



**HAL**  
open science

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

Emmanuel Boissard, Thibaut Le Gouic

► **To cite this version:**

Emmanuel Boissard, Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 2014, 10.1214/12-AIHP517 . hal-01291298

**HAL Id: hal-01291298**

**<https://hal.science/hal-01291298>**

Submitted on 8 Mar 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# ON THE MEAN SPEED OF CONVERGENCE OF EMPIRICAL AND OCCUPATION MEASURES IN WASSERSTEIN DISTANCE

EMMANUEL BOISSARD AND THIBAUT LE GOUIC

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.

## 1. INTRODUCTION

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

$$(1) \quad L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

denote the empirical measure associated with the i.i.d. sample  $(X_i)_{1 \leq i \leq n}$  of law  $\mu$ , then with probability 1,  $L_n \rightharpoonup \mu$  as  $n \rightarrow +\infty$  (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 [26].

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

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

where the infimum is taken on the set  $\mathcal{P}(\mu, \nu)$  of probability measures with first, resp. second, marginal  $\mu$ , resp.  $\nu$ . 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 [27].

Our purpose is to give bounds on the mean speed of convergence in  $W_p$  distance for the Glivenko-Cantelli theorem, i.e. bounds for the convergence  $\mathbb{E}(W_p(L_n, \mu)) \rightarrow 0$ . 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 [15]. 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 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 [25]. Of particular interest for us is R.M. Dudley's article [11], 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 [17] 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. Tusnady [1] and M. Talagrand [24] 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 [10] 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 [17] and [5]. In the case  $\mu \in \mathcal{P}(\mathbb{R})$ , del Barrio, Giné and Matran [7] obtain a central limit theorem for  $W_1(L_n, \mu)$  under the condition that  $\int_{-\infty}^{+\infty} \sqrt{F(t)(1-F(t))} dt < +\infty$  where  $F$  is the cumulative distribution function (c.d.f.) of  $\mu$ . 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.

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

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

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 the case of Gaussian measures (see Appendix A), but not in general. However, this is a well-trodden topic, and some results are gathered in [4].

*Acknowledgements.* We thank Patrick Cattiaux for his advice and careful reading of preliminary versions, and Charles Bordenave for introducing us to his work [2] and connected works.

### 1.1. Main result and first consequences.

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

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

Our main statement is summed up in the following proposition.

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

$$\mathbb{E}(W_p(L_n, \mu)) \leq c \left( t + n^{-1/2p} \int_t^{d/4} N(X, \delta)^{1/2p} d\delta \right).$$

with  $c \leq 64/3$ .

*Remark.* Proposition 1.1 is related in spirit and proof to the results of R.M. Dudley [11] in the case of the bounded Lipschitz metric

$$d_{BL}(\mu, \nu) = \inf_{f \text{ 1-Lip}, |f| \leq 1} \int f d(\mu - \nu).$$

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  $W_1$  :

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

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 [11] 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. [24] 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

$$(2) \quad N(E, \delta) \leq k_E (\text{Diam } E / \delta)^\alpha.$$

Here, the parameter  $\alpha$  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 :

$$\mathbb{E}[W_p(L_n, \mu)] \leq c \frac{\alpha}{\alpha - 2p} \text{Diam } E k_E^{1/\alpha} n^{-1/\alpha}$$

with  $c \leq 64/3$ .

*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  $d \geq 3$ , 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.

As opposed to Corollary 1.2, 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  $\mu$  is a centered Gaussian random variable with values in  $E$ , meaning that the image of  $\mu$  by every continuous linear functional  $f \in E^*$  is a centered Gaussian variable in  $\mathbb{R}$ . The couple  $(E, \mu)$  is called a (separable) Gaussian Banach space.

Let  $X$  be a  $E$ -valued r.v. with law  $\mu$ , and define the weak variance of  $\mu$  as

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

The small ball function of a Gaussian Banach space  $(E, \mu)$  is the function

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

We can associate to the couple  $(E, \mu)$  their Cameron-Martin Hilbert space  $H \subset E$ , see e.g. [19] 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. Kuelbs-Li [18] and Li-Linde [21], as well as with the approximation of  $\mu$  by measures with finite support in Wasserstein distance (the quantization or optimal quantization problem), see Fehringer's Ph.D. thesis [12], Dereich-Fehringer-Matoussi-Scheutzow [8], Graf-Luschgy-Pagès [16].

We make the following assumptions on the small ball function :

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

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 1.3.** *Let  $(E, \mu)$  be a Gaussian Banach space with weak variance  $\sigma$  and small ball function  $\psi$ . Assume that Assumptions (1) and (2) hold.*

*Then there exists a universal constant  $c$  such that for all*

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

*the following holds :*

$$(3) \quad \mathbb{E}(W_2(L_n, \mu)) \leq c \left[ \psi^{-1}\left(\frac{1}{6 + \kappa} \log n\right) + \sigma n^{-1/[4(6 + \kappa)]} \right].$$

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

$$(4) \quad \mathbb{E}(W_2(L_n, \mu)) \leq C \psi^{-1}(\log n).$$

*Moreover, for  $\lambda > 0$ ,*

$$(5) \quad W_2(L_n, \mu) \leq (C + \lambda)\psi^{-1}(\log n) \text{ with probability } 1 - \exp -n\psi^{-1}(\log n)\frac{\lambda^2}{2\sigma^2}.$$

*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

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

where the infimum runs on the set  $\mathcal{P}_n$  of probability measures with finite support of cardinal bounded by  $n$ . Under some natural assumptions, the upper bound of (5) is matched by a lower bound for the quantization error. Theorem 3.1 in [8] states the following : if for every  $0 < \zeta < 1$ ,

$$\mu((1 - \zeta)\varepsilon B) = o(\mu(\varepsilon B)) \text{ as } \varepsilon \rightarrow 0,$$

then

$$\delta_{n,r} \gtrsim \psi^{-1}(\log n)$$

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

In the terminology of quantization, Theorem 1.3 states that the empirical measure is a rate-optimal quantizer with high probability (under some assumptions on the small ball function). 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 leave aside the question of determining the sharp asymptotics for the average error  $\mathbb{E}(W_2(L_n, \mu))$ , that is of finding  $c$  such that  $\mathbb{E}(W_2(L_n, \mu)) \sim c\psi^{-1}(\log n)$ . Let us underline that the corresponding question for quantizers is tackled for example in [22].

**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  $\pi$ , but a Markov chain with stationary measure  $\pi$  is nevertheless available for simulation. This is the basic setup of the Markov Chain Monte Carlo framework, and a very frequent situation, even in finite dimension.

When looking at the proof of Proposition 1.1, it is apparent that the main ingredient missing in the dependent case is the argument following (18), 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 :

$$(6) \quad \text{Var}_\pi P^n f \leq C\lambda^n \text{Var}_\pi f.$$

In the reversible case, a bound of the type of (6) 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

$$(7) \quad \text{Var}_\pi f \leq C_P \int f(I - P^2)f d\pi$$

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

If (7) holds, we have

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

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

More generally, one may assume that we have a control of the decay of the variance in the following form :

$$(8) \quad \text{Var}_\pi P^n f \leq C \lambda^n \|f - \int f d\pi\|_{L^p}.$$

As soon as  $p > 2$ , these inequalities are weaker than (6). 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 [6].

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  $X \subset E$  with finite diameter,

$$(9) \quad N(X, \delta) \leq k_E (\text{Diam } X / \delta)^\alpha.$$

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

**Theorem 1.4.** *Assume that  $E$  has finite diameter  $d > 0$  and (9) holds. Let  $\pi \in \mathcal{P}(E)$ , and let  $(X_i)_{i \geq 0}$  be a  $E$ -valued Markov chain with initial law  $\nu$  such that  $\pi$  is its unique invariant probability. Assume also that (6) 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 :*

$$\mathbb{E}_\nu [W_p(L_n, \pi)] \leq c \frac{\alpha(1 + 1/r)}{\alpha(1 + 1/r) - 2p} k_E^{1/\alpha} d \left( \frac{C \|\frac{d\nu}{d\pi}\|_r}{(1 - \lambda)n} \right)^{1/[\alpha(1 + 1/r)]}$$

for some universal constant  $c \leq 64/3$ .

The previous theorem has the drawback of assuming that the state space has finite diameter. This can be circumvented, for example by truncation arguments. Our next theorem is an extension to the unbounded case under some moment conditions on  $\pi$ . The statement and the proof involve more technicalities than Theorem 1.4, so we separate the two in spite of the obvious similarities.

**Theorem 1.5.** *Assume that (9) holds. Let  $\pi \in \mathcal{P}(E)$ , and let  $(X_i)_{i \geq 0}$  be a  $E$ -valued Markov chain with initial law  $\nu$  such that  $\pi$  is its unique invariant probability. Assume also that (6) holds for some  $C > 0$  and  $\lambda < 1$ . Let  $x_0 \in E$  and for all  $\theta \geq 1$ , denote  $M_\theta = \int d(x_0, x)^\theta d\pi$ . Fix  $r$  and  $\zeta > 1$  and assume  $2p > \alpha(1 + 1/r)(1 + 1/\zeta)$ .*

*There exist two numerical constant  $C_1(p, r, \zeta)$  and  $C_2(p, r, \zeta)$  only depending on  $p, r$  and  $\zeta$  such that whenever*

$$\frac{C \|\frac{d\nu}{d\pi}\|_r}{(1-\lambda)n} \leq C_1(p, r, \zeta),$$

*the following holds :*

$$\mathbb{E}_\nu [W_p(L_n, \pi)] \leq C_1(p, r, \zeta) K(\zeta) \left( \frac{C \|\frac{d\nu}{d\pi}\|_r}{(1-\lambda)n} \right)^{1/[\alpha(1+1/r)(1+1/\zeta)]}$$

*where*

$$K(\zeta) = \frac{m_\zeta}{m_p^{\zeta/p}} \vee \frac{m_{\zeta+2p}}{m_p^{1+\zeta/p}} \vee k_E^{1/2p(1+1/r)} \frac{2p}{\alpha(1+1/r)} m_p^{\alpha/(2p^2)(1+1/r)}.$$

## 2. PROOFS IN THE INDEPENDENT CASE

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

$$(10) \quad m(j) \leq N(X, 4^{-j}s)$$

*and non-empty subsets  $X_{j,l}$  of  $X$ ,  $u \leq j \leq v$ ,  $1 \leq l \leq m(j)$ , such that the sets  $X_{j,l}$   $1 \leq l \leq m(j)$  satisfy*

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

*In other words, the sets  $X_{j,l}$  form a sequence of partitions of  $X$  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  $B_{j,l} = B(x_{j,l}, 4^{-j}s)$  with  $u \leq j \leq v$  and  $1 \leq l \leq N(X, 4^{-j}s)$ , such that for all  $j$ ,

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

Define  $X_{v,1} = B_{v,1}$ , and successively set  $X_{v,l} = B_{v,l} \setminus X_{v,l-1}$ . 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  $X_{j,l}$  have been built for  $k+1 \leq j \leq v$ . Set  $X_{k,1}$  to be the reunion of all  $X_{k+1,l'}$  such that  $X_{k+1,l'} \cap B_{k,1} \neq \emptyset$ . Likewise, define by induction on  $l$  the set  $X_{k,l}$  as the reunion of all  $X_{k+1,l'}$  such that  $X_{k+1,l'} \cap B_{k,l} \neq \emptyset$  and  $X_{k+1,l'} \not\subset X_{k,p}$  for  $1 \leq p < l$ . 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  $y \in X_{k+1,l'} \subset X_{k,l}$ . We have

$$d(x_{k,l}, y) \leq 4^{-k}s + \text{Diam } X_{k+1,l'} \leq 2 \times 4^{-k}s,$$

thus  $\text{Diam } X_{k,l} \leq 4^{-k+1}s$  as desired.  $\square$

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

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 formed by the points  $x_{j,l}$ ,  $1 \leq j \leq k$ , as a *metric tree*, where points are connected to their ancestor at the previous level, and 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  $L_j$ . Define the measures  $\tilde{m}_{j-1}$  and  $\tilde{n}_{j-1}$  on  $L_{j-1}$  by setting*

$$(11) \quad \tilde{m}_{j-1}(x_{j-1,l'}) = \sum_{(j-1,l') \rightarrow (j,l)} (m_j(x_{j,l}) - n_j(x_{j,l})) \wedge 0,$$

$$(12) \quad \tilde{n}_{j-1}(x_{j-1,l'}) = \sum_{(j-1,l') \rightarrow (j,l)} (n_j(x_{j,l}) - m_j(x_{j,l})) \wedge 0.$$

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

$$(13) \quad W_p(m_j, n_j) \leq 2 \times 4^{-j+2}d \|m_j - n_j\|_{TV}^{1/p} + W_p(\tilde{m}_{j-1}, \tilde{n}_{j-1}).$$

*Proof.* Set  $m_j \wedge n_j(x_{j,l}) = m_j(x_{j,l}) \wedge n_j(x_{j,l})$ . By the triangle inequality,

$$\begin{aligned} W_p(m, n) &\leq W_p(m_j, m_j \wedge n_j) + \tilde{m}_{j-1} + W_p(m_j \wedge n_j + \tilde{m}_{j-1}, m_j \wedge n_j + \tilde{n}_{j-1}) \\ &\quad + W_p(m_j \wedge n_j + \tilde{n}_{j-1}, n_j). \end{aligned}$$

We bound the term on the left. Introduce the transport plan  $\pi_m$  defined by

$$\begin{aligned} \pi_m(x_{j,l}, x_{j,l}) &= m_j \wedge n_j(x_{j,l}), \\ \pi_m(x_{j,l}, x_{j-1,l'}) &= (m_j(x_{j,l}) - n_j(x_{j,l}))_+ \text{ when } (j-1, l') \rightarrow (j, l). \end{aligned}$$

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

$$\begin{aligned} W_p(m_j, m_{\tilde{j}-1}) &\leq \left( \int d^p(x, y) \pi_m(dx, dy) \right)^{1/p} \\ &\leq 4^{-j+2} d \left( \sum_{l=1}^{m(j)} (m_j(x_{j,l}) - n_j(x_{j,l}))_+ \right)^{1/p}. \end{aligned}$$

Likewise,

$$W_p(n_j, m_j \wedge n_j + n_{\tilde{j}-1}) \leq 4^{-j+2} d \left( \sum_{l=1}^{m(j)} (n_j(x_{j,l}) - m_j(x_{j,l}))_+ \right)^{1/p}.$$

As for the term in the middle, it is bounded by  $W_p(\tilde{m}_{j-1}, \tilde{n}_{j-1})$ . Putting this together and using the inequality  $x + y \leq 2^{1-1/p}(x^p + y^p)^{1/p}$ , we get

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

□

**Lemma 2.3.** *Let  $m_j, n_j$  be measures with support in  $L_j$ . Define for  $1 \leq j' < j$  the measures  $m'_{j'}, n'_{j'}$  with support in  $L'_{j'}$  by*

$$(14) \quad m_{j'}(x_{j',l'}) = \sum_{(j',l') \rightarrow (j,l)} m_j(x_{j,l}), \quad n_{j'}(x_{j',l'}) = \sum_{(j',l') \rightarrow (j,l)} n_j(x_{j,l}).$$

The following bound holds :

$$(15) \quad W_p(m_j, n_j) \leq \sum_{j'=1}^j 2 \times 4^{-j'+2} d \|m'_{j'} - n'_{j'}\|_{TV}^{1/p}$$

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

Suppose that (15) holds for measures with support in  $L_{j-1}$ . By lemma 2.2, we have

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

where  $\tilde{m}_{j-1}$  and  $\tilde{n}_{j-1}$  are defined by (11) and (12) respectively. For  $1 \leq i < j-1$ , define following (14)

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

We have

$$W_p(m_j, n_j) \leq 2 \times 4^{-j+2} d \|m_j - n_j\|_{TV}^{1/p} + \sum_{j'=1}^{j-1} 2 \times 4^{-j'+2} d \|\tilde{m}_i - \tilde{n}_i\|_{TV}^{1/p}.$$

To conclude, it suffices to check that for  $1 \leq i \leq j-1$ ,  $\|\tilde{m}_i - \tilde{n}_i\|_{TV} = \|m_i - n_i\|_{TV}$ .  $\square$

*Proof of Proposition 1.1.* We pick some positive integer  $k$  whose value will be determined at a later point. Introduce the sequence of partitions  $(X_{j,l})_{1 \leq l \leq m(j)}$  for  $0 \leq j \leq k$  as in the lemmas above, as well as the points  $x_{j,l}$ . Define  $\mu_k$  as the measure with support in  $L(k)$  such that  $\mu_k(x_{k,l}) = \mu(X_{k,l})$  for  $1 \leq l \leq m(k)$ . The diameter of the sets  $X_{k,l}$  is bounded by  $4^{-k+1}d$ , therefore  $W_p(\mu, \mu_k) \leq 4^{-k+1}d$ .

Let  $L_n^k$  denote the empirical measure associated to  $\mu_k$ .

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

$$(16) \quad \mu_j(x_{j,l'}) = \sum_{(j,l') \rightarrow (k,l)} \mu_k(x_{k,l})$$

$$(17) \quad L_n^j(x_{j,l'}) = \sum_{(j,l') \rightarrow (k,l)} L_n^k(x_{k,l}).$$

It is simple to check that  $\mu_j(x_{j,l}) = \mu(X_{j,l})$ , and that  $L_n^j$  is the empirical measure associated with  $\mu_j$ . Applying (15), we get

$$(18) \quad W_p(\mu_k, L_n^k) \leq \sum_{j=1}^k 2 \times 4^{-j+2} d \|\mu_j - L_n^j\|_{TV}^{1/p}.$$

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

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

In the last inequality, we use Cauchy-Schwarz's inequality and the fact that  $(X_{j,l})_{1 \leq l \leq m(j)}$  is a partition of  $X$ . Putting this back in (18), we get

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

In the last line, we use a standard sum-integral comparison argument. By the triangle inequality, we have

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

We claim that  $\mathbb{E}(W_p(L_n^k, L_n)) \leq W_p(\mu, \mu_k)$ . 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.  $\mathbb{E}|X_i - X_i^k|^p = W_p^p(\mu, \mu^k)$ . We have the identities in law

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

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

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

and passing to expectation proves our claim.

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

*Proof of Corollary 1.2.* It suffices to use Proposition 1.1 along with (2) and to optimize in  $t$ .  $\square$

### 3. PROOF OF THEOREM 1.3.

*Proof of Theorem 1.3.* We begin by noticing that statement (5) is a simple consequence of statement (4) and the tensorization of  $\mathbf{T}_2$ : we have by Corollary A.2

$$\mathbb{P}(W_2(L_n, \mu) \geq \mathbb{E}(W_2(L_n, \mu)) + t) \leq e^{-nt^2/(2\sigma^2)},$$

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  $\mu$ , and by  $B$  the unit ball of  $E$ . According to the Gaussian isoperimetric inequality (see [19]), for all  $\lambda > 0$  and  $\varepsilon > 0$ ,

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

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

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

$$\mu' = \frac{1}{\mu(X)} \mathbf{1}_X \mu$$

the restriction of  $\mu$  to the enlarged ball.

The diameter of  $X$  is bounded by  $2(\sigma\lambda + \varepsilon)$ . The  $W_2$  distance between  $L_n$  and  $\mu$  is thus bounded as follows :

$$(19) \quad W_2(L_n, \mu) \leq 2W_2(\mu, \mu') + ct + cn^{-1/4} \int_t^{(\sigma\lambda + \varepsilon)/2} N(X, \delta)^{1/4} d\delta$$

Set

$$(20) \quad I_1 = W_2(\mu, \mu')$$

$$(21) \quad I_2 = t$$

$$(22) \quad I_3 = n^{-1/4} \int_t^{(\sigma\lambda + \varepsilon)/2} N(X, \delta)^{1/4} d\delta.$$

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

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

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

Introduce the tail function of the Gaussian distribution

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

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  $\Upsilon(t) \leq e^{-t^2/2}/2$ ,  $t \geq 0$  and its consequence

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

We have

$$\Phi^{-1}(e^{-\psi(t/2)}) = -\Upsilon^{-1}(e^{-\psi(t/2)}) \geq -\sqrt{2\psi(t/2)}$$

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

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

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

$$(23) \quad I_1 \leq \sqrt{2}\sigma e^{-(\lambda - \sqrt{2\psi(t/2)})^2/4}.$$

whenever

$$(24) \quad \psi(t/2) \geq \log 2 \text{ and } \lambda - \sqrt{2\psi(t/2)} \geq 0.$$

*Controlling  $I_3$ .* The term  $I_3$  is bounded by  $1/2n^{-1/4}(\sigma\lambda + t/2)N(X, t)^{1/4}$  (just bound the function inside by its value at  $t$ , which is minimal). Denote  $k = N(\lambda K, t - \varepsilon)$  the covering number of  $\lambda K$  (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  $\lambda K$ . From the triangle inequality we get the inclusion

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

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

We now use the well-known link between  $N(\lambda K, t/2)$  and the small ball function. Lemma 1 in [18] gives the bound

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

so that

$$(25) \quad I_3 \leq \frac{1}{2}(\sigma\lambda + t/2)e^{\frac{\lambda^2}{8} + \frac{\kappa}{4}\psi(t/2) - \frac{1}{4}\log n}.$$

Remark that we have used the doubling condition on  $\psi$ , so that we require

$$(26) \quad t/4 \leq t_0.$$

*Final step.* Set now  $t = 2\psi^{-1}(a \log n)$  and  $\lambda = 2\sqrt{2a \log n}$ , with  $a > 0$  yet undetermined. Using (23) and (25), we see that there exists a universal constant  $c$  such that

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

Choose  $a = 1/(6 + \kappa)$  and assume  $\log n \geq (6 + \kappa)(\log 2 \vee \psi(1) \vee \psi(t_0/2))$ . This guarantees that the technical conditions (24) and (26) are enforced, and that  $\psi^{-1}(a \log n) \leq 1$ . Summing up, we get :

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

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

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

and the proof is finished.  $\square$

#### 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 (6) for some  $C > 0$  and  $\lambda < 1$ . Then the following holds :*

$$(27) \quad \mathbb{E}_\nu(W_p(L_n, \pi)) \leq c \left( t + \left( \frac{C}{(1-\lambda)n} \left\| \frac{d\nu}{d\pi} \right\|_r \right)^{1/2p} \int_t^{d/4} N(X, t)^{1/2p(1+1/r)} dt \right).$$

*Proof.* An application of (15) as in (18) yields

$$(28) \quad \mathbb{E}(W_p(L_n, \pi)) \leq 2 \times 4^{-k+1} d + \sum_{j=1}^k 2 \times 4^{-j+2} d \left( \sum_{l=1}^{m(j)} \mathbb{E}|(L_n - \pi)(X_{j,l})| \right)^{1/p}.$$

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

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

Let  $\tilde{p}, \tilde{q}, r \geq 1$  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  $s$ ,

$$\mathbb{E}_\nu [f_A(X_i) f_A(X_j)] \leq \left\| \frac{d\nu}{d\pi} \right\|_r (\mathbb{E}_\pi |f_A(X_i) f_A(X_j)|^s)^{1/s}.$$

Use the Markov property and the fact that  $f \mapsto Pf$  is a contraction in  $L^s$  to get

$$\mathbb{E}_\nu [f_A(X_i) f_A(X_j)] \leq \left\| \frac{d\nu}{d\pi} \right\|_r \|f_A P^{j-i} f_A\|_s.$$

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

$$(29) \quad \mathbb{E}_\nu [f_A(X_i) f_A(X_j)] \leq \left\| \frac{d\nu}{d\pi} \right\|_r \|P^{j-i} f_A\|_{\tilde{p}} \|f_A\|_{\tilde{q}}.$$

Set  $\tilde{p} = 2$  and note that for  $1 \leq t \leq +\infty$ , we have  $\|f_A\|_t \leq 2\pi(A)^{1/t}$ . Use (6) applied to the centered function  $f_A$  to get

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

and as a consequence,

$$(30) \quad \mathbb{E}|(L_n - \pi)(A)| \leq \frac{1}{\sqrt{n}} \frac{2\sqrt{2C}}{\sqrt{1-\lambda}} \left\| \frac{d\nu}{d\pi} \right\|_r^{1/2} \pi(A)^{1/2-1/2r}.$$

Come back to (28) : we have

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

□

*Proof of Theorem 1.4.* Use (27) and (9) to get

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

where

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

Optimizing in  $t$  finishes the proof.

□

We now move to the proof in the unbounded case.

*Proof of Theorem 1.5.* We remind the reader that the following assumption stands : for  $X \subset E$  with diameter bounded by  $d$ ,

$$(31) \quad N(X, \delta) \leq k_E(d/\delta)^\alpha.$$

In the following lines, we will make use of the elementary inequalities

$$(32) \quad (x+y)^p \leq 2^{p-1}(x^p+y^p) \leq 2^{p-1}(x+y)^p.$$

*Step 1.*

Pick increasing sequence of numbers  $d_i > 0$  to be set later on, and some point  $x_0 \in E$ . Define  $C_1 = B(x_0, d_1)$ , and  $C_i = B(x_0, d_i) \setminus B(x_0, d_{i-1})$  for  $i \geq 2$ .



The idea is as follows : we decompose the state space  $E$  into a union of rings, and deal separately with  $C_1$  on the one hand, using the case of Theorem 1.4 as guideline, and with the union of the  $C_i$ ,  $i \geq 2$  on the other hand, where we use more brutal bounds.

We define partial occupation measures

$$L_n^i = 1/n \sum_{j=1}^n \delta_{X_j} \mathbf{1}_{X_j \in C_i}$$

and their masses  $m_i = L_n^i(E)$ . We have the inequality

$$(33) \quad W_p^p(L_n, \pi) \leq \sum_{i \geq 1} m_i W_p^p(1/m_i L_n^i, \pi).$$

On the other hand,

$$\begin{aligned} W_p(1/m_i L_n^i, \pi) &\leq \left( \int d(x_0, x)^p d\pi \right)^{1/p} + \left( \int d(x_0, x)^p d(1/m_i L_n^i) \right)^{1/p} \\ &\leq M_p^{1/p} + d_i, \end{aligned}$$

so that  $W_p^p(1/m_i L_n^i, \pi) \leq 2^{p-1} (M_p + d_i^p)$  using (32). Also, using (33) and (32) yields

$$W_p(L_n, \pi) \leq m_1^{1/p} W_p(1/m_1 L_n^1, \pi) + 2^{1-1/p} \left( \sum_{i \geq 2} m_i [M_p + d_i^p] \right)^{1/p}.$$

Pass to expectations to get

$$(34) \quad \mathbb{E}[W_p(L_n, \pi)] \leq \mathbb{E} \left[ m_1^{1/p} W_p(1/m_1 L_n^1, \pi) \right] + 2^{1-1/p} \left( \sum_{i \geq 2} \pi(C_i) [M_p + d_i^p] \right)^{1/p}$$

We bound separately the left and right term in the right-hand side of (34), starting with the right one.

*Step 2.*

Choose some  $q > p$  and use Chebyshev's inequality to bound the sum on the right by

$$(35) \quad \sum_{i \geq 2} \frac{M_q}{d_{i-1}^q} [M_p + d_i^p]$$

Take  $d_i = \rho^i M_p^{1/p}$ , (35) becomes

$$\begin{aligned} &M_q M_p^{1-q/p} \rho^q \sum_{i \geq 2} [\rho^{-qi} + \rho^{(p-q)i}] \\ &= M_q M_p^{1-q/p} \left[ \frac{\rho^{-q}}{1 - \rho^{-q}} + \frac{\rho^{2p-q}}{1 - \rho^{p-q}} \right]. \end{aligned}$$

Assume for example that  $\rho \geq 2$  : this implies

$$\sum_{i \geq 2} \pi(C_i) [M_p + d_i^p] \leq 4M_q M_p^{1-q/p} \rho^{2p-q}.$$

For later use, we set  $\zeta = q/p - 2$  and the above yields

$$(36) \quad 2^{1-1/p} \left( \sum_{i \geq 2} \pi(C_i) [M_p + d_i^p] \right)^{1/p} \leq 4M_{(\zeta+2)p}^{1/p} M_p^{-(1+\zeta)/p} \rho^{-\zeta}.$$

*Step 3.*

We now turn our attention to the term on the left in (34).

Once again, we apply (15) to obtain

$$W_p(1/m_1 L_n^1, \pi) \lesssim 4^{-k} d_1 + \sum_{j=1}^k 4^{-j} \left( \sum_{l=1}^{m(j)} |(1/m_1)L_n - \pi|(X_{j,l}) \right)^{1/p}$$

Multiply by  $m_1^{1/p}$  and pass to expectations :

$$\begin{aligned} \mathbb{E} \left[ m_1^{1/p} W_p(m_1^{-1} L_n^1, \pi) \right] &\lesssim \sum_{j=1}^k 4^{-j} \left( \sum_{l=1}^{m(j)} \mathbb{E} |(L_n - m_1 \pi)(X_{j,l})| \right)^{1/p} \\ &\quad + 4^{-k} d_1 \mathbb{E}(m_1^{1/p}). \end{aligned}$$

First, notice that  $0 \leq m_1 \leq 1$  a.s. so that  $\mathbb{E}(m_1^{1/p}) \leq 1$ . Next, write

$$\begin{aligned} \sum_{l=1}^{m(j)} \mathbb{E} |(L_n - m_1 \pi)(X_{j,l})| &\leq \sum_{l=1}^{m(j)} \mathbb{E} (|(L_n - \pi)(X_{j,l})| + |(m_1 \pi - \pi)(X_{j,l})|) \\ &\leq \sum_{l=1}^{m(j)} \mathbb{E} |(L_n - \pi)(X_{j,l})| + \mathbb{E} (|m_1 - 1|) \pi(C_1) \\ &\leq \sum_{l=1}^{m(j)} \mathbb{E} |(L_n - \pi)(X_{j,l})| + \mathbb{E} |L_n(C_1) - 1|. \end{aligned}$$

The first of these two terms is controlled using (30) : we have

$$\sum_{l=1}^{m(j)} \mathbb{E} |(L_n - \pi)(X_{j,l})| \leq \frac{1}{\sqrt{n}} \frac{2\sqrt{2C}}{\sqrt{1-\lambda}} \left\| \frac{d\nu}{d\pi} \right\|_r^{1/2} m(j)^{1/2+1/2r}$$

And on the other hand,

$$\begin{aligned} \mathbb{E} |L_n(C_1) - 1| &\leq \mathbb{E} |(L_n - \pi)(C_1)| + \pi(C_1^c) \\ &\leq \frac{1}{\sqrt{n}} \frac{2\sqrt{2C}}{\sqrt{1-\lambda}} \left\| \frac{d\nu}{d\pi} \right\|_r^{1/2} + \pi(C_1^c). \end{aligned}$$

Here we have used (30) again.

We skip over details here as they are similar to those in previous proofs. Choosing an appropriate value for  $k$  and using the estimates above allows us to recover the following :

(37)

$$\mathbb{E} \left[ m_1^{1/p} W_p(1/m_1 L_n^1, \pi) \right] \lesssim \left( \frac{C}{(1-\lambda)n} \left\| \frac{d\nu}{d\pi} \right\|_r \right)^{1/2p} \int_t^{d_1/4} N(C_1, \delta)^{1/2p(1+1/r)} d\delta + \pi(C_1^c) + t.$$

The term  $\pi(C_1^c)$  is bounded by the Chebyshev inequality :

$$\pi(C_1^c) \leq \int x^\zeta d\pi / d_1^\zeta = \int x^\zeta d\pi \left( \int x^p d\pi \right)^{-\zeta/p} \rho^{-\zeta}.$$

*Step 4.*

Use (36) and (37), along with assumption (31) : this yields

$$\mathbb{E}(W_p(L_n, \pi)) \lesssim K(\zeta) \left( \rho^{-\zeta} + t + A_n \rho^{\alpha/2p(1+1/r)} t^{1-\alpha/2p(1+1/r)} \right)$$

where  $A_n = \left( \frac{C}{(1-\lambda)n} \left\| \frac{d\nu}{d\pi} \right\|_r \right)^{1/2p}$ , and

$$K(\zeta) = \frac{m_\zeta}{m_p^{\zeta/p}} \vee \frac{m_{\zeta+2p}}{m_p^{1+\zeta/p}} \vee k_E^{1/2p(1+1/r)} \frac{2p}{\alpha(1+1/r)} m_p^{\alpha/(2p^2)(1+1/r)}.$$

The remaining step is optimization in  $t$  and  $\rho$ . We obtain the following result : there exists a constant  $C(p, r, \zeta)$  depending only on the values of  $p, r, \zeta$ , such that

$$\mathbb{E}(W_p(L_n, \pi)) \lesssim C(p, r, \zeta) K(\zeta) A_n^{2p/(\alpha(1+1/r)(1+1/\zeta))}.$$

There is a caveat : we have used the condition  $\rho \geq 2$  at some point, and with this restriction the optimization above is valid only when  $A_n \leq C'(p, r, \zeta)$ , where the constant  $C'(p, r, \zeta)$  only depends on the values of  $p, r, \zeta$ . □

#### APPENDIX A. TRANSPORTATION INEQUALITIES FOR GAUSSIAN MEASURES ON A BANACH SPACE

Transportation inequalities, also called transportation-entropy inequalities, have been introduced by K. Marton [23] to study the phenomenon of concentration of measure. M. Talagrand showed that the finite-dimensional Gaussian measures satisfy a  $\mathbf{T}_2$  inequality. The following appendix 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 [14] by N. Gozlan and C. Léonard.

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

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

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

$$W_p(\nu, \mu) \leq \sqrt{CH(\nu|\mu)} \quad \forall \nu \in \mathcal{P}_p(E)$$

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  $(E, \mu)$  be a Gaussian Banach space and  $X \sim \mu$  be a  $E$ -valued r.v.. The weak variance of  $\mu$  or  $X$  is defined by

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

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

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

*Proof.* According e.g. to [20], 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  $\mu_n$  of the partial sums  $\sum_{i=1}^n g_i x_i$  converge weakly to  $\mu$ .

As a consequence of the stability result of Djellout-Guillin-Wu (Lemma 2.2 in [9]) showing that  $\mathbf{T}_2$  is stable under weak convergence, it thus suffices to show that the measures  $\mu_n$  all satisfy the  $\mathbf{T}_2(2\sigma^2)$  inequality.

First, by definition of  $\sigma$ , 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

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

(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  $\sigma$ -Lipschitz, and we can use the second stability result of Djellout-Guillin-Wu (Lemma 2.1 in [9]) : the push forward of a measure satisfying  $\mathbf{T}_2(C)$  by a  $L$ -Lipschitz function satisfies  $\mathbf{T}_2(L^2 C)$ . As is well-known, the standard Gaussian measure  $\gamma^n$  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.  $\square$

*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-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).

Let  $L_n$  denote the empirical measure associated with  $\mu$ . As a consequence of Lemma A.1, we can give an inequality for the concentration of  $W_2(L_n, \mu)$  around its mean, using results from transportation inequalities. This is acutally a simple case of more general results of N. Gozlan and C. Léonard ([13], [14]), we reproduce a proof here for convenience.

**Corollary A.2.** *Let  $\mu$  be as above. The following holds :*

$$\mathbb{P}(W_2(L_n, \mu) \geq \mathbb{E}[W_2(L_n, \mu)] + t) \leq e^{-nt^2/(2\sigma^2)}.$$

*Proof.* The proof relies on the property of dimension-free tensorization of the  $\mathbf{T}_2$  inequality, see [14]. Since  $\mu$  satisfies  $\mathbf{T}_2(2\sigma^2)$ , the product measure  $\mu^{\otimes n}$  on the product space  $E^n$  endowed with the  $l_2$  metric

$$d_2((x_1, \dots, x_n), (y_1, \dots, y_n)) = \sqrt{|x_1 - y_1|^2 + \dots + |x_n - y_n|^2}$$

also satisfies a  $\mathbf{T}_2(2\sigma^2)$  inequality ([14], Corollary 4.4). Therefore, it also satisfies a  $\mathbf{T}_1$  inequality by Jensen's inequality, and this implies that we have the concentration inequality

$$\mu^{\otimes n}(f \geq \int f d\mu^{\otimes n} + t) \leq e^{-t^2/(2\sigma^2)}$$

for all 1-Lipschitz functions  $f : (E^n, d_2) \rightarrow \mathbb{R}$  ([14], Theorem 1.7). For  $x = (x_1, \dots, x_n) \in E^n$ , denote  $L_n^x = 1/n \sum_{i=1}^n \delta_{x_i}$ . To conclude it suffices to notice that  $(x_1, \dots, x_n) \rightarrow W_2(L_n^x, \mu)$  is  $\sqrt{n}$ -Lipschitz from  $(E^n, d_2)$  to  $\mathbb{R}$ .  $\square$

## REFERENCES

- [1] M. Ajtai, J. Komlos, and G. Tusnády. On optimal matchings. *Combinatorica*, 4(4):259–264, 1984.
- [2] F. Barthe and C. Bordenave. Combinatorial optimization over two random point sets, March 2011.
- [3] S.G. Bobkov, I. Gentil, and M. Ledoux. Hypercontractivity of Hamilton-Jacobi equations. *Journal des Mathématiques Pures et Appliquées*, 80(7):669–696, 2001.
- [4] E. Boissard and T. Le Gouic. Exact deviations in 1-wasserstein distance for empirical and occupation measures, March 2011.
- [5] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593, 2007.
- [6] P. Cattiaux, D. Chafai, and A. Guillin. Central limit theorems for additive functionals of ergodic Markov diffusions processes. *Arxiv preprint arXiv:1104.2198*, 2011.
- [7] E. Del Barrio, E. Giné, and C. Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, 27(2):1009–1071, 1999.
- [8] S. Dereich, F. Fehring, A. Matoussi, and M. Scheutzow. On the link between small ball probabilities and the quantization problem for Gaussian measures on Banach spaces. *Journal of Theoretical Probability*, 16(1):249–265, 2003.
- [9] H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities for random dynamical systems and diffusions. *Annals of Probability*, 32:2702–2732, 2004.
- [10] V. Dobric and J.E. Yukich. Exact asymptotics for transportation cost in high dimensions. *J. Theoretical Prob*, pages 97–118, 1995.

- [11] R.M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.
- [12] F. Fehring. *Kodierung von Gaußmaßen*. 2001.
- [13] N. Gozlan and C. Léonard. A large deviation approach to some transportation cost inequalities. *Probability Theory and Related Fields*, 139:235–283, 2007.
- [14] N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Processes and Related Fields 16 (2010) 635-736*, 2010.
- [15] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2000.
- [16] S. Graf, H. Luschgy, and G. Pagès. Functional quantization and small ball probabilities for Gaussian processes. *Journal of Theoretical Probability*, 16(4):1047–1062, 2003.
- [17] J. Horowitz and R.L. Karandikar. Mean rates of convergence of empirical measures in the Wasserstein metric. *Journal of Computational and Applied Mathematics*, 55(3):261–273, 1994.
- [18] J. Kuelbs and W.V. Li. Metric entropy and the small ball problem for Gaussian measures. *Journal of Functional Analysis*, 116(1):133–157, 1993.
- [19] M. Ledoux. Isoperimetry and Gaussian analysis. *Lectures on probability theory and statistics*, pages 165–294, 1996.
- [20] M. Ledoux and M. Talagrand. Probability in Banach spaces, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*, 1991.
- [21] W.V. Li and W. Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability*, 27(3):1556–1578, 1999.
- [22] H. Luschgy and G. Pagès. Sharp asymptotics of the functional quantization problem for Gaussian processes. *The Annals of Probability*, 32(2):1574–1599, 2004.
- [23] K. Marton. Bounding  $\bar{d}$ -distance by informational divergence: a method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.
- [24] M. Talagrand. Matching random samples in many dimensions. *The Annals of Applied Probability*, 2(4):846–856, 1992.
- [25] A.W. Van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer Verlag, 1996.
- [26] V.S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics*, 19(1):23–26, 1958.
- [27] C. Villani. *Optimal transport. Old and new*. Grundlehren der Mathematischen Wissenschaften 338. Berlin: Springer. xxii., 2009.