathrm{d}\pi\|_{L^r(\pi)}}{(1-\lambda)n}\right)^{1/[\alpha(1+1/r)]}$$

with $C_1 \le 64/3(\frac{2p}{\alpha(1+1/r)-2p})^{2p/(\alpha(1+1/r))}$.

Our next theorem is an extension to the unbounded case under some moment conditions on π.

Theorem 1.6. *Assume that* (10) *holds. Let* $\pi \in \mathcal{P}(E)$, *and let* $(X_i)_{i\ge 0}$ *be an* E*-valued Markov chain with initial law* ν *such that* π *is its unique invariant probability. Assume also that* (7) *holds for some* $C > 0$ *and* $\lambda < 1$. *Let* $x_0 \in E$ *and for all* $\theta \ge 1$, *denote* $M_\theta = \int d(x_0,x)^\theta \,\mathrm{d}\pi$. *Fix* r *and assume* $2p < \alpha(1+1/r)$. *Assume that* π *admits a finite moment of order* $q > 2p/(1-1/r) \vee \alpha p/(\alpha - p)$.

Then there exists $C_1 > 0$ *depending on* E, α, p, r, q *and on* M_q *such that*

$$\mathbb{E}_\nu\big[W_p(L_n,\pi)\big] \le C_1\left(\frac{C\|\mathrm{d}\nu/\mathrm{d}\pi\|_{L^r(\pi)}}{(1-\lambda)n}\right)^{1/[\alpha(1+1/r)]}.$$

Remark. *As in the case of Corollary* 1.3, *a more precise expression may be found in the proof.*

To conclude this section, we include without proof a possible variant of the results above. We no longer assume that $(X_n)_{n\ge 0}$ is the trajectory of a Markov chain, but instead that $(X_n)_{n\in\mathbb{Z}}$ is a general π-stationary sequence of variables with controlled ρ-mixing coefficients. Remember that the ρ-mixing coefficient of two sub-σ-algebras $\mathcal{F}$ and $\mathcal{G}$ over a common probability space is

$$\rho(\mathcal{F},\mathcal{G}) = \sup_{X\in L^2(\mathcal{F}), Y\in L^2\mathcal{G}} \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}},$$

and that the sequence of ρ-mixing coefficients for the π-stationary sequence $(X_n)_{n\in\mathbb{Z}}$ is the sequence given by $\rho_0 = 1$ and

$$\rho_n = \rho\big(\sigma(X_s, s\le 0), \sigma(X_t, t\ge n)\big)$$

for $n \ge 1$. The proof of the following proposition may be obtained along the same lines as in the Markov case.

Proposition 1.7. *Assume that* (X_n) *is a* π*-stationary sequence as above, such that* $\rho_n \to 0$ *as* $n \to +\infty$. *Define*

$$\chi_n = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^i \rho_j.$$

Then if $\Delta = \mathrm{Diam}\, E$ *and* $t > 0$ *is fixed*, *we have*

$$\mathbb{E} W_p(L_n,\pi) \le c\left(t + \chi_n^{1/2p}\int_t^{\Delta/4} N(E,\delta)^{1/2p}\,\mathrm{d}\delta\right).$$

If for example E *has finite-dimensional type with parameter* α *as defined earlier*, *we would get a convergence rate of the form*

$$\mathbb{E} W_p(L_n,\pi) \le c\Delta\chi_n^{1/\alpha}.$$

Remark. *If* (ρ_n) *is exponentially decaying we retrieve our more usual case since* χ_n *is of order* n^{-1}.

2. Proofs in the independent case

Lemma 2.1. *Let $S \subset E$, $s > 0$ and $u, v \in \mathbb{N}$ with $u < v$. Suppose that $N(S, 4^{-v}s) < +\infty$. For $u \le j \le v$, there exist integers*

$$m(j) \le N\big(S, 4^{-j}s\big) \tag{11}$$

and non-empty subsets $S_{j,l}$ of S, $u \le j \le v$, $1 \le l \le m(j)$, such that the sets $S_{j,l}$ $1 \le l \le m(j)$ satisfy

(1) *for each j, $(S_{j,l})_{1\le l\le m(j)}$ is a partition of S,*
(2) $\operatorname{Diam} S_{j,l} \le 4^{-j+1}s$,
(3) *for each $j > u$, for each $1 \le l \le m(j)$ there exists $1 \le l' \le m(j-1)$ such that $S_{j,l} \subset S_{j-1,l'}$.*

In other words, the sets $S_{j,l}$ form a sequence of partitions of S that get coarser as j decreases (tiles at the scale $j-1$ are unions of tiles at the scale j).

Proof. We begin by picking a set of balls $B_{j,l} = B(x_{j,l}, 4^{-j}s) \cap S$ with $u \le j \le v$ and $1 \le l \le N(S, 4^{-j}s)$, such that for all j,

$$S \subset \bigcup_{l=1}^{N(S,4^{-j}s)} B_{j,l}.$$

Define $S_{v,1} = B_{v,1}$, and successively set

$$S_{v,l} = B_{v,l} \Big\backslash \bigcup_{1\le l'\le l-1} S_{v,l'}.$$

Discard the possible empty sets and relabel the existing sets accordingly. We have obtained the finest partition, obviously satisfying conditions (1)–(2).

Assume now that the sets $S_{j,l}$ have been built for $k+1 \le j \le v$. Set $S_{k,1}$ to be the reunion of all $S_{k+1,l'}$ such that $S_{k+1,l'} \cap B_{k,1} \ne \varnothing$. Likewise, define by induction on l the set $S_{k,l}$ as the reunion of all $S_{k+1,l'}$ such that $S_{k+1,l'} \cap B_{k,l} \ne \varnothing$ and $S_{k+1,l'} \not\subset S_{k,p}$ for $1 \le p < l$. Again, discard the possible empty sets and relabel the remaining tiles. It is readily checked that the sets obtained satisfy assumptions (1) and (3). We check assumption (2): let $x_{k,l}$ denote the center of $B_{k,l}$ and let $y \in S_{k+1,l'} \subset S_{k,l}$. We have

$$d(x_{k,l}, y) \le 4^{-k}s + \operatorname{Diam} S_{k+1,l'} \le 2 \times 4^{-k}s,$$

thus $\operatorname{Diam} S_{k,l} \le 4^{-k+1}s$ as desired. □

Consider as above a subset S of E with finite diameter Δ_S, and assume that $N(S, 4^{-k}\Delta_S) < +\infty$. Pick a sequence of partitions $(S_{j,l})_{1\le l\le m(j)}$ for $1 \le j \le k$, as per Lemma 2.1. For each (j,l) choose a point $x_{j,l} \in S_{j,l}$. Define the set of points of level j as the set $L(j) = \{x_{j,l}\}_{1\le l\le m(j)}$. Say that $x_{j',l'}$ is an ancestor of $x_{j,l}$ if $S_{j,l} \subset S_{j',l'}$: we will denote this relation by $(j,l) \preceq (j',l')$.

The next two lemmas study the cost of transporting a finite measure m_k to another measure n_k when these measures have support in $L(k)$. The underlying idea is that we consider the finite metric space

$$\mathcal{T} = \big\{x_{j,l}, 1 \le l \le m(j), 1 \le j \le k\big\}$$

as a *metric tree*, where the ancestry relationship $\preceq$ defined above corresponds to the hierarchical structure of the tree. This tree is naturally endowed with a metric by considering that the distance from a point to its child (i.e. one-step descendant) is given by their distance on the original space (E,d), and the distance between two points of $\mathcal{T}$ is the sum of distances on the unique tree path that joins them. By the triangle inequality it is immediate to see that this tree metric dominates the metric inherited from (E,d). We consider the problem of transportation between two masses at the leaves of the tree. The transportation algorithm we consider consists in allocating as much mass as possible at each point, then moving the remaining mass up one level in the tree, and iterating the procedure.

A technical warning: please note that the transportation cost is usually defined between two probability measures; however there is no difficulty in extending its definition to the transportation between two finite measures of equal total mass, and we will freely use this fact in the sequel.

Lemma 2.2. *Let m_j, n_j be measures with support in L_j with same mass. Define the measures $\tilde{m}_{j-1}$ and $\tilde{n}_{j-1}$ on L_{j-1} by setting*

$$\tilde{m}_{j-1}(x_{j-1,l'}) = \sum_{(j,l)\preceq(j-1,l')} \bigl(m_j(x_{j,l}) - n_j(x_{j,l})\bigr) \vee 0, \tag{12}$$

$$\tilde{n}_{j-1}(x_{j-1,l'}) = \sum_{(j,l)\preceq(j-1,l')} \bigl(n_j(x_{j,l}) - m_j(x_{j,l})\bigr) \vee 0. \tag{13}$$

The measures $\tilde{m}_{j-1}$ and $\tilde{n}_{j-1}$ have same mass, so the transportation cost between them may be defined. Moreover, if $\Delta_S = \operatorname{Diam} S$, the following bound holds:

$$W_p(m_j, n_j) \le 2 \times 4^{-j+2} \Delta_S \|m_j - n_j\|_{\mathrm{TV}}^{1/p} + W_p(\tilde{m}_{j-1}, \tilde{n}_{j-1}). \tag{14}$$

Proof. Set $m_j \wedge n_j(x_{j,l}) = m_j(x_{j,l}) \wedge n_j(x_{j,l})$. We introduce the measure $m_j \wedge n_j + \tilde{m}_{j-1}$ defined by

$$[m_j \wedge n_j + \tilde{m}_{j-1}](x_{j,l}) = m_j \wedge n_j(x_{j,l}),$$
$$[m_j \wedge n_j + \tilde{m}_{j-1}](x_{j-1,l'}) = \tilde{m}_{j-1}(x_{j-1,l'}).$$

Likewise, we define the measure $m_j \wedge n_j + \tilde{n}_{j-1}$. We stress that they have the same mass as m_j, n_j.

By the triangle inequality,

$$W_p(m_j, n_j) \le W_p(m_j, m_j \wedge n_j + \tilde{m}_{j-1}) + W_p(m_j \wedge n_j + \tilde{m}_{j-1}, m_j \wedge n_j + \tilde{n}_{j-1}) + W_p(m_j \wedge n_j + \tilde{n}_{j-1}, n_j).$$

We bound the term on the left. Introduce the transport plan π_m defined by

$$\pi_m(x_{j,l}, x_{j,l}) = m_j \wedge n_j(x_{j,l}),$$
$$\pi_m(x_{j,l}, x_{j-1,l'}) = \bigl(m_j(x_{j,l}) - n_j(x_{j,l})\bigr)_+ \quad \text{when } (j,l) \preceq \bigl(j-1,l'\bigr).$$

The reader can check that $\pi_m \in \mathcal{P}(m_j, m_j \wedge n_j + \tilde{m}_{j-1})$. Moreover,

$$W_p(m_j, m_j \wedge n_j + \tilde{m}_{j-1}) \le \left(\int d^p(x,y)\pi_m(\mathrm{d}x, \mathrm{d}y)\right)^{1/p} \le 4^{-j+2}\Delta_S \left(\sum_{l=1}^{m(j)} \bigl(m_j(x_{j,l}) - n_j(x_{j,l})\bigr)_+\right)^{1/p}.$$

Likewise,

$$W_p(n_j, m_j \wedge n_j + \tilde{n}_{j-1}) \le 4^{-j+2}\Delta_S \left(\sum_{l=1}^{m(j)} \bigl(n_j(x_{j,l}) - m_j(x_{j,l})\bigr)_+\right)^{1/p}.$$

As for the term in the middle, it is bounded by $W_p(\tilde{m}_{j-1}, \tilde{n}_{j-1})$: indeed, this follows by considering a transport plan that leaves the mass $m_j \wedge n_j$ in place and optimally maps $\tilde{m}_{j-1}$ towards $\tilde{n}_{j-1}$. Putting this together and using the inequality $x + y \le 2^{1-1/p}(x^p + y^p)^{1/p}$, we get

$$W_p(m_j, n_j) \le 2^{1-1/p} 4^{-j+2} \Delta_S \left(\sum_{l=1}^{m(j)} \left| m_j(x_{j,l}) - n_j(x_{j,l}) \right| \right)^{1/p} + W_p(\tilde{m}_{j-1}, \tilde{n}_{j-1}).$$

□

Lemma 2.3. *Let m_j, n_j be measures with support in L_j. Define for $1 \le i < j$ the measures $\mathfrak{m}_i, \mathfrak{n}_i$ with support in L_i by*

$$\mathfrak{m}_i(x_{i,l'}) = \sum_{(j,l) \preceq (i,l')} m_j(x_{j,l}), \qquad \mathfrak{n}_i(x_{i,l'}) = \sum_{(j,l) \preceq (i,l')} n_j(x_{j,l}). \tag{15}$$

The following bound holds:

$$W_p(m_j, n_j) \le \sum_{i=1}^{j} 2 \times 4^{-i+2} \Delta_S \|\mathfrak{m}_i - \mathfrak{n}_i\|_{\mathrm{TV}}^{1/p}. \tag{16}$$

Proof. We proceed by induction on j. For $j = 1$, the result is obtained by using the simple bound $W_p(m_1, n_1) \le \Delta_S \|m_1 - n_1\|_{\mathrm{TV}}^{1/p}$.

Suppose that (16) holds for measures with support in L_{j-1}. By Lemma 2.2, we have

$$W_p(m_j, n_j) \le 2 \times 4^{-j+2} \Delta_S \|m_j - n_j\|_{\mathrm{TV}}^{1/p} + W_p(\tilde{m}_{j-1}, \tilde{n}_{j-1}),$$

where $\tilde{m}_{j-1}$ and $\tilde{n}_{j-1}$ are defined by (12) and (13) respectively. For $1 \le i < j - 1$, define following (15)

$$\tilde{\mathfrak{m}}_i(x_{i,l'}) = \sum_{(j-1,l) \preceq (i,l')} \tilde{m}_{j-1}(x_{j-1,l}), \qquad \tilde{\mathfrak{n}}_i(x_{i,l'}) = \sum_{(j-1,l) \preceq (i,l')} \tilde{n}_{j-1}(x_{j-1,l}).$$

We have by induction hypothesis

$$W_p(m_j, n_j) \le 2 \times 4^{-j+2} \Delta_S \|m_j - n_j\|_{\mathrm{TV}}^{1/p} + \sum_{i=1}^{j-1} 2 \times 4^{-i+2} \Delta_S \|\tilde{\mathfrak{m}}_i - \tilde{\mathfrak{n}}_i\|_{\mathrm{TV}}^{1/p}.$$

To conclude, it suffices to check that for $1 \le i \le j - 1$, $\|\tilde{\mathfrak{m}}_i - \tilde{\mathfrak{n}}_i\|_{\mathrm{TV}} = \|\mathfrak{m}_i - \mathfrak{n}_i\|_{\mathrm{TV}}$. □

Proof of Theorem 1.1. We pick some positive integer k whose value will be determined further on. Introduce the sequence of partitions $(S_{j,l})_{1 \le l \le m(j)}$ for $0 \le j \le k$ as in the lemmas above, as well as the points $x_{j,l}$. Define μ_k as the measure with support in $L(k)$ such that $\mu_k(x_{k,l}) = \mu(S_{k,l})$ for $1 \le l \le m(k)$. The diameter of the sets $S_{k,l}$ is bounded by $4^{-k+1}\Delta_S$, therefore $W_p(\mu, \mu_k) \le 4^{-k+1}\Delta_S$.

Let L_n^k denote the empirical measure associated to μ_k.

For $0 \le j \le k - 1$, define as in Lemma 2.3 the measures μ_j and L_n^j with support in $L(j)$ by

$$\mu_j(x_{j,l'}) = \sum_{(k,l) \preceq (j,l')} \mu_k(x_{k,l}), \tag{17}$$

$$L_n^j(x_{j,l'}) = \sum_{(k,l) \preceq (j,l')} L_n^k(x_{k,l}). \tag{18}$$

It is simple to check that $\mu_j(x_{j,l}) = \mu(S_{j,l})$, and that L_n^j is the empirical measure associated with μ_j. Applying (16), we get

$$W_p\big(\mu_k, L_n^k\big) \le \sum_{j=1}^{k} 2 \times 4^{-j+2} \Delta_S \big\|\mu_j - L_n^j\big\|_{\mathrm{TV}}^{1/p}. \tag{19}$$

Observe that $nL_n^j(x_{j,l})$ is a binomial law with parameters n and $\mu(S_{j,l})$. The expectation of $\|\mu_j - L_n^j\|_{\mathrm{TV}}$ is bounded as follows:

$$\begin{aligned}
\mathbb{E}\big(\big\|\mu_j - L_n^j\big\|_{\mathrm{TV}}\big) &= 1/2 \sum_{l=1}^{m(j)} \mathbb{E}\big(\big|\big(L_n^j - \mu_j\big)(x_{j,l})\big|\big) \\
&\le 1/2 \sum_{l=1}^{m(j)} \sqrt{\mathbb{E}\big(\big|\big(L_n^j - \mu_j\big)(x_{j,l})\big|^2\big)} \\
&= 1/2 \sum_{l=1}^{m(j)} \sqrt{\frac{\mu(S_{j,l})(1-\mu(S_{j,l}))}{n}} \\
&\le 1/2 \sqrt{\frac{m(j)}{n}}.
\end{aligned}$$

In the last inequality, we use Cauchy–Schwarz's inequality and the fact that $(S_{j,l})_{1\le l\le m(j)}$ is a partition of S. Plugging this back in (19), we get

$$\begin{aligned}
\mathbb{E}\big(W_p\big(\mu_k, L_n^k\big)\big) &\le n^{-1/2p} \sum_{j=1}^{k} 2^{1-1/p} 4^{(-j+2)} \Delta_S m(i)^{1/2p} \\
&\le 2^{5-1/p} n^{-1/2p} \sum_{j=1}^{k} 4^{-j} \Delta_S N\big(S, 4^{-j}\Delta_S\big)^{1/2p} \\
&\le 2^{6-1/p}/3 n^{-1/2p} \int_{4^{-(k+1)}\Delta_S}^{\Delta_S/4} N(S,\delta)^{1/2p}\, d\delta.
\end{aligned}$$

In the last line, we use a standard sum-integral comparison argument.

By the triangle inequality, we have

$$W_p(\mu, L_n) \le W_p(\mu, \mu_k) + W_p\big(\mu_k, L_n^k\big) + W_p\big(L_n^k, L_n\big).$$

We claim that $\mathbb{E}(W_p(L_n^k, L_n)) \le W_p(\mu, \mu_k)$. Indeed, choose n i.i.d. couples (X_i, X_i^k) such that $X_i \sim \mu$, $X_i^k \sim \mu_k$, and the joint law of (X_i, X_i^k) achieves an optimal coupling, i.e. $\mathbb{E}|X_i - X_i^k|^p = W_p^p(\mu, \mu^k)$. Observe that existence of this optimal coupling is guaranteed e.g. by Theorem 4.1 in [32]. We have the identities in law

$$L_n \sim \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad L_n^k \sim \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i^k}.$$

Choose the transport plan that sends X_i to X_i^k: this gives the upper bound

$$W_p^p\big(L_n, L_n^k\big) \le 1/n \sum_{i=1}^{n} \big|X_i - X_i^k\big|^p$$

and passing to expectation proves our claim.

Thus, $\mathbb{E}(W_p(\mu, L_n)) \le 2W_p(\mu, \mu_k) + \mathbb{E}(W_p(\mu_k, L_n^k))$. Choose now k as the largest integer such that $4^{-k}\Delta_S > t$. This imposes $4^{-k+1}\Delta_S \le 16t$, and this finishes the proof. □

Proof of Corollary 1.2. The proof simply consists in plugging the bound 2 on covering numbers in the estimate of Theorem 1.1, and choosing an optimal $t > 0$. □

Lemma 2.4. *Let $1 \le p < +\infty$, $\mu, \nu \in \mathcal{P}_p(E)$, and for $i \ge 1$, let $\mu_i, \nu_i \in \mathcal{P}_p(E)$, $\lambda_i \ge 0$ be such that $\sum_{i\ge 1} \lambda_i = 1$ and*

$$\sum_{i\ge 1} \lambda_i \mu_i = \mu, \qquad \sum_{i\ge 1} \lambda_i \nu_i = \nu.$$

Then the following bound holds:

$$W_p(\mu, \nu) \le \sum_{i\ge 1} \lambda_i^{1/p} W_p(\mu_i, \nu_i).$$

Proof. Let us pick random variables (X_i, Y_i) realizing optimal couplings of μ_i, ν_i for all $i \ge 1$. Define a random variable I over $\mathbb{N}$, independent of the previous variables, such that $\mathbb{P}(I = i) = \lambda_i$. Then the variables μ_I, ν_I have respective laws μ and ν.

By definition of W_p,

$$\begin{aligned} W_p(\mu, \nu) &\le \left[\mathbb{E} d^p(X_I, Y_I)\right]^{1/p} \\ &= \left[\mathbb{E}\left(\sum_{i\ge 1} \mathbf{1}_{I=i} d(X_i, Y_i)\right)^p\right]^{1/p} \\ &\le \sum_{i\ge 1} \left[\mathbb{E}\mathbf{1}_{I=i} d^p(X_i, Y_i)\right]^{1/p} = \sum_{i\ge 1} \lambda_i^{1/p} W_p(\mu_i, \nu_i). \end{aligned}$$

Here we have used the triangle inequality for L^p-norms to go from the second to the third line. □

Proof of Corollary 1.3. Let r_i, $i \ge 1$, be an increasing sequence of positive numbers to be specified later. We will use a decomposition of the space E into the union of rings

$$K_1 = B(x_0, r_1), \qquad K_i = B(x_0, r_i) \setminus B(x_0, r_{i-1}) \quad \text{for } i \ge 2.$$

Define the conditional measures

$$\mu^{K_i} = \frac{\mathbf{1}_{K_i}\mu}{\mu(K_i)}, \qquad L_n^i = \frac{\mathbf{1}_{K_i} L_n}{L_n(K_i)}$$

(whenever $L_n(K_i) \ne 0$). Notice that

$$\begin{aligned} L_n &= \sum_{i\ge 1} L_n(K_i) L_n^i, \\ \mu &= \sum_{i\ge 1} \mu(K_i) \mu^{K_i}, \end{aligned}$$

which we rewrite as

$$\begin{aligned} L_n &= \sum_{i\ge 1} L_n(K_i) \wedge \mu(K_i) L_n^i + \sum_{i\ge 1} \left[\left(L_n(K_i) - \mu(K_i)\right) \vee 0\right] L_n^i, \\ \mu &= \sum_{i\ge 1} L_n(K_i) \wedge \mu(K_i) \mu^{K_i} + \sum_{i\ge 1} \left[\left(\mu(K_i) - L_n(K_i)\right) \vee 0\right] \mu^{K_i}. \end{aligned}$$

In the two preceding relations, we make the convention that whenever $L_n(K_i)=0$, we set $0\times L_n^i=0$, so that no terms are ill-defined. Set

$$\lambda=\sum_{i\ge 1}\big(L_n(K_i)-\mu(K_i)\big)\vee 0=\sum_{i\ge 1}\big(\mu(K_i)-L_n(K_i)\big)\vee 0,$$
$$\lambda_i=L_n(K_i)\wedge\mu(K_i)$$

(the two definitions of λ agree since $\sum_{i\ge 1}L_n(K_i)=\sum_{i\ge 1}\mu(K_i)=1$). Set

$$\mathfrak{m}=\frac{1}{\lambda}\sum_{i\ge 1}\big[\big(L_n(K_i)-\mu(K_i)\big)\vee 0\big]L_n^i,$$
$$\mathfrak{n}=\sum_{i\ge 1}\big[\big(\mu(K_i)-L_n(K_i)\big)\vee 0\big]\mu^{K_i},$$

so that

$$L_n=\sum_{i\ge 1}\lambda_i L_n^i+\lambda\mathfrak{m},$$
$$\mu=\sum_{i\ge 1}\lambda_i\mu^{K_i}+\lambda\mathfrak{n}.$$

According to Lemma 2.4, we have

$$\begin{aligned}W_p(L_n,\mu)&\le\sum_{i\ge 1}\lambda_i^{1/p}W_p\big(L_n^i,\mu^{K_i}\big)+\lambda^{1/p}W_p(\mathfrak{m},\mathfrak{n})\\&\le\sum_{i\ge 1}\lambda_i^{1/p}W_p\big(L_n^i,\mu^{K_i}\big)+\lambda^{1/p}W_p(\mu,\mathfrak{m})+\lambda^{1/p}W_p(\mu,\mathfrak{n}).\end{aligned}$$

We take expectations and bound the terms in the right-hand side. Let us start with the second one. With another application of Lemma 2.4, we get

$$\lambda^{1/p}W_p(\mu,\mathfrak{m})\le\sum_{i\ge 1}\big[\big(L_n(K_i)-\mu(K_i)\big)\vee 0\big]^{1/p}W_p\big(\mu,L_n^i\big).$$

We bound the Wasserstein distance with the classical inequality

$$W_p\big(\mu,L_n^i\big)\le M_p^{1/p}+\left(\int d(x_0,x)\,\mathrm{d}L_n^i(x)\right)^{1/p}\le M_p^{1/p}+r_i.$$

Taking expectations, and choosing some $\xi\ge 1$, we have

$$\begin{aligned}\mathbb{E}\lambda^{1/p}W_p(\mu,\mathfrak{m})&\le\sum_{i\ge 1}\big(\mathbb{E}|L_n(K_i)-\mu(K_i)|^2\big)^{1/2p}\big(M_p^{1/p}+r_i\big)\\&\le\sum_{i\ge 1}\left(\frac{\mu(K_i)}{n}\right)^{1/2p}\big(M_p^{1/p}+r_i\big)\\&\le n^{-1/2p}\left[M_p^{1/p}+r_1+\sum_{i\ge 2}\frac{{M_{2\xi p}}^{1/2p}}{r_{i-1}^{\xi}}\big(M_p^{1/p}+r_i\big)\right].\end{aligned}$$

We have used Jensen's inequality in the first line, the fact that $nL_n(K_i)$ is binomial in the second line and Markov's inequality in the third line. Set $r_i = 2^i$: we get

$$\mathbb{E}\lambda^{1/p} W_p(\mu, \mathfrak{m}) \le n^{-1/2p} C(\mu, \xi),$$

where

$$C(\mu, \xi) = M_p^{1/p} + 2 + M_{2\xi p}^{1/2p}\left(M_p^{1/p}\frac{2^{-\xi}}{1-2^{-\xi}} + \frac{2^{1-\xi}}{1-2^{1-\xi}}\right).$$

The third term is bounded likewise and this yields the same bound. We turn our attention to the first term. In order to control it, we will apply our result for bounded spaces on the spaces K_i, conditionally on the value of $L_n(K_i)$. Let us bound λ_i with $L_n(K_i)$: we get

$$\begin{aligned}\mathbb{E}\lambda_i^{1/p} W_p\big(L_n^i, \mu^{K_i}\big) &\le \mathbb{E}\big[L_n(K_i)^{1/p} c(E, p, \alpha) r_i \big(nL_n(K_i)\big)^{-1/\alpha}\big] \\ &\le 2^i \mathbb{E}\big(L_n(K_i)^{1/p-1/\alpha}\big) c(E, p, \alpha) n^{-1/\alpha} \\ &\le 2^i \mu(K_i)^{(\alpha-p)/\alpha p} c(E, p, \alpha) n^{-1/\alpha}.\end{aligned}$$

Here $c(E, p, \alpha)$ is the constant given in Corollary 1.2. By an application of Markov's inequality, picking $\zeta > 1$ yields

$$\begin{aligned}\sum_{i\ge1} 2^i \mu(K_i)^{(\alpha-p)/\alpha p} &\le 1 + M_{\zeta\alpha p/(\alpha-p)}^{(\alpha-p)/\alpha p} \sum_{i\ge2} 2^\zeta 2^{i(1-\zeta)} \\ &= 1 + M_{\zeta\alpha p/(\alpha-p)}^{(\alpha-p)/\alpha p} \frac{2}{1-2^{1-\zeta}}.\end{aligned}$$

This implies that the first term is bounded as follows:

$$\sum_{i\ge1} \mathbb{E}\lambda_i^{1/p} W_p\big(L_n^i, \mu^{K_i}\big) \le C'(E, p, \alpha, \mu, \zeta) n^{-1/\alpha},$$

where

$$C'(E, p, \alpha, \mu, \zeta) = c(E, p, \alpha) M_{\zeta\alpha p/(\alpha-p)}^{(\alpha-p)/\alpha p} \frac{2}{1-2^{1-\zeta}}.$$

This completes the proof. □

3. Proof of Theorem 1.4

We begin by noticing that statement (6) is a simple consequence of statement (5) and the tensorization of the transportation inequality $\mathbf{T}_2$ (see Appendices A and B): we have by Corollary A.1

$$\mathbb{P}\big(W_2(L_n, \mu) \ge \mathbb{E}\big(W_2(L_n, \mu)\big) + t\big) \le \mathrm{e}^{-nt^2/(2\sigma^2)},$$

and it suffices to choose $t = \lambda\psi^{-1}(\log n)$ to conclude. We now turn to the other claims.

Denote by K the unit ball of the Cameron–Martin space associated to E and μ, and by B the unit ball of E. According to the Gaussian isoperimetric inequality (see [22]), for all $\lambda > 0$ and $\varepsilon > 0$,

$$\mu(\lambda K + \varepsilon B) \ge \Phi\big(\lambda + \Phi^{-1}\big(\mu(\varepsilon B)\big)\big),$$

where $\Phi(t) = \int_{-\infty}^t \mathrm{e}^{-u^2/2}\,\mathrm{d}u/\sqrt{2\pi}$ is the Gaussian c.d.f.

Choose $\lambda > 0$ and $\varepsilon > 0$, and set $S = \lambda K + \varepsilon B$. Note

$$\mu' = \frac{1}{\mu(S)} \mathbf{1}_S \mu$$

the restriction of μ to the enlarged ball.

The diameter of S is bounded by $2(\sigma\lambda + \varepsilon)$. By Theorem 1.1, the W_2 distance between L_n and μ is thus bounded as follows:

$$\mathbb{E} W_2(L_n, \mu) \le 2W_2(\mu, \mu') + ct + cn^{-1/4} \int_t^{(\sigma\lambda+\varepsilon)/2} N(S, \delta)^{1/4} \, \mathrm{d}\delta. \tag{20}$$

Let us denote

$$I_1 = W_2(\mu, \mu'), \tag{21}$$

$$I_2 = t, \tag{22}$$

$$I_3 = n^{-1/4} \int_t^{(\sigma\lambda+\varepsilon)/2} N(S, \delta)^{1/4} \, \mathrm{d}\delta. \tag{23}$$

To begin with, set $\varepsilon = t/2$.

Controlling I_1. We use transportation inequalities and the Gaussian isoperimetric inequality. By Lemma B.1, μ satisfies a $\mathbf{T}_2(2\sigma^2)$ inequality, so that we have

$$\begin{aligned} W_2(\mu, \mu') &\le \sqrt{2\sigma^2 H(\mu'|\mu)} = \sqrt{-2\sigma^2 \log \mu(\lambda K + \varepsilon B)} \\ &\le \sqrt{-2\sigma^2 \log \Phi\bigl(\lambda + \Phi^{-1}\bigl(\mu(\varepsilon B)\bigr)\bigr)} \\ &= \sqrt{2}\sigma \sqrt{-\log \Phi\bigl(\lambda + \Phi^{-1}\bigl(\mathrm{e}^{-\psi(t/2)}\bigr)\bigr)}. \end{aligned}$$

Introduce the tail function of the Gaussian distribution

$$\Upsilon(x) = \sqrt{2\pi}^{-1} \int_x^{+\infty} \mathrm{e}^{-y^2/2} \, \mathrm{d}y.$$

We will use the fact that $\Phi^{-1} + \Upsilon^{-1} = 0$, which comes from symmetry of the Gaussian distribution. We will also use the bound $\Upsilon(t) \le \mathrm{e}^{-t^2/2}/2$, $t \ge 0$ and its consequence

$$\Upsilon^{-1}(u) \le \sqrt{-2\log u}, \quad 0 < u \le 1/2.$$

We have

$$\Phi^{-1}\bigl(\mathrm{e}^{-\psi(t/2)}\bigr) = -\Upsilon^{-1}\bigl(\mathrm{e}^{-\psi(t/2)}\bigr) \ge -\sqrt{2\psi(t/2)}$$

as soon as $\psi(t/2) \ge \log 2$. The elementary bound $\log \frac{1}{1-x} \le 2x$ for $x \le 1/2$ yields

$$\begin{aligned} \sqrt{-2\log \Phi(u)} &= \sqrt{2}\left(\log \frac{1}{1 - \Upsilon(u)}\right)^{1/2} \\ &\le 2\mathrm{e}^{-u^2/4} \end{aligned}$$

whenever $u \ge \Upsilon^{-1}(1/2) = 0$. Putting this together, we have

$$I_1 \le 2\sigma \mathrm{e}^{-(\lambda - \sqrt{2\psi(t/2)})^2/4} \tag{24}$$

whenever

$$\psi(t/2) \geq \log 2 \quad \text{and} \quad \lambda - \sqrt{2\psi(t/2)} \geq 0. \tag{25}$$

Controlling I_3. The term I_3 is bounded by $1/2n^{-1/4}(\sigma\lambda + t/2)N(S,t)^{1/4}$ (just bound the function inside by its value at t, which is minimal). Denote by $k = N(\lambda K, t-\varepsilon)$ the covering number of λK (w.r.t. the norm of E). Let $x_1, \dots, x_k \in K$ be such that union of the balls $B(x_i, t-\varepsilon)$ contains λK. From the triangle inequality we get the inclusion

$$\lambda K + \varepsilon B \subset \bigcup_{i=1}^{k} B(x_i, t).$$

Therefore, $N(S,t) \leq N(\lambda K, t-\varepsilon) = N(\lambda K, t/2)$.

We now use the well-known link between $N(\lambda K, t/2)$ and the small ball function. Lemma 1 in [21] gives the bound

$$N(\lambda K, t/2) \leq \mathrm{e}^{\lambda^2/2+\psi(t/4)} \leq \mathrm{e}^{\lambda^2/2+\kappa\psi(t/2)}$$

so that

$$I_3 \leq \frac{1}{2}(\sigma\lambda + t/2)\mathrm{e}^{\lambda^2/8+(\kappa/4)\psi(t/2)-(1/4)\log n}. \tag{26}$$

Remark that we have used the doubling condition on ψ, so that we require

$$t/4 \leq t_0. \tag{27}$$

Final step. Set now $t = 2\psi^{-1}(a\log n)$ and $\lambda = 2\sqrt{2a\log n}$, with $a > 0$ yet undetermined. Using (24) and (26), we see that there exists a universal constant c such that

$$\begin{aligned}\mathbb{E}\big(W_2(L_n,\mu)\big) \leq c\big[&\psi^{-1}(a\log n) + \sigma\mathrm{e}^{-(a/2)\log n} \\ &+ \big(\sigma\sqrt{a\log n} + \psi^{-1}(a\log n)\big)\mathrm{e}^{[a(1+\kappa/4)-1/4]\log n}\big].\end{aligned}$$

Choose $a = 1/(6+\kappa)$ and assume $\log n \geq (6+\kappa)(\log 2 \vee \psi(1) \vee \psi(2t_0))$. This guarantees that the technical conditions (25) and (27) are enforced, and that $\psi^{-1}(a\log n) \leq 1$. Summing up, we get:

$$\mathbb{E}\big(W_2(L_n,\mu)\big) \leq c\left[\psi^{-1}\left(\frac{1}{6+\kappa}\log n\right) + \left(1 + \sigma\sqrt{\frac{1}{6+\kappa}\log n}\right)n^{-1/(12+2\kappa)}\right].$$

Impose $\log n \geq (6+\kappa)/\sigma^2$: this ensures $\sigma\sqrt{\frac{1}{6+\kappa}\log n} \geq 1$. And finally, there exists some $c > 0$ such that for all $x \geq 1$, $\sqrt{\log x}x^{-1/4} \leq c$: this implies

$$\sqrt{\frac{1}{6+\kappa}\log n}n^{-1/(24+4\kappa)} \leq c.$$

This gives

$$\left(1 + \sigma\sqrt{\frac{1}{6+\kappa}\log n}\right)n^{-1/(12+2\kappa)} \leq c\sigma n^{-1/[4(6+\kappa)]}$$

and the proof is finished.

4. Proofs in the dependent case

We consider hereafter a Markov chain $(X_n)_{n\in\mathbb{N}}$ defined by $X_0 \sim \nu$ and the transition kernel P. Let us denote by

$$L_n = \sum_{i=1}^{n} \delta_{X_i}$$

its occupation measure.

Proposition 4.1. *Suppose that the Markov chain satisfies* (7) *for some* $C > 0$ *and* $\lambda < 1$. *Then the following holds*:

$$\mathbb{E}_\nu\big(W_p(L_n,\pi)\big) \le c\left(t + \left(\frac{C}{(1-\lambda)n}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\right)^{1/2p}\int_t^{\Delta/4} N(E,\delta)^{1/2p(1+1/r)}\,\mathrm{d}\delta\right). \tag{28}$$

Proof. Introduce a sequence of $k \ge 1$ nested partitions $(S_{j,l})_{1\le j\le k}$ as in Lemma 2.1. An application of (16) as in the proof of Theorem 1.1 (see (19)) yields

$$\mathbb{E}\big(W_p(L_n,\pi)\big) \le 2\times 4^{-k+1}\Delta + \sum_{j=1}^{k} 2\times 4^{-j+2}\Delta\left(\sum_{l=1}^{m(j)} \mathbb{E}\big|(L_n-\pi)(S_{j,l})\big|\right)^{1/p}. \tag{29}$$

Let A be a measurable subset of E, and set $f_A(x) = \mathbf{1}_A(x) - \pi(A)$. We have

$$\begin{aligned}\mathbb{E}\big|(L_n-\pi)(A)\big| &= 1/n\mathbb{E}_\nu\left|\sum_{i=1}^{n} f_A(X_i)\right| \\ &\le 1/n\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_\nu\big[f_A(X_i)f_A(X_j)\big]}.\end{aligned}$$

Let $\tilde{p},\tilde{q},r \ge 1$ be such that $1/\tilde{p} + 1/\tilde{q} + 1/r = 1$, and let s be defined by $1/s = 1/\tilde{p} + 1/\tilde{q}$. Now, using Hölder's inequality with r and s,

$$\mathbb{E}_\nu\big[f_A(X_i)f_A(X_j)\big] \le \left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\big(\mathbb{E}_\pi|f_A(X_i)f_A(X_j)|^s\big)^{1/s}.$$

Take $j \ge i$. Use the Markov property and the fact that $f \mapsto Pf$ is a contraction in $L^s(\pi)$ to get

$$\mathbb{E}_\nu\big[f_A(X_i)f_A(X_j)\big] \le \left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\big\|f_A P^{j-i} f_A\big\|_{L^s(\pi)}.$$

Finally, use Hölder's inequality with $\tilde{p},\tilde{q}$: we get

$$\mathbb{E}_\nu\big[f_A(X_i)f_A(X_j)\big] \le \left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\big\|P^{j-i} f_A\big\|_{L^{\tilde{p}}(\pi)}\|f_A\|_{L^{\tilde{q}}(\pi)}. \tag{30}$$

Set $\tilde{p} = 2$ and note that for $1 \le t \le +\infty$, we have $\|f_A\|_{L^t(\pi)} \le 2\pi(A)^{1/t}$. Use (7) applied to the centered function f_A to get

$$\mathbb{E}_\nu\big[f_A(X_i)f_A(X_j)\big] \le 4C\lambda^{j-i}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\pi(A)^{1-1/r},$$

and as a consequence,

$$\mathbb{E}\big|(L_n-\pi)(A)\big| \le \frac{1}{\sqrt{n}}\frac{2\sqrt{2C}}{\sqrt{1-\lambda}}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}^{1/2}\pi(A)^{1/2-1/2r}. \tag{31}$$

Come back to (29): we have

$$\begin{aligned}
\mathbb{E}\big(W_p(L_n,\pi)\big) &\le 2\times 4^{-k+1}\Delta + 32\left(\frac{2\sqrt{2C}}{\sqrt{1-\lambda}}\right)^{1/p}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}^{1/2p} n^{-1/2p} \\
&\quad \times \sum_{j=1}^{k} 4^{-j}\Delta\left(\sum_{l=1}^{m(j)}\pi(X_{j,l})^{1/2-1/2r}\right)^{1/p} \\
&\le 2\times 4^{-k+1}\Delta + c\left(\frac{C}{(1-\lambda)n}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\right)^{1/2p} \\
&\quad \times \sum_{j=1}^{k} 4^{-j}\Delta m(j)^{1/2p(1+1/r)} \\
&\le c\left(t + \left(\frac{C}{(1-\lambda)n}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\right)^{1/2p}\int_t^{\Delta/4} N(E,\delta)^{1/2p(1+1/r)}\,\mathrm{d}\delta\right).
\end{aligned}$$

□

Proof of Theorem 1.5. Use (28) and (10) to get

$$\mathbb{E}W_p(L_n,\mu) \le c\big[t + At^{-\alpha/2p(1+1/r)+1}\big],$$

where

$$A = \frac{2p}{\alpha(1+1/r)}\big(C/(1-\lambda)\big)^{1/2p}\left\|\frac{d\nu}{d\pi}\right\|_{L^r(\pi)}^{1/2p} n^{-1/2p}\Delta^{\alpha/2p(1+1/r)}.$$

Optimizing in t finishes the proof. □

We now move to the proof in the unbounded case.

Proof of Theorem 1.6. We will mimic the proof of Corollary 1.3: we give an outline and indicate where changes are appropriate. First introduce the rings K_i and the conditional measures

$$\pi|_{K_i} = \frac{\mathbf{1}_{K_i}\pi}{\pi(K_i)}, \qquad L_n^i = \frac{\mathbf{1}_{K_i}L_n}{L_n(K_i)}.$$

Also let

$$\begin{aligned}
\lambda_0 &= \sum_{i\ge 1}\big(L_n(K_i)-\pi(K_i)\big)\vee 0 = \sum_{i\ge 1}\big(\pi(K_i)-L_n(K_i)\big)\vee 0, \\
\lambda_i &= L_n(K_i)\wedge\pi(K_i)
\end{aligned}$$

and

$$\begin{aligned}
\mathfrak{m} &= \frac{1}{\lambda_0}\sum_{i\ge 1}\big[\big(L_n(K_i)-\pi(K_i)\big)\vee 0\big]L_n^i, \\
\mathfrak{n} &= \sum_{i\ge 1}\big[\big(\pi(K_i)-L_n(K_i)\big)\vee 0\big]\pi|_{K_i}.
\end{aligned}$$

Following the proof of Corollary 1.3 we get

$$W_p(L_n,\pi)\le\sum_{i\ge1}\lambda_i^{1/p}W_p\big(L_n^i,\pi|_{K_i}\big)+\lambda_0^{1/p}W_p(\pi,\mathfrak{m})+\lambda_0^{1/p}W_p(\pi,\mathfrak{n}).$$

We will take expectations and bound the terms separately. The second term is bounded as follows:

$$\mathbb{E}\lambda_0^{1/p}W_p(\mu,\mathfrak{m})\le\sum_{i\ge1}\big(\mathbb{E}|L_n(K_i)-\mu(K_i)|^2\big)^{1/2p}\big(M_p^{1/p}+r_i\big).$$

Here we must depart from the independent case. Using instead relation (31), we have

$$\big(\mathbb{E}|(L_n-\pi)(K_i)|^2\big)^{1/2p}\le n^{-1/2p}\left(\frac{8C}{1-\lambda}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\right)^{1/2p}\pi(K_i)^{(1-1/r)/2p}.$$

Take $r_i=2^i$, set $\xi>1$ and use Markov's inequality. After summation, we see that

$$\mathbb{E}\lambda_0^{1/p}W_p(\mu,\mathfrak{m})\le\left(\frac{8C}{(1-\lambda)n}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right\|_{L^r(\pi)}\right)^{1/2p}C(\mu,p,r,\xi)\tag{32}$$

where

$$C(\mu,p,r,\xi)=M_p^{1/p}+2+M_{2p\xi/(1-1/r)}^{(1-1/r)/2p}\left[M_p^{1/p}\frac{2^{-\xi}}{1-2^{-\xi}}+\frac{2^{1-\xi}}{1-2^{1-\xi}}\right].$$

The third term is bounded identically. As for the first one, it will require a little more work. For now fix $i\ge1$ and $k\ge1$, and introduce nested tesselations $(S_{j,l})_{1\le j\le k}$ of K_i as in the proof of Proposition 4.1. Following the line of reasoning of this proof, we obtain

$$\begin{aligned}\mathbb{E}\big(\lambda_i^{1/p}W_p\big(L_n^i,\pi|_{K_i}\big)\big)&\le 2\times4^{-k+1}r_i\pi(K_i)^{1/p}\\&\quad+\sum_{j=1}^{k}2\times4^{-j+2}r_i\left(\sum_{l=1}^{m(j)}\mathbb{E}L_n(K_i)\big|\big(L_n^i-\pi|_{K_i}\big)(S_{j,l})\big|\right)^{1/p}.\end{aligned}$$

Let us deal with the term inside parentheses. Observe that

$$\begin{aligned}L_n(K_i)\big|\big(L_n^i-\pi|_{K_i}\big)(S_{j,l})\big|&=\left|\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_{S_{j,l}}(X_t)-\frac{L_n(K_i)}{\pi(K_i)}\pi(S_{j,l})\right|\\&\le\left|\frac{1}{n}\sum_{t=1}^{n}\big(\mathbf{1}_{S_{j,l}}(X_t)-\pi(S_{j,l})\big)\right|+\pi(S_{j,l})\left|\frac{L_n(K_i)}{\pi(K_i)}-1\right|.\end{aligned}$$

After summing over all l and taking expectations, we get

$$\sum_{l=1}^{m(j)}\mathbb{E}L_n(K_i)\big|\big(L_n^i-\pi|_{K_i}\big)(S_{j,l})\big|\le\sum_{l=1}^{m(j)}\mathbb{E}\big|(L_n-\pi)(S_{j,l})\big|+\mathbb{E}\big|L_n(K_i)-\pi(K_i)\big|.$$

With help of (31), setting $Z=(\frac{8C}{1-\lambda}\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\|_{L^r(\pi)})^{1/2}$ for convenience, we bound the above by

$$\begin{aligned}\sum_{l=1}^{m(j)}\mathbb{E}L_n(K_i)\big|\big(L_n^i-\pi|_{K_i}\big)(S_{j,l})\big|&\le\frac{Z}{\sqrt{n}}\left(\sum_{l=1}^{m(j)}\pi(S_{j,l})^{1/2-1/2r}+\pi(K_i)^{1/2-1/2r}\right)\\&\le\frac{Z}{\sqrt{n}}\pi(K_i)^{1/2-1/2r}\big(m(j)^{1/2+1/2r}+1\big).\end{aligned}$$

With this bound and arguments of integral approximation that are by now usual, we obtain

$$\mathbb{E}\big(\lambda_i^{1/p} W_p\big(L_n^i, \pi|_{K_i}\big)\big) \le C_1 \times \big(\pi(K_i)^{1/p-1/\alpha} 2^i\big) \bigg(\frac{C}{(1-\lambda)n} \bigg\| \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \bigg\|_{L^r(\pi)}\bigg)^{1/[\alpha(1+1/r)]},$$

where $C_1 > 0$ is some constant depending on E, α, p, r. The end goes as in the proof of Corollary 1.3: for $\zeta > 1$, we get

$$\sum_{i\ge 1} \mathbb{E}\big(\lambda_i^{1/p} W_p\big(L_n^i, \pi|_{K_i}\big)\big) \le C_2(\mu, p, r, \zeta, \alpha) \bigg(\frac{C}{(1-\lambda)n} \bigg\| \frac{\mathrm{d}\nu}{\mathrm{d}\pi} \bigg\|_{L^r(\pi)}\bigg)^{1/[\alpha(1+1/r)]}, \tag{33}$$

where

$$C_2(\mu, p, r, \zeta, \alpha) = C_1 \bigg(1 + M_{\zeta\alpha p/(\alpha-p)}^{(\alpha-p)/\alpha p} \frac{2^{1-\zeta}}{1-2^{1-\zeta}}\bigg)$$

(and C_1 depends on E, α, p, r). This completes the proof. □

Appendix A: Some results from measure concentration

In this appendix, we provide results for the deviation of $W_p(L_n, \mu)$ from its mean. We consider only the independent case here. Together with our main results, they give quantitative bounds for the convergence in probability of the empirical measure.

It is an easy observation that when E^n is endowed with the l_p metric

$$d_p\big((x_1, \dots, x_n), (y_1, \dots, y_n)\big) = \big(d(x_1, y_1)^p + \cdots + d(x_n, y_n)^p\big)^{1/p}, \tag{34}$$

the application $(x_1, \dots, x_n) \to L_n^x = 1/n \sum_{i=1}^n \delta_{x_i}$ is Lipschitz with constant $n^{-1/p}$, when the arrival space $\mathbb{P}_p(E)$ is endowed with the metric W_p. Therefore, it is natural to look for concentration inequalities for Lipschitz functions on the space E^n endowed with the product measure $\mu^{\otimes n}$, under which L_n^x is the empirical measure associated with μ. One suitable choice is to look for transportation inequalities.

Transportation inequalities or transportation-entropy inequalities were introduced by K. Marton [28] in order to study the phenomenon of concentration of measure. M. Talagrand showed that the finite-dimensional Gaussian measures satisfy a $\mathbf{T}_2$ inequality. Appendix B contains a simple extension of this result to the infinite-dimensional case. For much more on the topic of transportation inequalities, the reader may refer to the survey [15] by N. Gozlan and C. Léonard.

For $\mu \in \mathcal{P}(E)$, let $H(\cdot|\mu)$ denote the relative entropy with respect to μ:

$$H(\nu|\mu) = \int_E \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \,\mathrm{d}\mu$$

if $\nu \ll \mu$, and $H(\nu|\mu) = +\infty$ otherwise.

We say that $\mu \in \mathcal{P}_p(E)$ satisfies a $\mathbf{T}_p(C)$ transportation inequality when

$$W_p(\nu, \mu) \le \sqrt{C H(\nu|\mu)} \quad \forall \nu \in \mathcal{P}_p(E).$$

Let L_n denote the empirical measure associated with $\mu \in \mathcal{P}(E)$. The next result states that a $\mathbf{T}_p$ inequality on μ implies a Gaussian concentration inequality for $W_p(L_n, \mu)$. We reproduce a particular case of more general results of N. Gozlan and C. Léonard [14,15].

Theorem A.1 ([14], Theorem 12). *Let $\mu \in \mathcal{P}(E)$ satisfy a $\mathbf{T}_p(C)$ inequality. The following holds*:

$$\mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t\big) \leq \mathrm{e}^{-nt^2/C}. \tag{35}$$

Likewise, if μ satisfies the modified inequality $W_p(\mu,\nu) \leq (CH(\nu|\mu))^{1/2p}$, the following holds:

$$\mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t\big) \leq \mathrm{e}^{-nt^{2p}/C}. \tag{36}$$

Remark. *Actually, it is not difficult to check that in Theorem A.1 we can actually replace $W_p(L_n,\mu)$ with $W_p(L_n,\nu)$ for any $\nu \in \mathcal{P}(E)$ (the only important point being that $x \mapsto W_p(L_n^x,\nu)$ is always Lipschitz).*

The bound (35) is used for Gaussian measures in Appendix B. As for the bound (36), its interest is made clear by the following result of F. Bolley and C. Villani ([6], Particular case 2.5): whenever $\mu \in \mathcal{P}_p(E)$ has support with finite diameter D, it satisfies

$$W_p(\nu,\mu) \leq \big(2D^{2p}H(\nu|\mu)\big)^{1/2p} \quad \forall \nu \in \mathcal{P}_p(E). \tag{37}$$

With this in hand, we are in a position to give deviation bounds for measures satisfying only some boundedness or moment condition.

Proposition A.2. *If $\mu \in \mathcal{P}_p(E)$ satisfies $D = \operatorname{Diam}\operatorname{Supp}\mu < +\infty$, we have*

$$\mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t\big) \leq \mathrm{e}^{-nt^{2p}/(2D^{2p})}.$$

Likewise, if μ has finite moment of order β, i.e. $M_\beta = \int d(x_0,x)^\beta \,\mathrm{d}\mu < +\infty$, and $\beta > 2p$, we have

$$\mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t\big) \leq C_\beta n^{1-\beta/2p} t^{-\beta}\big(1 + \big(\log n^{\beta/2p-1} t^\beta\big)^{\beta/2p}\big),$$

where the constant C_β is bounded by $2^{\beta(1+1/2p)} M_\beta$.

Remark. *In contrast with the case of transportation inequalities, we get a polynomial speed of convergence under a polynomial moment assumption. This is not too surprising if one ponders the fact that a transportation inequality implies the finiteness of a square-exponential moment (see [10]), and more generally that any convex transportation-entropy inequality, as defined in [14], requires at least the finiteness of an exponential moment.*

Proof. The majorization in the bounded case is a straightforward consequence of (36) and (37).

In the unbounded case, we use a conditioning argument. Let X_i denote i.i.d. variables of law μ. Let us call $M = \max_{1\leq i\leq n} d(x_0,X_i)$ with x_0 some fixed point, and $L_n = \sum_{i=1}^n \delta_{X_i}$. Let $R > 0$ and denote by B the ball $B(x_0,R)$: conditionally to $M \leq R$, L_n is the empirical measure associated with the measure $\mu^B = \mu \mathbf{1}_B/\mu(B)$.

We have

$$\begin{aligned}\mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t\big) &\leq \mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t | M \leq R\big) \\ &\quad + \mathbb{P}(M \geq R).\end{aligned}$$

Thanks to the first result, we know that

$$\mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t | M \leq R\big) \leq \mathrm{e}^{-nt^{2p}/(2(2R)^{2p})}.$$

Observe that we used our first result with $W_p(L_n,\mu)$ instead of $W_p(L_n,\mu^B)$, but it is still valid in this case for the same reasons as in the remark following Theorem A.1.

On the other hand,

$$\begin{aligned}\mathbb{P}(M \geq R) &= 1 - \mathbb{P}(X_1 \leq R)n = 1 - \big(1 - \mathbb{P}(X_1 \geq R)\big)^n \\ &\leq n\mathbb{P}(X_1 \geq R) \\ &\leq nM_\beta / R^\beta.\end{aligned}$$

Altogether,

$$\mathbb{P}\big(W_p(L_n,\mu) \geq \mathbb{E}\big[W_p(L_n,\mu)\big] + t\big) \leq \mathrm{e}^{-nt^{2p}/(2^{2p+1}R^{2p})} + nM_\beta/R^\beta.$$

Set $y = nt^{2p}/(2^{2p+1}R^{2p})$: the right-hand side is equal to

$$\mathrm{e}^{-y} + n^{1-\beta/2p}t^{-\beta}M_\beta 2^{\beta(2p+1)/2p}y^{\beta/2p}.$$

We pick a value for y by setting $y = -\log n^{1-\beta/2p}t^{-\beta}$, which is positive as soon as $n \geq t^{-2p\beta/(\beta-2p)}$. We get the announced result. □

Remark. *At least in the case $p = 1$, the result of Proposition A.2 in the bounded support case can be recovered in an alternate fashion, using Azuma's inequality (also known as the method of bounded martingale differences). To do so, one should note that the function*

$$(x_1,\dots,x_n) \to W_1\big(L_n^x,\mu\big)$$

has increments bounded by D/n in all its variables.

We do not go any further in the discussion of this topic. However, it is clear that there exist many more functional inequalities yielding concentration-of-measure estimates, such as Poincaré and log-Sobolev inequalities and their weak or weighted forms, and the behaviour under tensorization of these inequalities, which is crucial in the argument above, is generally well understood. References may be found e.g. in the book [23].

Remark. *We have left aside the case of dependent samples, which requires results on dependent tensorization of concentration inequalities. Results in this case are not as numerous as in the independent framework. The reader may refer to Theorem* 2.5 *in* [10] *as well as to* [4,20] *in the W_1 case.*

Appendix B: Transportation inequalities for Gaussian measures on a Banach space

We identify what kind of transport inequality is satisfied by a Gaussian measure on a Banach space. We remind the reader of the following definition: let (E,μ) be a Gaussian–Banach space and $X \sim \mu$ be an E-valued r.v. The weak variance of μ or X is defined by

$$\sigma^2 = \sup_{f\in E^*, |f|\leq 1} \mathbb{E}\big(f^2(X)\big).$$

The lemma below is optimal, as shown by the finite-dimensional case.

Lemma B.1. *Let (E,μ) be a Gaussian–Banach space, and let σ^2 denote the weak variance of μ. Then μ satisfies a $\mathbf{T}_2(2\sigma^2)$ inequality.*

Proof. According e.g. to [24], there exists a sequence $(x_i)_{i\geq 1}$ in E and an orthogaussian sequence $(g_i)_{i\geq 1}$ (meaning a sequence of i.i.d. standard normal variables) such that

$$\sum_{i\geq 1} g_i x_i \sim \mu,$$

where convergence of the series holds a.s. and in all the L^p's. In particular, the laws μ_n of the partial sums $\sum_{i=1}^n g_i x_i$ converge weakly to μ.

As a consequence of the stability result of Djellout, Guillin and Wu (Lemma 2.2 in [10]) showing that $\mathbf{T}_2$ is stable under weak convergence, it thus suffices to show that the measures μ_n all satisfy the $\mathbf{T}_2(2\sigma^2)$ inequality.

First, by definition of σ, we have

$$\sigma^2 = \sup_{f \in E^*, |f| \leq 1} \mathbb{E}\left(\sum_{i=1}^{+\infty} f(x_i) g_i\right)^2$$

and since (g_i) is an orthogaussian sequence, the sum is equal to $\sum_{i=1}^{+\infty} f^2(x_i)$.

Consider the mapping

$$T : (\mathbb{R}^n, N) \to (E, \|\cdot\|)$$
$$(a_1, \dots, a_n) \mapsto \sum_{i=1}^n a_i x_i.$$

(here $\mathbb{R}^n$ is equipped with the Euclidean norm N). With the remark above it is easy to check that $\|T(a)\| \leq \sigma N(a)$ for $a \in \mathbb{R}^n$. Consequently, T is σ-Lipschitz, and we can use the second stability result of Djellout, Guillin and Wu (Lemma 2.1 in [10]): the push forward of a measure satisfying $\mathbf{T}_2(C)$ by a L-Lipschitz function satisfies $\mathbf{T}_2(L^2C)$. As is well-known, the standard Gaussian measure γ^n on $\mathbb{R}^n$ satisfies $\mathbf{T}_2(2)$ and thus $T_\# \gamma^n$ satisfies $\mathbf{T}_2(2\sigma^2)$. But it is readily checked that $T_\# \gamma^n = \mu_n$, which concludes this proof. □

Remark. *M. Ledoux indicated to us another way to obtain this result. First, one shows that the Gaussian measure satisfies a $\mathbf{T}_2(2)$ inequality when considering the cost function $c = d_H^2$, where d_H denotes the Cameron–Martin metric on E inherited from the scalar product on the Cameron–Martin space. This can be done in a number of ways, for example by tensorization of the finite-dimensional $\mathbf{T}_2$ inequality for Gaussian measures or by adapting the Hamilton–Jacobi arguments of Bobkov, Gentil and Ledoux [3] in the infinite-dimensional setting. It then suffices to observe that this transport inequality implies the one we are looking for since we have the bound $d \leq \sigma d_H$ (here d denotes the metric inherited from the norm of the Banach space).*

Acknowledgments

We thank the anonymous referees for their careful reading and many suggestions, that led to a major overhaul of the presentation and improvements on several results of this paper. We thank Patrick Cattiaux and Philippe Berthet for their advice and careful reading of preliminary versions, and Charles Bordenave for introducing us to his work [2] and connected works.

References

[1] M. Ajtai, J. Komlos and G. Tusnády. On optimal matchings. *Combinatorica* **4** (1984) 259–264. MR0779885
[2] F. Barthe and C. Bordenave. Combinatorial optimization over two random point sets. Preprint, 2011. Available at arXiv:1103.2734v1.
[3] S. G. Bobkov, I. Gentil and M. Ledoux. Hypercontractivity of Hamilton-Jacobi equations. *J. Math. Pures Appl.* **80** (2001) 669–696. MR1846020
[4] E. Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-Wasserstein distance. *Electron J. Probab* **16** (2011) 2296–2333. MR2861675
[5] F. Bolley, A. Guillin and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Related Fields* **137** (2007) 541–593. MR2280433
[6] F. Bolley and C. Villani. Weighted Csiszár–Kullback–Pinsker inequalities and applications to transportation inequalities. *Ann. Fac. Sci. Toulouse Math.* **14** (2005) 331–351. MR2172583
[7] P. Cattiaux, D. Chafai and A. Guillin. Central limit theorems for additive functionals of ergodic Markov diffusion processes. Preprint, 2011. Available at arXiv:1104.2198.

[8] E. Del Barrio, E. Giné and C. Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.* **27** (1999) 1009–1071. MR1698999

[9] S. Dereich, F. Fehringer, A. Matoussi and M. Scheutzow. On the link between small ball probabilities and the quantization problem for Gaussian measures on Banach spaces. *J. Theoret. Probab.* **16** (2003) 249–265. MR1956830

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

[11] V. Dobric and J. E. Yukich. Exact asymptotics for transportation cost in high dimensions. *J. Theoret. Probab.* **8** (1995) 97–118. MR1308672

[12] R. M. Dudley. The speed of mean Glivenko–Cantelli convergence. *Ann. Math. Statist.* **40** (1969) 40–50. MR0236977

[13] F. Fehringer. *Kodierung von Gaußmaßen*. Ph.D. manuscript, 2001, available at http://d-nb.info/962880116.

[14] N. Gozlan and C. Léonard. A large deviation approach to some transportation cost inequalities. *Probab. Theory Related Fields* **139** (2007) 235–283. MR2322697

[15] N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Process. Related Fields* **16** (2010) 635–736. MR2895086

[16] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. *Lecture Notes in Mathematics* **1730**. Springer, Berlin, 2000. MR1764176

[17] S. Graf and H. Luschgy. Rates of convergence for the empirical quantization error. *Ann. Probab.* **30** (2002) 874–897. MR1905859

[18] S. Graf, H. Luschgy and G. Pagès. Functional quantization and small ball probabilities for Gaussian processes. *J. Theoret. Probab.* **16** (2003) 1047–1062. MR2033197

[19] J. Horowitz and R. L. Karandikar. Mean rates of convergence of empirical measures in the Wasserstein metric. *J. Comput. Appl. Math.* **55** (1994) 261–273. MR1329874

[20] A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38** (2010) 2418–2442. MR2683634

[21] J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116** (1993) 133–157. MR1237989

[22] M. Ledoux. Isoperimetry and Gaussian analysis. In *Lectures on Probability Theory and Statistics* (Saint-Flour, 1994) 165–294. *Lecture Notes in Math.* **1648**. Springer, Berlin, 1996. MR1600888

[23] M. Ledoux. *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. *Am. Math. Soc.*, Providence, RI, 2001. MR1849347

[24] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]* **23**. Springer, Berlin, 1991. MR1102015

[25] W. V. Li and W. Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.* **27** (1999) 1556–1578. MR1733160

[26] H. Luschgy and G. Pagès. Sharp asymptotics of the functional quantization problem for Gaussian processes. *Ann. Probab.* **32** (2004) 1574–1599. MR2060310

[27] H. Luschgy and G. Pagès. Sharp asymptotics of the Kolmogorov entropy for Gaussian measures. *J. Funct. Anal.* **212** (2004) 89–120. MR2065239

[28] K. Marton. Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration. *Ann. Probab.* **24** (1996) 857–866. MR1404531

[29] M. Talagrand. Matching random samples in many dimensions. *Ann. Appl. Probab.* **2** (1992) 846–856. MR1189420

[30] A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996. MR1385671

[31] V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā* **19** (1958) 23–26. MR0094839

[32] C. Villani. *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften* **338**. Springer, Berlin, 2009. MR2459454