Empirical measures: regularity is a counter-curse to dimensionality

Benoît R. Kloeckner *

February 11, 2018

We propose a "decomposition method" to prove non-asymptotic bound for the convergence of empirical measures in various dual norms. The main point is to show that if one measures convergence in duality with sufficiently regular observables, the convergence is much faster than for, say, merely Lipschitz observables. Actually, assuming s derivatives with $s > d/2$ (d the dimension) ensures an optimal rate of convergence of $1/\sqrt{n}$ (n the number of samples). The method is flexible enough to apply to Markov chains which satisfy a geometric contraction hypothesis, assuming neither stationarity nor reversibility, with the same convergence speed up to a power of logarithm factor.

Our results are stated as controls of the expected distance between the empirical measure and its limit, but we explain briefly how the classical method of bounded difference can be used to deduce concentration estimates.

1 Introduction

1.1 Empirical measures and quadrature

Consider a discrete-time stochastic process $(X_k)_{k\geq 0}$ taking its values in some phase space Ω, assumed to be a Polish space endowed with its Borel σ-algebra. We are concerned with the random atomic measure

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k},$$

called the *empirical measure* of the process, and its convergence. We shall either assume that the $(X_k)_{k\geq 0}$ are independent identically distributed of some law μ, or assume some weak long-range dependence and convergence of the law of X_k to μ as $k \to \infty$.

*Université Paris-Est, Laboratoire d'Analyse et de Matématiques Appliquées (UMR 8050), UPEM, UPEC, CNRS, F-94010, Créteil, France

To quantify the convergence, we are interested in distances on the set $\mathcal{P}(\Omega)$ of probability measures defined by duality. Given a class $\mathscr{F}$ of functions $f : \Omega \to \mathbb{R}$ (sometime called "test functions" or "observables"), one defines for $\nu_0, \nu_1 \in \mathcal{P}(\Omega)$:

$$\|\nu_0 - \nu_1\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \left|\nu_0(f) - \nu_1(f)\right|$$

(note that we write indifferently $\nu_0(f)$ or $\int f \, \mathrm{d}\nu_0$).

One particularly important case is obtained by taking $\mathscr{F} = \mathrm{Lip}_1(\Omega)$, the set of 1-Lipschitz functions. The corresponding metric is the 1-Wasserstein metric $\mathrm{W}_1 = \|\cdot\|_{\mathrm{Lip}_1}$, which by virtue of *Kantorovich duality* can be written equivalently as

$$\mathrm{W}_1(\nu_0, \nu_1) := \inf_{X \sim \nu_0, Y \sim \nu_1} \mathbb{E}\left[\|X - Y\|\right]$$

where $\|\cdot\|$ here is the Euclidean norm and the infimum is over all pairs of random variable with the given measures as individual laws. It is long-known [AKT84] that, when the $(X_k)_{k \geq 0}$ are independent and uniformly distributed on $[0,1]^d$, we have

$$\mathbb{E}\left[\mathrm{W}_1(\hat{\mu}_n, \lambda)\right] \asymp \begin{cases} \dfrac{1}{\sqrt{n}} & \text{if } d = 1, \\ \sqrt{\dfrac{\log n}{n}} & \text{if } d = 2, \\ \dfrac{1}{n^{\frac{1}{d}}} & \text{if } d \geq 3. \end{cases} \tag{1}$$

where $\asymp$ expresses upper and lower bounds up to multiplicative constants and λ denotes the Lebesgue measure. This problem and generalizations have been studied in several works, e.g. [Tal92, Tal94, BLG14, DSS13, FG15, AST16, WB17].

The bounds (1) are interesting theoretically, but are rather negative for the practical application to quadrature. Computations of integrals are in many cases impractical using deterministic methods, and one often has to resort to Monte Carlo methods, i.e. approximate the unknown $\mu(f)$ by $\hat{\mu}_n(f)$. When one has to compute the integrals of a large number of functions $(f_m)_{1 \leq m \leq M}$ with respect to a fixed measure μ, one would rather draw the random quadrature points $X_1, \ldots, X_k$ once and for all, and use them for all functions f_m; while usual Monte Carlo bound will ensure each individual estimate $\hat{\mu}_n(f_m)$ has small probability to be far from $\mu(f_m)$, if M is large compared to n these bounds will not ensure that *all* estimates are good with high probability. On the contrary, convergence in W_1 (or in duality with some other class $\mathscr{F}$) ensures good estimates simultaneously for all f_m, as long as they belong to the given class, independently of M. This makes such convergence potentially useful; but the *rate* given above, $n^{-\frac{1}{d}}$, is hopelessly slow in high dimension which is precisely the setting where Monte Carlo methods are most needed. We shall prove that if the functions of interest are regular, then this "curse of dimensionality" can be overcome. We shall be interested in the duality with $\mathcal{C}^{\mathrm{s}}_1$ the set of functions with $\mathcal{C}^{\mathrm{s}}$ norm at most 1 (precise definitions are given

below; when $s = 1$ this is the set of 1-Lipschitz functions); but other spaces could be considered, e.g. Sobolev or Besov spaces.

Another issue is that in many cases, drawing independent samples $(X_k)_{k\geq 0}$ of law μ is not feasible, and one is lead to instead rely on a Markov chain having μ as its stationary measure; this is the Markov Chain Monte Carlo method (MCMC). While the empirical measure of Markov chains have been considered by Fournier and Guillin [FG15], these authors need quite strong assumptions: a spectral gap in the L^2 space (or similarly large spaces), and a "warm start" hypothesis (X_0 should have a law absolutely continuous with respect to μ). In good cases, one can achieve this by a burn-in period (start with arbitrary X_0, and consider $(X_{k_0+k})_{k\geq 0}$ for some large k_0); but in some cases, each X_k has a singular law with respect to μ (for example the natural random walk generated by an Iterated Function System). We shall consider Markov chains satisfying a certain geometric contraction property, but again the method can certainly be adapted to other assumptions.

1.2 Markov chains

Our main result handles Markov chains of arbitrary starting distribution and with a spectral gap in Lip (e.g. positively curved chains in the sense of Ollivier [Oll09]).

Theorem A. *Assume that $(X_k)_{k\geq 0}$ is a Markov chain defined on a bounded domain Ω of $\mathbb{R}^d$, whose iterated transition kernel $(m_x^t)_{x\in\Omega,t\in\mathbb{N}}$ defined by*

$$m_x^t(A) = \mathbb{P}(X_{k+t} \in A \mid X_k = x)$$

is exponentially contracting in the Wasserstein metric W_1, i.e. there are constants $D \geq 1$ and $\theta \in (0,1)$ such that

$$\mathrm{W}_1(m_x^t, m_y^t) \leq D\theta^t \|x - y\|.$$

Denote by μ the (unique) stationary measure of the transition kernel.

Then for some constant $C = C(\Omega, d, D, s)$ and all large enough n, letting $\bar{n} = (1-\theta)n$, we have

$$\mathbb{E}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s}\right] \leq C \begin{cases} \dfrac{(\log \bar{n})^{\frac{d}{2s+1}}}{\sqrt{\bar{n}}} & \text{when } s > d/2 \\ \dfrac{\log \bar{n}}{\sqrt{\bar{n}}} & \text{when } s = d/2 \\ \dfrac{(\log \bar{n})^{d-2s+\frac{s}{d}}}{\bar{n}^{\frac{s}{d}}} & \text{when } s < d/2 \end{cases} \tag{2}$$

Let us stress two strengths of this result:

- for $s = 1$, recalling $\|\cdot\|_{\mathcal{C}_1^1} = \|\cdot\|_{\mathrm{Lip}_1} = \mathrm{W}_1$, the bounds are only a power of logarithm factor away from the optimal bounds for IID random variables,
- for s large enough, we almost obtain the optimal convergence rate $\asymp 1/\sqrt{n}$

- we assume neither reversibility, stationarity, nor warm start hypotheses (the distribution of X_0 can be arbitrary),
- the rate of convergence does not depend on the specific feature of the Markov chain, only on D and θ.

Note that for fixed θ, $\bar{n}$ has the same order than n, but if θ is close to 1, $1/(1-\theta)$ is the typical time scale for the decay of correlations. One thus cannot expect less than $(1-\theta)n$ Markov samples to achieve the bound obtained for n independent samples.

Examples of Markov chains which are exponentially contracting in W_1 (equivalently, that have a spectral gap in the space of Lipschitz observables) are numerous; it is a slightly more general condition than "positive curvature" in the sense of Ollivier [Oll09], see e.g. [JO10] and [Klo17b] for concrete examples, or in the context of dynamical systems [KLS15] and [Klo17a].

Under the assumption of Theorem A, it is well-known that uniform estimates

$$\sup_{f\in\mathscr{F}} \mathbb{P}\left(|\hat{\mu}_n(f)-\mu(f)|>\varepsilon\right)\to 0 \qquad \text{and} \qquad \sup_{f\in\mathscr{F}} \mathbb{E}\left[|\hat{\mu}_n(f)-\mu(f)|\right]\to 0 \tag{3}$$

hold, here with $\mathscr{F}=\mathrm{Lip}_1$ (or any smaller class), with a Gaussian rate.

The problem of convergence in duality to the class $\mathscr{F}$ is thus to invert the supremum and the probability (or expectancy), to bound from above

$$\mathbb{P}\Big(\sup_{f\in\mathscr{F}}|\hat{\mu}_n(f)-\mu(f)|>\varepsilon\Big) \qquad \text{or} \qquad \mathbb{E}\Big[\sup_{f\in\mathscr{F}}|\hat{\mu}_n(f)-\mu(f)|\Big].$$

We shall disregard the potential issue of non-measurability: as we shall only deal with classes $\mathscr{F}$ having a countable subset which is dense in the uniform norm, we can always replace the supremum with a supremum over a countable set of functions.

The idea of the proof of Theorem A is to take an arbitrary $f\in\mathcal{C}^{\mathrm{s}}_1(\Omega)$ and decompose it using Fourier series. The regularity hypothesis gives us a control on both the uniform approximation by a truncated Fourier series, and on the Fourier coefficients. Combining these controls, we bound from above $|\hat{\mu}_n(f)-\mu(f)|$ by a quantity that does not depend on f at all, but depends on the Fourier basis elements $(e_k)_{k\in\mathbb{Z}^d}$ up to some index size. Taking a supremum and an expectation, this leaves us with the simple task to optimize where to truncate the Fourier series.

This decomposition method can in principle be used under various assumptions on the process $(X_k)_{k\geq 0}$, the point being to identify a decomposition suited to the assumption; in particular, one can easily adapt the method to study geometrically ergodic Markov chains. I chose to present Theorem A in part because its hypothesis is relevant to several Markov chains I am interested in, and in part because it presents specific difficulties: a blunt computation leads to non-optimal powers of n. To obtain good rates, we translate the contraction hypothesis to frame part of the argument in the space Hol_α, where the Fourier basis has smaller norm; and instead of bounding the Fourier coefficients of a Lipschitz function directly, we use Parceval's formula and the injection $\mathcal{C}^{\mathrm{s}}\to H^s$ which turns out to give a better estimate. Another functional decomposition, and another path in computations might improve the power in the logarithmic factor.

We restrict to the compact case, but the method can in principle be adapted, or truncation argument be used, to deal with non-compactly supported measure.

In order to introduce the decomposition method and show its flexibility, we shall state two simpler results below.

1.3 Explicit bounds in the i.i.d case, for the Wasserstein metric

The decomposition method enables one to get a very explicit version of (1) with a few computations but very little sophistication.

Theorem B. *If μ is any probability measure on $[0,1]^d$ and $(X_k)_{k\geq 0}$ are i.i.d. random variable with law μ, then for all $n \in \mathbb{N}$ we have*

$$\mathbb{E}\Big[\,\mathrm{W}_1(\hat{\mu}_n,\mu)\Big] \leq \begin{cases} \dfrac{1}{2(\sqrt{2}-1)}\cdot\dfrac{1}{\sqrt{n}} & \text{when } d=1 \\ \dfrac{\log_2(n)+8}{\sqrt{8n}} & \text{when } d=2 \\ \dfrac{C_d}{n^{\frac{1}{d}}} & \text{when } d\geq 3 \end{cases} \tag{4}$$

where $C_3 \leq 6.3$, $C_d \leq 3\sqrt{d}$ for all $d \geq 4$, and $C_d/\sqrt{d} \to 2$ as $d \to \infty$.

The order of magnitude of these bounds is sharp in many regimes:

- in dimension 1, the order of magnitude $1/\sqrt{n}$ is optimal; however the constant $1/(2(\sqrt{2}-1))$ is *not* asymptotically optimal when μ is Lebesgue measure,
- when $d=2$ and μ is Lebesgue measure, as previously mentioned the correct order is $\sqrt{\log n/n}$, but to the best of my knowledge it is an open question to determine whether this better order holds for arbitrary measures (a positive answer is strongly expected). See Section 2.4 for an example showing that in a more general setting the order $\log n/\sqrt{n}$ cannot be improved,
- when $d \geq 3$, both orders of magnitude $n^{-1/d}$ as $n \to \infty$ and $\sqrt{d}$ as $d \to \infty$ are sharp up to multiplicative constants (see Remark 2.2). The asymptotic constant 2 is certainly quite larger than the asymptotic constant

$$\lim_{d\to\infty}\lim_{n\to\infty}\frac{n^{\frac{1}{d}}}{\sqrt{d}}\,\mathbb{E}\Big[\,\mathrm{W}_1(\hat{\mu}_n,\lambda)\Big]$$

  which has been computed for the related, but slightly different *matching problem* by Talagrand [Tal92]; but our bound holds for all n and all d (and also all μ). An even more general bound has been given by Boissard and Le Gouic [BLG14], but their constant is larger by a factor approximately 10.

Let us stress that the main purpose of this result will be to expose our method in an elementary setting: indeed many previous similar bounds are available in this case. For example more general non-asymptotic results have been obtained by Fournier and Guillin [FG15], building on previous work by Dereich, Scheutzow and Schottstedt [DSS13]. They are more general in that they consider q-Wasserstein metric for any $q > 0$ (while we will only be able to consider $q \leq 1$), and apply to non-compactly supported measures μ under moment assumptions. However their constants, though non-asymptotic, have not been made explicit, and their behavior when the dimension grows has not been studied.

1.4 Regular observables and independent samples

In the i.i.d. case, we can improve Theorem A by removing most of the logarithmic factors.

Theorem C. *If μ is any probability measure on $[0,1]^d$ and $(X_k)_{k\geq 0}$ are i.i.d. random variable with law μ, then for all $s \geq 1$, for some constant $C = C(d,s) > 0$ (not depending upon μ), and all integer $n \geq 2$ we have*

$$\mathbb{E}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s}\right] \leq C \begin{cases} \dfrac{1}{\sqrt{n}} & \text{when } s > \frac{d}{2} \\ \dfrac{\log n}{\sqrt{n}} & \text{when } s = \frac{d}{2} \\ \dfrac{1}{n^{s/d}} & \text{when } s < \frac{d}{2} \end{cases} \tag{5}$$

It is possible to prove this result with previous, more classical methods. Indeed, combining the "entropy bound" for the class $\mathcal{C}_1^s$ [VdVW96, Thm 2.7.1] and the "chaining method" (see e.g. [vH96, Ex 5.11, p. 138]) leads to Theorem C; I am indebted to Jonathan Weed for pointing this out to me. The proof by the decomposition method we provide here is very simple, but non-elementary as it relies on a wavelet decomposition. It is well-known that all functions in $\mathcal{C}_1^s$ can be written as a linear combination of a few elements of a wavelet basis, with small coefficients, up to a small error. Then controlling $|\hat{\mu}_n(f) - \mu(f)|$ for all $f \in \mathcal{C}_1^s$ simultaneously reduces to controlling this quantity for the few needed elements of the wavelet basis.

1.5 concentration inequalities

Up to know, we have restricted to estimates on the expectancy, while in many practical situations one would need concentration estimates. This is in fact not a restriction, as we shall explain briefly in Section 5: the classical bounded difference method enable one to get concentration near the expectancy. In particular, we get the following.

Corollary D. *Under the assumptions of Theorem A, for some ϵ depending on $\theta, D, \operatorname{diam}\Omega$, for all large enough n and all $M \geq C = C(\Omega, d, D, \theta)$ we have:*

- *when* $s > d/2$

$$\mathbb{P}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s} \geq M\frac{(\log n)^{\frac{d}{2s+1}}}{\sqrt{n}}\right] \leq e^{-\epsilon(M-C)^2(\log n)^{\frac{d}{2s+1}}} \tag{6}$$

- *when* $s = d/2$

$$\mathbb{P}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s} \geq M\frac{\log n}{\sqrt{n}}\right] \leq e^{-\epsilon(M-C)^2(\log n)^2} \tag{7}$$

- *when* $s < d/2$

$$\mathbb{P}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s} \geq M\frac{(\log n)^{d-2s+\frac{s}{d}}}{n^{\frac{s}{d}}}\right] \leq e^{-\epsilon(M-C)^2 n^{1-2s/d}}. \tag{8}$$

(The last inequality is not optimal as we relaxed the poly-logarithmic factor for simplicity.)

For example, when $s \geq d/2$ we deduce that $\frac{\sqrt{n}}{\log n}\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s}$ is bounded almost surely.

Structure of the paper Sections 2, 3 and 4 are independent and contain the proofs of the main Theorems (B, C and A respectively: we start with the most elementary proof, follow with the simplest one, and end with the most sophisticated).

Section 5, dealing with concentration estimates, is mostly independent from the previous ones, which are only used to deduce Corollary D.

We shall write $a \lesssim b$ for $a \leq Cb$, the dependency of the constant C being left implicit unless it feels necessary; the constants denoted by C will be allowed to change from line to line.

2 Wasserstein convergence and dyadic decomposition

The goal of this Section is to prove (a refinement of) Theorem B. We consider a sequence $(X_k)_{1\leq k}$ of independent, identically distributed random points whose common law shall be denoted by μ; we assume that μ is supported on the cube $[0,1]^d$ and consider the convergence of the empirical measure $\hat{\mu}_n := \sum_{k=1}^n \frac{1}{n}\delta_{X_k}$ in the q-Wasserstein distance where $q \in (0,1]$, i.e.

$$\mathrm{W}_q(\mu_0,\mu_1) := \inf_{f\in \mathrm{Hol}_1^q} \left|\mu_0(f) - \mu_1(f)\right|$$

where Hol_1^q is the set of functions $f : [0,1]^d \to \mathbb{R}$ such that for all $x, y \in [0,1]^d$:

$$|f(x) - f(y)| \leq \|x-y\|^q$$

While we are mostly interested in the Euclidean norm $\|\cdot\|$, our method is sharper in the case of the supremum norm[1] $\|\cdot\|_\infty$, with respect to which the analogue of the

[1]The same notation is used for the uniform norm of functions, but the type of the argument will prevent any confusion.

aforementioned objects are denoted by $\mathrm{W}_{q,\infty}$ and $\mathrm{Hol}_1^{q,\infty}$. We will work with $\|\cdot\|_\infty$, and then deduce directly the corresponding result for the Euclidean norm by using that $\|\cdot\| \le \sqrt{d}\|\cdot\|_\infty$ (and thus $\mathrm{W}_q \le d^{\frac{q}{2}}\,\mathrm{W}_{q,\infty}$).

Our most precise result is the following.

Theorem 2.1. *For all $q \in (0,1]$ and all n, it holds:*

$$\mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n,\mu)\right] \le \begin{cases} \dfrac{2^{\frac{d}{2}-2q}}{1-2^{\frac{d}{2}-q}}\cdot\dfrac{1}{\sqrt{n}} & \text{when } d<2q, \\ \left(2+\dfrac{\log_2(n)}{2^{q+1}q}\right)\dfrac{1}{\sqrt{n}} & \text{when } d=2q \\ 2\left(\dfrac{\frac{d}{2}-q}{2q(1-2^{q-\frac{d}{2}})}\right)^{\frac{2q}{d}}\left(1+\dfrac{q}{2^q(\frac{d}{2}-q)}\right)\dfrac{1}{n^{\frac{q}{d}}} & \text{when } d>2q. \end{cases}$$

We deduce several more compact formulas below, including Theorem B. Observe that for fixed q and large d, the complicated front constant converges to 2.

Remark 2.2. It is not difficult to see that for μ the Lebesgue measure and an optimal, deterministic approximation $\tilde{\mu}_n$ with $n = k^d$ Dirac masses, one has

$$\mathrm{W}_{1,\infty}(\tilde{\mu}_n,\mu) \ge \frac{d}{(d+q)2^q}\frac{1}{n^{\frac{q}{d}}}$$

so that in high dimension, for the ℓ^∞ norm and in the worst case $q=1$ our estimate is off by a factor of approximately 4 compared to a best approximation.

With the Euclidean norm, an easy lower bound in the case of the Lebesgue measure is obtained by observing that a mass at most

$$\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}R^d n$$

is at distance R or less of one of the n points (be they random or not). This leads, for *any* measure $\tilde{\mu}_n$ supported on n points, to

$$\mathrm{W}_1(\tilde{\mu}_n,\mu) \ge n\int_0^{R_0} d\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}R^d\,\mathrm{d}R = n\frac{d\pi^{\frac{d}{2}}}{(d+1)\Gamma(\frac{d}{2}+1)}R_0^{d+1}$$

where R_0 is defined by $n\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}R_0^d = 1$. Finally,

$$W_1(\tilde{\mu}_n,\mu) \ge \underbrace{\frac{d\Gamma(\frac{d}{2}+1)^{\frac{1}{d}}}{(d+1)\sqrt{\pi}}}_{\underset{d\to\infty}{\sim}\sqrt{\frac{d}{2e\pi}}}\cdot\frac{1}{n^{\frac{1}{d}}}$$

and again our order of magnitude $C_d \asymp \sqrt{d}$ is the correct one.

The results of [Tal92] show that, at least for the bipartite matching problem, this seemingly crude lower bounds are in fact attained asymptotically, taking renormalized limits as $n \to \infty$ and then $d \to \infty$. This indicates that our constant are not optimal, and it would be interesting to have a non-asymptotic bound with optimal asymptotic behavior.

2.1 Decomposition of Hölder functions

The method to prove Theorem 2.1 consists in a multiscale decomposition of the functions $f \in \mathrm{Hol}_1^{q,\infty}$. In its spirit, it seems quite close to arguments of [BLG14], [DSS13] and [FG15]; our interest is mostly in setting this multiscale analysis in a functional decomposition framework.

We fix a positive integer J to be optimized later, representing the depth of the decomposition. For each $j \in \{0, \dots, J\}$, set $\Lambda_j = \{j\} \times \{0, \dots, 2^j - 1\}^d$; then define $\Lambda = \bigcup_{j=0}^{J} \Lambda_j$, acting as the set of indices for the decomposition.

For each $j \in \{0, \dots, J\}$, let $\{C_\lambda : \lambda \in \Lambda_j\}$ be the regular decomposition of $[0,1]^d$ into cubes of side-length 2^{-j}; the boundary points are attributed in an arbitrary (measurable) manner, with the constraint that $\{C_\lambda : \lambda \in \Lambda_j\}$ is a partition of $[0,1]^d$ that refines the previous partition $\{C_\lambda : \lambda \in \Lambda_{j-1}\}$. Denote by x_λ the center of the cube C_λ, and by $\psi_\lambda := \mathbf{1}_{C_\lambda}$ the characteristic function of C_λ (so that for each j, $\sum_{\lambda \in \Lambda_j} \psi_\lambda = \mathbf{1}_{[0,1]^d}$).

Lemma 2.3. *For all function $f \in \mathrm{Hol}_1^{q,\infty}$ and all J, there exists coefficients $\alpha(\lambda) \in \mathbb{R}$ such that*

$$f = \sum_{j=1}^{J} \sum_{\lambda \in \Lambda_j} \alpha(\lambda) \psi_\lambda + c + g \tag{9}$$

where c is a constant and g is a function $[0,1]^d \to \mathbb{R}$, such that

$$|\alpha(\lambda)| \le 2^{-(j+1)q} \qquad \forall \lambda \in \Lambda_j$$
$$\|g\|_\infty \le 2^{-(J+1)q}.$$

Proof. Replacing f with $f - c$ where $c = f(x_{0,0})$, we assume that f vanishes at the center $x_{0,0}$ of $C_{0,0} = [0,1]^d$. Observe that $f \in \mathrm{Hol}_1^{q,\infty}$ then implies that $\|f\|_\infty \le 2^{-q}$ and $|f(x_\lambda)| \le 2^{-2q}$ for all $\lambda \in \Lambda_1$.

For $\lambda \in \Lambda_1$, we define $\alpha(\lambda) = f(x_\lambda)$ and set $f_1 = \sum_{\lambda \in \Lambda_1} \alpha(\lambda)\psi_\lambda$; we have $|\alpha(\lambda)| \le 2^{-2q}$, the function $f - f_1$ is $\mathrm{Hol}_1^{q,\infty}$ on C_λ and vanishes at x_λ. Since C_λ is a $\|\cdot\|_\infty$ ball of center x_λ and radius $1/4$, it follows that $\|f - f_1\|_\infty \le 2^{-2q}$ on each C_λ, and thus on the whole of $[0,1]^d$. Moreover for all $\lambda \in \Lambda_2$ it holds $|(f - f_1)(x_\lambda)| \le 2^{-3q}$.

Similarly, we define $f_j : [0,1]^d \to \mathbb{R}$ recursively by setting $\alpha(\lambda) = (f - f_{j-1})(x_\lambda)$ for all $\lambda \in \Lambda_j$ and $f_j = f_{j-1} + \sum_{\lambda \in \Lambda_j} \alpha(\lambda)\psi_\lambda$. Then $|\alpha(\lambda)| \le 2^{-(j+1)q}$ for all $\lambda \in \Lambda_j$ and $\|f - f_J\|_\infty \le 2^{-(J+1)q}$. □

2.2 Wasserstein distance estimation

With the notation of Lemma 2.3, for any $f \in \mathrm{Hol}_1^q$ we have:

$$\begin{aligned}\left|\hat{\mu}_n(f) - \mu(f)\right| &\leq 2\|g\|_\infty + \sum_{j=1}^{J} \sum_{\lambda \in \Lambda_j} |\alpha(\lambda)||\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)| \\ &\leq 2^{1-(J+1)q} + \sum_{j=1}^{J} 2^{-(j+1)q} \sum_{\lambda \in \Lambda_j} |\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|\end{aligned}$$

where the last right-hand term does not depend on f in any way. We can thus take a supremum and an expectation to obtain

$$\mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n, \mu)\right] \leq 2^{1-(J+1)q} + \sum_{j=1}^{J} 2^{-(j+1)q} \sum_{\lambda \in \Lambda_j} \mathbb{E}\left[|\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|\right]$$

Remark 2.4. This is the core of the decomposition method. Observe that we used no hypothesis on the (X_k) yet; any stochastic process for which one can control $\mathbb{E}[|\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|]$ can be applied the method.

Setting $p_\lambda = \mu(\psi_\lambda)$, the random variable $n\hat{\mu}_n(\psi_\lambda)$ is binomial of parameters n and p_λ. A standard estimation of the mean absolute deviation yields

$$\begin{aligned}\mathbb{E}\left[|n\hat{\mu}_n(\psi_\lambda) - n\mu(\psi_\lambda)|\right] &\leq \sqrt{np_\lambda(1-p_\lambda)} \\ \sum_{\lambda \in \Lambda_j} \mathbb{E}\left[|\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|\right] &\leq \frac{1}{\sqrt{n}} \sum_{\lambda \in \Lambda_j} \sqrt{p_\lambda}\end{aligned}$$

By concavity of the square-root function, we have

$$2^{-dj} \sum_{\lambda \in \Lambda_j} \sqrt{p_\lambda} \leq \sqrt{2^{-dj} \sum_{\lambda \in \Lambda_j} p_\lambda} = 2^{-\frac{dj}{2}} \tag{10}$$

and we deduce

$$\begin{aligned}\sum_{\lambda \in \Lambda_j} \mathbb{E}\left[|\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|\right] &\leq \frac{2^{\frac{dj}{2}}}{\sqrt{n}} \\ \mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n, \mu)\right] &\leq 2^{1-(J+1)q} + \sum_{j=1}^{J} \frac{2^{j(\frac{d}{2}-q)-q}}{\sqrt{n}},\end{aligned} \tag{11}$$

leaving us with the simple task to optimize the choice of J.

2.3 Optimization of the depth parameter

We shall distinguish three cases: $d < 2q$, $d = 2q$ and $d > 2q$. The first case is only possible for $d = 1$, but we let it phrased that way because for some measures μ the dimension d of the ambient space can be replaced by the "dimension" of the measure itself, see Section 2.4 for an example.

2.3.1 Small dimension

If $d < 2q$, then the sum in (11) is bounded independently of J and we can let $J \to \infty$ to obtain:

$$\begin{aligned}\mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n, \mu)\right] &\leq \frac{2^{-q}}{\sqrt{n}} \sum_{j=1}^{\infty} 2^{j(\frac{d}{2}-q)} \\ &\leq \frac{2^{\frac{d}{2}-2q}}{1-2^{\frac{d}{2}-q}} \cdot \frac{1}{\sqrt{n}} \end{aligned} \tag{12}$$

In particular, for $d = 1$, $q = 1$:

$$\mathbb{E}\left[\mathrm{W}_1(\hat{\mu}_n, \mu)\right] \leq \frac{1}{2(\sqrt{2}-1)} \cdot \frac{1}{\sqrt{n}} \tag{13}$$

Remark 2.5. For $\frac{d}{2} - q$ close to 0, the constant in (12) goes to infinity; in this regime, for moderate n letting $J \to \infty$ is sub-optimal and one should optimize J in (11) as we shall do in the next cases.

2.3.2 Critical dimension

If $d = 2q$ (or in fact $d \leq 2q$) we can rewrite (11) as

$$\mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n, \mu)\right] \leq 2^{1-(J+1)q} + \frac{2^{-q} J}{\sqrt{n}}.$$

To optimize J, we formally differentiate the right-hand side with respect to J, equate to zero and solve for J. Reminding that J is an integer, and keeping only the leading term (when $n \to \infty$) to simplify, this leads us to choose

$$J = \left\lfloor \frac{\log_2 n}{2q} \right\rfloor$$

in particular implying $2^{1-(J+1)q} \leq 2/\sqrt{n}$. We deduce the claimed bound

$$\mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n, \mu)\right] \leq \left(2 + \frac{\log_2(n)}{2^{q+1} q}\right) \frac{1}{\sqrt{n}} \lesssim \frac{\log n}{\sqrt{n}} \tag{14}$$

immediately implying the bound of Theorem B for $d = 2$ and $q = 1$ (where a $\sqrt{2}$ comes from the comparison between the supremum and Euclidean norms):

$$\mathbb{E}\left[\mathrm{W}_1(\hat{\mu}_n, \mu)\right] \leq \frac{\log_2(n) + 8}{\sqrt{8n}} \tag{15}$$

2.3.3 Large dimension

If $d > 2q$, equation (11) becomes

$$\mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n,\mu)\right] \leq 2^{1-(J+1)q} + \frac{2^{J(\frac{d}{2}-q)}-1}{1-2^{q-\frac{d}{2}}}\cdot\frac{1}{2^q\sqrt{n}} \leq 2^{1-(J+1)q} + \frac{2^{J(\frac{d}{2}-q)}}{2^q(1-2^{q-\frac{d}{2}})}\cdot\frac{1}{\sqrt{n}}$$

Following the same optimization process as in the critical dimension case, we choose J such that

$$\frac{1}{2}n^{\frac{1}{d}}\left(\frac{2q(1-2^{q-\frac{d}{2}})}{\frac{d}{2}-q}\right)^{\frac{2}{d}} \leq 2^J \leq n^{\frac{1}{d}}\left(\frac{2q(1-2^{q-\frac{d}{2}})}{\frac{d}{2}-q}\right)^{\frac{2}{d}}$$

leading to

$$\mathbb{E}\left[\mathrm{W}_{q,\infty}(\hat{\mu}_n,\mu)\right] \leq 2\left(\frac{\frac{d}{2}-q}{2q(1-2^{q-\frac{d}{2}})}\right)^{\frac{2q}{d}}\left(1+\frac{q}{2^q(\frac{d}{2}-q)}\right)\frac{1}{n^{\frac{q}{d}}}$$

For $q=1$ and $d\geq 3$, it comes $\mathbb{E}\left[\mathrm{W}_{1,\infty}(\hat{\mu}_n,\mu)\right] \leq C'_d n^{-\frac{1}{d}}$ where

$$C'_d = 2\left(\frac{\frac{d}{2}-1}{2-2^{2-\frac{d}{2}}}\right)^{\frac{2}{d}}\left(1+\frac{1}{d-2}\right)\frac{1}{n^{\frac{1}{d}}}$$

We have notably $C'_4 = 3$. Relaxing our bound for $d\geq 4$ to

$$C'_d \leq 2\left(\frac{d}{4}\right)^{\frac{2}{d}}\left(1+\frac{1}{d-2}\right)$$

it is more easily seen that it is decreasing (and still takes the value 3 at $d=4$). We also see that we can take $C'_d \to 2$ as $d\to\infty$. The last part of Theorem B follows with $C_d = \sqrt{d}C'_d$, and a numerical computation shows $C_3 \leq 6.3$.

2.4 The four-corners Cantor measure

We conclude this section with an example showing that the critical case order $\log n/\sqrt{n}$ is sharp if one generalizes its scope.

The *four-corner* Cantor set K is the compact subset of the plane defined as the attractor of the Iterated Function System (T_1,T_2,T_3,T_4) where T_i are homotheties of ratio $1/4$ centered at $(0,0)$, $(0,1)$, $(1,1)$ and $(1,0)$ (see figure 1). It has a natural measure μ_K, which can be defined as the fixed point of the map

$$\begin{aligned}\mathcal{T}\colon\ \mathcal{P}([0,1]^2) &\to \mathcal{P}([0,1]^2)\\ \nu &\mapsto \frac{1}{4}(T_1)_*\nu + \frac{1}{4}(T_2)_*\nu + \frac{1}{4}(T_3)_*\nu + \frac{1}{4}(T_4)_*\nu\end{aligned}$$

($\mathcal{T}$ is contracting in the complete metric W_1, so that it has a unique fixed point). The measure μ_K can also be described as follows. In the 4-adic decomposition of the square, at depth $j>0$ there are 16^j squares, among which 4^j intersect K in their interior; μ_K gives each of these squares a mass $1/4^j$.

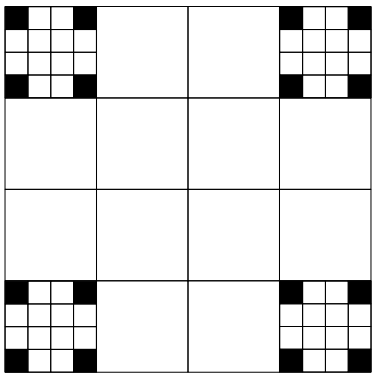

Figure 1: Second stage of the construction of the four-corners Cantor set (contained in the filled black area).

K has Hausdorff dimension 1 (and positive, finite 1-dimensional Hausdorff measure), and one should expect μ_K to have dimension $d = 1$ in any reasonable sense of the term. It is thus interesting to have a look at $\mathrm{W}_q(\hat{\mu}_n, \mu_K)$ in the critical case $q = 1/2$.

Proposition 2.6. *If $(X_k)_{k\geq 0}$ are i.i.d. of law μ_K, then*

$$\mathbb{E}\left[\mathrm{W}_{\frac{1}{2}}(\hat{\mu}_n, \mu_K)\right] \asymp \frac{\log n}{\sqrt{n}}.$$

Proof. The proof of the upper bound follows the proof of Theorem 2.1, using a 4-adic decomposition and discarding all λ such that C_λ does not intersects K in its interior. This replaces d by 1 as there are 4^j relevant squares of size 4^{-j} (indeed the only place where d is used is in (10), only through the number of dyadic squares to be considered), so that with $q = 1/2$ we end up in the critical case.

To prove the lower bound, we first record the proportions p_1, p_2, p_3, p_4 of the random points X_k lying in each of the four relevant depth-one squares (of side-length $1/4$). For large n, each p_i is close to $1/4$ with typical fluctuations of the order of $1/\sqrt{n}$. The discrepancy of mass in each of these squares compared to the mass $1/4$ given to each of them by μ_K induces a cost of at least $1/\sqrt{2n}$, since the distance between depth-one squares is at least $1/2$ and $q = 1/2$. The same reasoning applies at depth two inside each depth-one square, but with $np_i \simeq n/4$ points, thus fluctuations are of the order of $1/\sqrt{n/4} = 2/\sqrt{n}$, inducing a total cost of the order of $1/\sqrt{2n}$ (distances are now $1/4 \times 1/2$, and a square root is taken since $q = 1/2$). The fact that the number of points is np_i rather than precisely $n/4$ is not an issue, an uneven distribution improving the bound.

At each depth j up to $\log_4 n$, there is a typical induced cost of the order of $1/\sqrt{n}$ from the uneven distribution of points among the 4 subsquares of each depth j square, yielding the desired bound of the order of $\log n/\sqrt{n}$. □

3 Wavelet decomposition and convergence against regular test functions

3.1 Wavelet decomposition

Let us give a short account of the results about wavelets we will use (see e.g. Meyer's book [Mey92] for proofs and references).

It will be convenient to use wavelets of compact support with arbitrary regularity $\mathcal{C}^{\mathrm{r}}$, whose construction is due to Daubechies [Dau88]. The construction yields *compactly supported* functions $\phi, \psi^\epsilon : \mathbb{R}^d \to \mathbb{R}$ where ϵ takes any of $2^d - 1$ values ($\epsilon \in E := \{0,1\}^d \setminus \{(0,0,\dots,0)\}$), with particular properties of which only those we will use will be described.

One defines from these "father and mother" wavelets a larger family of *wavelets* by

$$\begin{aligned} \phi_\tau(x) &= \phi(x-\tau), && (\tau \in \mathbb{Z}^d) \\ \psi_\lambda(x) &= 2^{\frac{dj}{2}} \psi^\epsilon(2^j x - \tau), && (\lambda = (j,\tau,\epsilon) \in \Lambda = \mathbb{Z} \times \mathbb{Z}^d \times E); \end{aligned} \tag{16}$$

one important property of the construction is that the union of $(\phi_\tau)_{\tau\in\mathbb{Z}^d}$ and $(\psi_\lambda)_{\lambda\in\Lambda}$ form an *orthonormal basis* of $L^2(\mathbb{R}^d)$. For $f \in L^2(\mathbb{R}^d)$ we can thus write

$$f = \sum_{\tau\in\mathbb{Z}^d} \langle f, \phi_\tau\rangle \phi_\tau + \sum_{j=0}^{\infty} \sum_{\lambda\in\Lambda_j} \langle f, \psi_\lambda\rangle \psi_\lambda$$

where $\Lambda_j = \{j\} \times \mathbb{Z}^d \times E$ and $\langle\cdot,\cdot\rangle$ denotes the L^2 scalar product (with respect to Lebesgue measure).

One stunning property is that many functional spaces can be *characterized* in term of the wavelet coefficients $\alpha(\lambda) = \langle f, \psi_\lambda\rangle$ and $\beta(\tau) = \langle f, \phi_\tau\rangle$. We shall only use upper bounds on the $\alpha(\lambda)$ and $\beta(\tau)$ in a specific case.

The Hölder space $\mathcal{C}^{\mathrm{s}}$ is defined as the space of k times continuously differentiable with γ-Hölder partial derivatives of order k, with k a non-negative integer, $\gamma \in (0,1]$ and $k+\gamma = s$ (e.g. $\mathcal{C}^1$ is the space of Lipschitz functions, $\mathcal{C}^{3/2}$ the space of once continuously differentiable functions with 1/2-Hölder first-order partial derivatives, $\mathcal{C}^5$ is the space of four-times continuously differentiable functions with Lipschitz fourth-order partial derivatives, etc.). Note that "1-Hölder", meaning "Lipschitz", could be slightly enlarged to "Zygmund" (and should, if one is interested in two-sided bounds), but we need not enter this subtlety here.

The space $\mathcal{C}^{\mathrm{s}}$ is endowed with the norm

$$\|f\|_{\mathcal{C}^{\mathrm{s}}} = \max_{j\in\{0,\dots,k\}} \max_{\omega\in\{1,\dots,d\}^j} \left\| \frac{\partial^j f}{\partial x_{\omega_1} \cdots \partial x_{\omega_j}} \right\|_\star$$

where the decomposition $s = k + \gamma$ is defined as above and $\|\cdot\|_\star$ is the uniform norm if $j < k$ and is the γ-Hölder constant if $j = k$. We denote by $\mathcal{C}^{\mathrm{s}}_1$ the set of functions with $\mathcal{C}^{\mathrm{s}}$ norm at most 1.

If the regularity of the wavelets is larger than the regularity of the considered Hölder space ($r > s$) then

$$\begin{aligned} |\beta(\tau)| &\leq C_{d,s}\|f\|_\infty & \forall \tau \in \mathbb{Z}^d \\ |\alpha(\lambda)| &\leq C_{d,s}\|f\|_{\mathcal{C}^s} 2^{-\frac{dj}{2}} 2^{-js} & \forall \lambda \in \Lambda_j, \end{aligned}$$

where the constant $C_{d,s}$ depends implicitely on the choice of father and mother wavelets ϕ and ψ^ϵ; but we can fix for each s such a choice with suitable regularity, e.g. $r = s+1$ and the constants then truly depends only on d and s. The $\mathcal{C}^s$ norm in the $\alpha(\lambda)$ coefficient could be relaxed to the "regularity part" of the norm but we do not use this.

Note that the explicit computation of these constants would in particular need a very fine analysis of the chosen wavelet construction, and I do not know whether such a task has been conducted.

3.2 Decomposition of regular functions

Let us now use wavelet decomposition to prove good convergence properties for the empirical measure against smooth enough test functions; the strategy is similar to the one used in Section 2. We assume here that $(X_k)_{k\geq 0}$ is a sequence of i.i.d. random variables whose law μ is supported on a bounded set $\Omega \subset \mathbb{R}^d$ (e.g. $\Omega = [0,1]^d$); note that $\mathcal{C}^s_1 = \mathcal{C}^s_1(\mathbb{R}^d)$ makes no reference to Ω. We consider a fixed family of wavelet of regularity $r > s$ as in Section 3.1; all constants C below implicitly depend on d, s and Ω (only through its diameter).

Since the wavelets have compact support, there exist some constant C such that for each j:

- for each point $x \in [0,1]^d$, there are at most C different λ corresponding to a ψ_λ that does not vanish at x; the set of those λ is denoted by $\Lambda_j(x) \subset \Lambda_j$,
- the union $\Lambda_j(\Omega) := \bigcup_{x\in\Omega} \lambda_j(x)$ has at most $C2^{dj}$ elements.

We denote by Z the set of parameters $\tau \in \mathbb{Z}^d$ corresponding to a ϕ_τ whose support intersects Ω (observe that Z is finite).

We fix a function $f \in \mathcal{C}^s_1$ and decompose it in our wavelet basis:

$$f = \sum_{\tau\in\mathbb{Z}^d} \beta(\tau)\phi_\tau + \sum_{j=0}^{\infty} \sum_{\lambda\in\Lambda_j} \alpha(\lambda)\psi_\lambda$$

with

$$\begin{aligned} |\beta(\tau)| &\lesssim 1 & \forall \tau \in \mathbb{Z}^d \\ |\alpha(\lambda)| &\lesssim 2^{-\frac{dj}{2}} 2^{-js} & \forall \lambda \in \Lambda_j. \end{aligned}$$

Cutting the second term of the decomposition to some depth J we get:

$$f = \sum_{\tau\in Z} \beta(\tau)\phi_\tau + \sum_{j=0}^{J} \sum_{\lambda\in\Lambda_j} \alpha(\lambda)\psi_\lambda + g$$

where

$$g = \sum_{\tau \notin Z} \beta(\tau)\phi_\tau + \sum_{j>J} \sum_{\lambda \in \Lambda_j} \alpha(\lambda)\psi_\lambda.$$

Using the bound on the α coefficients and the formula (16) for ψ_λ, we get:

$$\|g\mathbf{1}_\Omega\|_\infty \lesssim 2^{-sJ}$$

and it follows:

$$|\hat{\mu}_n(f) - \mu(f)| \lesssim 2^{-Js} + \sum_{\tau \in Z} |\hat{\mu}_n(\phi_\tau) - \mu(\phi_\tau)| + \sum_{j=0}^{J} \sum_{\lambda \in \Lambda_j(\Omega)} 2^{-(\frac{d}{2}+s)j} |\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|$$

where the right-hand side does not depend on f. Taking a supremum and an expectation, it then comes:

$$\mathbb{E}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s}\right] \lesssim 2^{-sJ} + \sum_{\tau \in Z} \mathbb{E}\left[|\hat{\mu}_n(\phi_\tau) - \mu(\phi_\tau)|\right] + \sum_{j=0}^{J} \sum_{\lambda \in \Lambda_j(\Omega)} 2^{-(\frac{d}{2}+s)j}\, \mathbb{E}\left[|\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|\right] \tag{17}$$

and to conclude, we simply need to estimate the last two terms above.

3.3 Convergence for basis elements

Lemma 3.1. *We have*

$$\sum_{\tau \in Z} \mathbb{E}\left[|\hat{\mu}_n(\phi_\tau) - \mu(\phi_\tau)|\right] \lesssim \frac{1}{\sqrt{n}}$$

$$\textit{and} \qquad \sum_{\lambda \in \Lambda_j(\Omega)} \mathbb{E}\left[|\hat{\mu}_n(\psi_\lambda) - \mu(\psi_\lambda)|\right] \lesssim \frac{2^{dj}}{\sqrt{n}}$$

Proof. For each $\tau \in Z$, the random variable $\hat{\mu}_n(\phi_\tau)$ is the average of n independent identically distributed, bounded random variables of expectation $\mu(\phi_\tau)$, so that $\mathbb{E}\left[|\hat{\mu}_n(\phi_\tau) - \mu(\phi_\tau)|\right] \leq C/\sqrt{n}$. Since Z is finite, the first claim is proved.

To prove the second claim, we cannot argue in the exact same way because ψ_λ depends on j. To ease notation we introduce $\bar{\psi}_\lambda := 2^{-\frac{dj}{2}}\psi_\lambda$ and $Y_\lambda := \hat{\mu}_n(\bar{\psi}_\lambda) - \mu(\bar{\psi}_\lambda)$, and recall that $\bar{\psi}_\lambda$ is bounded independently of j. Also, a bounded number of different $\bar{\psi}_\lambda$ ($\lambda \in \Lambda_j$) are non-zero at any point $x \in \Omega$; we denote by p_λ the mass given by μ to the support of ψ_λ and observe that Y_λ is the average of n i.i.d. centered random variables of variance less than $Cp_\lambda + \mu(\bar{\psi}_\lambda)^2$. We have

$$\mathrm{Var}(Y_\lambda) \leq \frac{1}{n}\Big(Cp_\lambda + \mu(\bar{\psi}_\lambda)^2\Big) \qquad \sum_{\lambda \in \Lambda_j(\Omega)} p_\lambda \lesssim 1 \qquad \sum_{\lambda \in \Lambda_j(\Omega)} \mu(\bar{\psi}_\lambda) \lesssim 1$$

so that

$$\begin{aligned}\sum_{\lambda \in \Lambda_j(\Omega)} \mathrm{Var}(Y_k) &\leq \frac{1}{n}\Big(C \sum_{\lambda \in \Lambda_j(\Omega)} p_\lambda + \Big(\sum_{\lambda \in \Lambda_j(\Omega)} \mu(\bar{\psi}_\lambda)\Big)^2\Big) \\ &\lesssim \frac{1}{n}.\end{aligned}$$

Now it comes

$$\begin{aligned}\sum_{\lambda\in\Lambda_j(\Omega)}\mathbb{E}\left[|\hat{\mu}_n(\psi_\lambda)-\mu(\psi_\lambda)|\right] &= 2^{\frac{dj}{2}}\sum_{\lambda\in\Lambda_j(\Omega)}\mathbb{E}\left[|Y_\lambda|\right]\\ &\leq 2^{\frac{dj}{2}}\sum_{\lambda\in\Lambda_j(\Omega)}\sqrt{\mathbb{E}\left[Y_\lambda^2\right]}\\ &\leq 2^{\frac{dj}{2}}\sqrt{|\Lambda_j(\Omega)|}\sqrt{\sum_{\lambda\in\Lambda_j(\Omega)}\mathrm{Var}(Y_\lambda)}\\ &\lesssim \frac{2^{dj}}{\sqrt{n}}\end{aligned}$$

□

Remark 3.2. Lemma 3.1 is the only place where we use that the $(X_k)_{k\in\mathbb{N}}$ are i.i.d. The method can therefore be applied to any stochastic process satisfying the conclusion of Lemma 3.1.

3.4 Conclusion of the proof

Plugin Lemma 3.1 into (17) yields

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{c_1^s}\right]\lesssim 2^{-Js}+\frac{1}{\sqrt{n}}\sum_{j=0}^{J}\left(2^{\frac{d}{2}-s}\right)^j$$

and we get the same trichotomy as before. If $s>d/2$, then we can let $J\to\infty$ to obtain

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{c_1^s}\right]\leq\frac{C}{\sqrt{n}},$$

if $s=d/2$ we can take J such that $2^{-Js}\simeq 1/\sqrt{n}$ and get

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{c_1^s}\right]\leq C\frac{\log n}{\sqrt{n}},$$

and if $s<d/2$ we can choose J such that $2^J\simeq n^{\frac{1}{d}}$ to get

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{c_1^s}\right]\leq\frac{C}{n^{s/d}},$$

ending the proof of Theorem C.

4 Markov chains

In this section we assume $(X_k)_{k\geq 0}$ is a Markov chain on a bounded domain; since we will use Fourier series, it will make things simpler to embed this domain into a torus,

so we assume $\Omega \subset \mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ (we do not lose generality in doing so, as scaling down Ω makes it possible to make the embedding isometric). We still denote by $\|x-y\|$ the distance between two points induced by the Euclidean norm.

Our main assumption is that the iterated transition kernel of $(X_k)_{k\geq 0}$, defined by

$$m_x(A) = \mathbb{P}(X_{k+1} \in A \mid X_k = x) \qquad m_x^t(A) = \mathbb{P}(X_{k+t} \in A \mid X_k = x)$$

is exponentially contracting in W_1, i.e. there are constants $D \geq 1$ and $\theta \in (0,1)$ such that

$$\mathrm{W}_1(m_x^t, m_y^t) \leq D\theta^t \|x-y\|. \tag{18}$$

Let us denote by L the averaging operator, i.e.

$$\mathrm{L}f(x) = \int f(y)\, \mathrm{d}m_x(y)$$

and by L^* its dual acting on probability measure, i.e. $\mathrm{L}^*\nu$ is the law of X_{k+1} conditioned on X_k having law ν. The linearity of W_1 enables one to rewrite (18) as

$$\mathrm{W}_1(\mathrm{L}^{*t}\nu_0, \mathrm{L}^{*t}\nu_1) \leq D\theta^t\, \mathrm{W}_1(\nu_0, \nu_1) \tag{19}$$

so that there is a unique stationary measure μ, and the law of X_k converges exponentially fast (in W_1) to μ, whatever the law of X_0 is.

We shall prove Theorem A, which we restate for convenience.

Theorem 4.1. *For some constant $C = C(\Omega, d, D, s)$ and all large enough n, letting $\bar{n} = (1-\theta)n$, we have*

$$\mathbb{E}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s}\right] \leq C \begin{cases} \dfrac{(\log \bar{n})^{\frac{d}{2s+1}}}{\sqrt{\bar{n}}} & \text{when } s > d/2 \\ \dfrac{\log \bar{n}}{\sqrt{\bar{n}}} & \text{when } s = d/2 \\ \dfrac{(\log \bar{n})^{d-2s+\frac{s}{d}}}{\bar{n}^{\frac{s}{d}}} & \text{when } s < d/2 \end{cases} \tag{20}$$

Following the decomposition method, we shall find a suitable decomposition basis for any $f \in \mathcal{C}_1^s$, seeking for a compromise between precision of a truncated decomposition and number of basis elements. Here using wavelets seems inefficient, as we do not have a precise enough analogue of Lemma 3.1, which uses independence to take advantage of the localization property of wavelets; without this, the number and size of the ψ_λ are overwhelming. We shall use Fourier series instead, as they will be more easily controlled under our assumptions. For simplicity we consider complex-valued functions here, and denote the Fourier basis by $e_k(x) := e^{2i\pi k\cdot x}$ where $k \in \mathbb{Z}^d$ and the dot $\cdot$ denotes the canonical inner product.

The key is thus to control $|\hat{\mu}_n(e_k) - \mu(e_k)|$; our hypothesis may seem perfectly suited to this since e_k is Lipschitz, but its Lipschitz constant grows too rapidly with k for a direct approach to be efficient. We shall combine the following two observations (the first of which is pretty trivial, the second of which is folklore).

Lemma 4.2. *For all $\alpha \in (0,1)$, we have the following control of e_k's α-Hölder constant:*

$$\mathrm{Hol}_\alpha(e_k) \lesssim |k|_\infty^\alpha$$

where $|k|_\infty = \max\Big\{k_i : i \in \{1,\dots,d\}\Big\}$.

Proof. We have $\mathrm{Lip}(e_k) \le 2\pi\sqrt{d}|k|_\infty$ and $\|e_k\|_\infty \le 1$ so that for all $x \neq y \in \mathbb{T}^d$:

$$\frac{|e_k(x)-e_k(y)|}{\|x-y\|^\alpha} \le \min\Big(\frac{2}{\|x-y\|^\alpha}, 2\pi\sqrt{d}|k|_\infty\|x-y\|^{1-\alpha}\Big) \le 2\pi^\alpha d^{\frac{\alpha}{2}}|k|_\infty^\alpha$$

□

Lemma 4.3. *For all $\alpha \in (0,1]$, denoting by W_α the α-Wasserstein metric (i.e. the 1-Wasserstein metric associated with the modified distance $\|\cdot\|^\alpha$), we have*

$$\mathrm{W}_\alpha(\mathrm{L}_0^{*t}\nu_0, \mathrm{L}_0^{*t}\nu_1) \le D^\alpha\theta^{\alpha t}\,\mathrm{W}_\alpha(\nu_0,\nu_1) \tag{21}$$

As a consequence, for all α-Hölder functions $f : \Omega \to \mathbb{C}$ and all $\ell, m \in \mathbb{N}$ it holds

$$\Big|\mathbb{E}[f(X_\ell)] - \mu(f)\Big| \lesssim \mathrm{Hol}_\alpha(f)\,\theta^{\alpha\ell}$$
$$\Big|\mathbb{E}[f(X_m)f(X_\ell)] - \mathbb{E}[f(X_m)]\,\mathbb{E}[f(X_\ell)]\Big| \lesssim \mathrm{Hol}_\alpha(f)^2\,\theta^{\alpha|m-\ell|}$$

where the implied constants depends only on Ω and the constant C in (18).

Proof. By linearity we only have to check (21) when $\nu_0 = \delta_x$ and $\nu_1 = \delta_y$ for some $x, y \in \Omega$, and by concavity

$$\mathrm{W}_\alpha(\mathrm{L}^{*t}\delta_x, \mathrm{L}^{*t}\delta_y) \le \Big(\mathrm{W}_1(\mathrm{L}^{*t}\delta_x, \mathrm{L}^{*t}\delta_y)\Big)^\alpha \le D^\alpha\theta^{\alpha t}\|x-y\|^\alpha = D^\alpha\theta^{\alpha t}\,\mathrm{W}_\alpha(\delta_x,\delta_y).$$

To prove convergence toward the average and decay of correlation, we first use the contraction and that μ is the stationary measure to get

$$\begin{aligned}
\Big|\mathrm{L}^t f(x) - \mu(f)\Big| &= \Big|\int \mathrm{L}^t f\,\mathrm{d}\delta_x - \int f\,\mathrm{d}\mu\Big| \\
&= \Big|\int f\,\mathrm{d}\Big(\mathrm{L}^{*t}\delta_x\Big) - \int f\,\mathrm{d}\Big(\mathrm{L}^{*t}\mu\Big)\Big| \\
&\le \mathrm{Hol}_\alpha(f)\,\mathrm{W}_\alpha(\mathrm{L}^{*t}\delta_x, \mathrm{L}^{*t}\mu) \\
&\le \mathrm{Hol}_\alpha(f)\,D^\alpha\theta^{\alpha t}\,\mathrm{W}_\alpha(\delta_x,\mu) \\
\Big|\mathrm{L}^t f(x) - \mu(f)\Big| &\lesssim \mathrm{Hol}_\alpha(f)\,\theta^{\alpha t}.
\end{aligned}$$

Assuming without lost of generality $\mu(f) = 0$ we have $\|f\|_\infty \lesssim \mathrm{Hol}_\alpha(f)$ ($\mu(f) = 0$ implies that f takes both non-positive and non-negative values, and Ω is bounded).

Assume further $m \geq \ell$ and write $m = \ell + t$. Combining all previous observations we get:

$$\begin{aligned}
\|\mathrm{L}^t f\|_\infty &\lesssim \mathrm{Hol}_\alpha(f)\, \theta^{\alpha t},\\
\Big|\mathbb{E}[f(X_m)]\Big| &= \Big|\mathbb{E}\Big[\mathrm{L}^m f(X_0)\Big]\Big|\\
&\lesssim \mathrm{Hol}_\alpha(f)\, \theta^{\alpha m},\\
\Big|\mathbb{E}[f(X_\ell)]\Big| &\lesssim \mathrm{Hol}_\alpha(f)\, \theta^{\alpha \ell},\\
\Big|\mathbb{E}[f(X_m) f(X_\ell)]\Big| &= \Big|\mathbb{E}\Big[\mathrm{L}^t f(X_\ell)\, f(X_\ell)\Big]\Big|\\
&\lesssim \|\mathrm{L}^t f\|_\infty\, \mathbb{E}[|f(X_\ell)|]\\
&\lesssim \mathrm{Hol}_\alpha(f)^2 \theta^{\alpha t}
\end{aligned}$$

and the conclusion follows. □

We deduce the following from these two Lemmas.

Corollary 4.4. *For all k, α and all $n \geq 1/(1-\theta^\alpha)$ it holds*

$$\mathbb{E}\Big[|\hat{\mu}_n(e_k) - \mu(e_k)|^2\Big] \lesssim \frac{|k|_\infty^{2\alpha}}{(1-\theta^\alpha)n}$$

Proof. We have:

$$\begin{aligned}
\mathbb{E}\Big[|\hat{\mu}_n(e_k) - \mu(e_k)|^2\Big] &= \mathbb{E}\left[\Big(\frac{1}{n}\sum_{\ell=1}^n e_k(X_\ell) - \mu(e_k)\Big)^2\right]\\
&= \frac{1}{n^2}\sum_{1\leq \ell, m\leq n} \mathbb{E}[e_k(X_\ell) e_k(X_m)] - \frac{2}{n}\sum_{\ell=1}^n \mathbb{E}[e_k(X_\ell)]\mu(e_k) + \mu(e_k)^2\\
&\leq \frac{1}{n^2}\Big(\sum_{1\leq \ell, m\leq n} \mathbb{E}[e_k(X_\ell)]\, \mathbb{E}[e_k(X_m)] + C\, \mathrm{Hol}_\alpha(e_k)^2\, \theta^{\alpha|\ell - m|}\Big)\\
&\qquad - \frac{2}{n}\sum_{\ell=1}^n \mathbb{E}[e_k(X_\ell)]\mu(e_k) + \mu(e_k)^2\\
&\leq \frac{C\, \mathrm{Hol}_\alpha(e_k)^2}{n^2} \sum_{1\leq \ell, m\leq n} \theta^{\alpha|\ell-m|} + \frac{1}{n^2}\Big(\sum_{\ell=1}^n \Big(\mathbb{E}[e_k(X_\ell)] - \mu(e_k)\Big)\Big)^2\\
&\lesssim \frac{\mathrm{Hol}_\alpha(e_k)^2}{n^2} \cdot \sum_{\ell=1}^n 2\sum_{t=0}^\infty \theta^{\alpha t} + \frac{\mathrm{Hol}_\alpha(e_k)^2}{n^2}\Big(\sum_{\ell=1}^n \theta^{\alpha\ell}\Big)^2\\
&\lesssim \frac{\mathrm{Hol}_\alpha(e_k)^2}{n^2} \cdot \frac{n}{1-\theta^\alpha} + \frac{\mathrm{Hol}_\alpha(e_k)^2}{n^2(1-\theta^\alpha)^2}\\
&\lesssim \frac{|k|_\infty^{2\alpha}}{(1-\theta^\alpha)n}
\end{aligned}$$

whenever $n \geq 1/(1-\theta^\alpha)$. □

Fix some threshold $J \geq 3$ and some exponent $\alpha \in (0,1]$, to be determined explicitly later on.

Let $f : \mathbb{T}^d \to \mathbb{R}$ be in $\mathcal{C}_1^s$. From the multidimensional version of Jackson's theorem [Sch69], we know that there is a trigonometric polynomial $T_J(f)$ which is a linear combination of the e_k for $|k|_\infty \leq J$, such that

$$\|f - T_J(f)\|_\infty \lesssim \frac{1}{J^s}$$

We have no clear control on the coefficient of this optimal trigonometric polynomial, which need not be the Fourier coefficients of f. But it is also known that the Fourier series of f is within a factor $\simeq \|f\|_\infty (\log J)^d$ of the best approximation (see [Mas80] for an optimal constant), so that denoting by $F_J(f) := \sum_{|k|_\infty \leq J} \hat{f}_k e_k$ the J-truncation of the Fourier series of f, we get

$$\|f - F_J(f)\|_\infty \lesssim \frac{(\log J)^d}{J^s}.$$

We can assume $\hat{f}_0 = 0$ by translating f, and what precedes yields:

$$\begin{aligned}
|\hat{\mu}_n(f) - \mu(f)| &\leq |\hat{\mu}_n(f) - \hat{\mu}_n(F_J(f))| + |\hat{\mu}_n(F_J(f)) - \mu(F_J(f))| + |\mu(F_J(f)) - \mu(f)| \\
&\leq 2\|f - F_J(f)\|_\infty + \sum_{0<|k|_\infty \leq J} |\hat{f}_k||\hat{\mu}_n(e_k) - \mu(e_k)| \qquad (22) \\
&\lesssim \frac{(\log J)^d}{J^s} + \left(\sum_{0<|k|_\infty \leq J} |\hat{f}_k|^2 |k|_\infty^{2s} \right)^{\frac{1}{2}} \left(\sum_{0<|k|_\infty \leq J} \frac{|\hat{\mu}_n(e_k) - \mu(e_k)|^2}{|k|_\infty^{2s}} \right)^{\frac{1}{2}} \\
&\lesssim \frac{(\log J)^d}{J^s} + \|f\|_{H^s} \left(\sum_{0<|k|_\infty \leq J} \frac{|\hat{\mu}_n(e_k) - \mu(e_k)|^2}{|k|_\infty^{2s}} \right)^{\frac{1}{2}} \\
|\hat{\mu}_n(f) - \mu(f)| &\lesssim \frac{(\log J)^d}{J^s} + \left(\sum_{0<|k|_\infty \leq J} \frac{|\hat{\mu}_n(e_k) - \mu(e_k)|^2}{|k|_\infty^{2s}} \right)^{\frac{1}{2}} \qquad (23)
\end{aligned}$$

Where the right-hand side does not depend on f in any way (note that $\|\cdot\|_{H^s}$ is the Sobolev norm, controlled by the $\mathcal{C}^s$ norm).

Remark 4.5. At line (22), one could be tempted to bound directl $|\hat{f}_k|$ instead of using the Cauchy-Schwarz inequality, in order to make better use of our assumption on f. This would be effective if $|\hat{\mu}_n(e_k) - \mu(e_k)|$ were of the order of $1/n$, but it is actually of the order of $1/\sqrt{n}$, ultimately leading to a weaker bound than the one we aim for.

Taking a supremum and an expectation in (23) and using concavity, it comes:

$$\begin{aligned}\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{\mathcal{C}_1^s}\right] &\lesssim \frac{(\log J)^d}{J^s}+\left(\sum_{0<|k|_\infty\le J}\frac{\mathbb{E}\left[|\hat{\mu}_n(e_k)-\mu(e_k)|^2\right]}{|k|_\infty^{2s}}\right)^{\frac{1}{2}}\\ &\lesssim \frac{(\log J)^d}{J^s}+\left(\sum_{0<|k|_\infty\le J}\frac{|k|^{2\alpha}}{(1-\theta^\alpha)n|k|_\infty^{2s}}\right)^{\frac{1}{2}}\\ &\lesssim \frac{(\log J)^d}{J^s}+\left(\sum_{\ell=1}^{J}\frac{\ell^{d-1+2\alpha-2s}}{(1-\theta^\alpha)n}\right)^{\frac{1}{2}}\end{aligned}$$

Choose now $\alpha = 1/\log J$ so that $\ell^{2\alpha}\lesssim 1$ for all $\ell\in\{1,\dots,J\}$, use $1-\theta^\alpha\ge\alpha(1-\theta)$ and set $\bar{n}:=(1-\theta)n$ to obtain

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{\mathcal{C}_1^s}\right]\lesssim\frac{(\log J)^d}{J^s}+\sqrt{\frac{\log J}{\bar{n}}}\Big(\sum_{\ell=1}^{J}\ell^{d-1-2s}\Big)^{\frac{1}{2}} \tag{24}$$

For $s<d/2$, we get:

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{\mathcal{C}_1^s}\right]\lesssim\frac{(\log J)^d}{J^s}+\frac{(\log J)^{\frac{1}{2}}J^{\frac{d}{2}-s}}{\sqrt{\bar{n}}} \tag{25}$$

Trying to balance the contribution of the two terms, we first see that taking $J\simeq\bar{n}^{\frac{1}{d}}$ would optimize the power of $\bar{n}$ in the final expression; refining to $J=(\log\bar{n})^\beta\bar{n}^{\frac{1}{d}}$, developing and ignoring lower order terms shows that the choice $\beta=2-\frac{1}{d}$ optimizes the final power of $\log\bar{n}$, and we thus set

$$J=\left\lfloor(\log\bar{n})^{2-\frac{1}{d}}\bar{n}^{\frac{1}{d}}\right\rfloor$$

Any large enough n (the bound depending on both θ and d) satisfies the requirement $n\ge 1/(1-\theta^\alpha)$ since the right-hand side is of the order of $\log n$. It then comes:

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{\mathcal{C}_1^s}\right]\lesssim\frac{(\log\bar{n})^{d-2s+\frac{s}{d}}}{\bar{n}^{\frac{s}{d}}}\qquad(n\text{ large enough}).$$

For $2s=d$ we get

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{\mathcal{C}_1^s}\right]\lesssim\frac{(\log J)^d}{J^s}+\frac{\log J}{\sqrt{\bar{n}}}$$

and taking $J=\lfloor\bar{n}^{\frac{1}{2s}}(\log\bar{n})^{(d-1)/s}\rfloor$ yields

$$\mathbb{E}\left[\,\mathrm{W}_1(\hat{\mu}_n,\mu)\right]\lesssim\frac{\log\bar{n}}{\sqrt{\bar{n}}}.$$

Finally, for $s>d/2$ we get

$$\mathbb{E}\left[\|\hat{\mu}_n-\mu\|_{\mathcal{C}_1^s}\right]\lesssim\frac{(\log J)^d}{J^s}+\frac{(\log J)^{\frac{1}{2}}}{\sqrt{\bar{n}}}$$

and taking $J = \lfloor \bar{n}^{\frac{1}{2s}} (\log \bar{n})^{\frac{d}{s+1/2}} \rfloor$ yields

$$\mathbb{E}\Big[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^s}\Big] \lesssim \frac{(\log \bar{n})^{\frac{d}{2s+1}}}{\sqrt{\bar{n}}},$$

ending the proof of Theorem A.

5 Concentration near the expectancy

Let us detail how classical bounded martingale difference methods can be used to prove that the empirical measure concentrates very strongly around its expectancy. When $(X_k)_{k\geq 0}$ are independent identically distributed, this is long-known (see [Tal92], and also [WB17] for more general Wasserstein metrics W_p, $p \geq 1$). In the case of Markov chains, such arguments have been developed notably in [CR09] and, in a dynamical context, [CG12]. Our approach is very similar and thus cannot pretend to novelty, but we write it down to show how to handle functional spaces more general than just Lipschitz and Hölder.

The fundamental result to be used is the Azuma-Hoeffding inequality, which we recall.

Theorem (Azuma-Hoeffding inequality)**.** *Let Y be a random variable, let*

$$\{\varnothing, \Omega\} = \mathscr{B}_0 \subset \mathscr{B}_1 \subset \cdots \subset \mathscr{B}_n = \mathscr{B}(\Omega)$$

be a filtration and for each $k \in [\![1, n]\!]$ set $\Delta_k = \mathbb{E}[Y|\mathscr{B}_k] - \mathbb{E}[Y|\mathscr{B}_{k-1}]$. Assume that for all k and some numbers $a_k \in \mathbb{R}$, $c_k > 0$ we have $\Delta_k \in [a_k, a_k + c_k]$ almost surely. Then for all $t > 0$,

$$\mathbb{P}\Big[Y \geq \mathbb{E}[Y] + t\Big] \leq \exp\Big(-\frac{2t^2}{\sum_k c_k^2}\Big).$$

5.1 The independent case

In the case of i.i.d. random variables, the Azuma-Hoeffding inequality famously yields the following concentration inequality.

Theorem (McDiarmid's inequality)**.** *Let $F : \Omega^n \to \mathbb{R}$ be a function such that for some $c_1, \ldots, c_n$ and all $k \in [\![1, n]\!]$ and all $(x_1, \ldots, x_n, x_k') \in \Omega^{n+1}$ it holds*

$$\Big|F(x_1, \ldots, x_k, \ldots, x_n) - F(x_1, \ldots, x_k', \ldots, x_n)\Big| \leq c_k.$$

Let $(X_k)_{1\leq k\leq n}$ be a sequence of independent random variables. Then for all $t > 0$ it holds

$$\mathbb{P}\Big[F(X_1, \ldots, X_n) \geq \mathbb{E}[F(X_1, \ldots, X_n)] + t\Big] \leq \exp\Big(-\frac{2t^2}{\sum_k c_k^2}\Big).$$

Applying this to

$$F(X_1,\dots,X_n) = \|\hat{\mu}_n - \mu\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mu(f) \right|$$

we can take

$$c_k = \frac{1}{n} \sup_{f\in\mathscr{F}, x, x'\in\Omega} |f(x) - f(x')| =: \frac{1}{n}\operatorname{osc}(\mathscr{F})$$

and it comes

$$\mathbb{P}\Big[F(X_1,\dots,X_n) \geq \mathbb{E}[F(X_1,\dots,X_n)] + t\Big] \leq \exp\Big(-\frac{2nt^2}{\operatorname{osc}(\mathscr{F})^2}\Big).$$

For example if $\mathscr{F} \subset \mathrm{Lip}_1(\Omega)$ (e.g. $\mathscr{F} = \mathcal{C}_1^{\mathrm{s}}$) we have $\operatorname{osc}(\mathscr{F}) \leq \operatorname{diam}\Omega$; if moreover $\Omega = [0,1]^d$ it thus comes

$$\mathbb{P}\Big[\|\hat{\mu}_n - \mu\|_{\mathscr{F}} \geq \mathbb{E}\left[\|\hat{\mu}_n - \mu\|_{\mathscr{F}}\right] + t\Big] \leq \exp\Big(-\frac{2}{d}\cdot nt^2\Big). \tag{26}$$

This, combined with Theorem C, yields good concentration estimates.

Corollary 5.1. *If $(X_k)_{k\geq 0}$ are i.i.d.random variables with law μ, then for all $s \geq 1$, for some constant $C = C(d,s) > 0$ (not depending upon μ), all integer $n \geq 2$ and all $M \geq C$ we have:*

- *if $s > d/2$*

$$\mathbb{P}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^{\mathrm{s}}} \geq \frac{M}{\sqrt{n}}\right] \leq e^{-\frac{2}{d}(M-C)^2}; \tag{27}$$

- *if $s = d/2$*

$$\mathbb{P}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^{\mathrm{s}}} \geq \frac{M\log n}{\sqrt{n}}\right] \leq e^{-\frac{2}{d}(M-C)^2(\log n)^2}; \tag{28}$$

- *if $s < d/2$*

$$\mathbb{P}\left[\|\hat{\mu}_n - \mu\|_{\mathcal{C}_1^{\mathrm{s}}} \geq \frac{M}{n^{\frac{s}{d}}}\right] \leq e^{-\frac{2}{d}(M-C)^2 n^{1-2s/d}}; \tag{29}$$

Similarly, with Theorem B we can obtain entirely explicit, non-asymptotic concentration bounds.

5.2 Markov Chains

To tackle Markov chains we will need some hypothesis to replace independence; we choose a framework that covers the case of W_1, but also more general dual metrics $\|\cdot\|_{\mathscr{F}}$.

Assume that Ω is endowed with a metric d with finite diameter (d is assumed to be lower-semi-continuous, but not necessarily to induce the given topology on Ω). We still denote by $\mathrm{Lip}_1(\Omega)$ be the space of functions $\Omega \to \mathbb{R}$ which are 1-Lipschitz with respect to d.

Let $(X_k)_{\geq 0}$ be a Markov chain on Ω which is exponentially contracting (see the beginning of Section 4) with constant D and rate θ, in the metric d instead of the euclidean norm; this can be rewritten in a coupling formulation as follows: for all $x, x' \in \Omega$, all $i, t \in \mathbb{N}$ there are random variables $(X'_k)_{k\geq i}$ with the same law as $(X'_k)_{k\geq i}$ and such that for all t:

$$\mathbb{E}[d(X_{i+t}, X'_{i+t}) \mid X_i = x, X'_i = x'] \leq D\theta^t d(x, x').$$

Note that the flexibility in the choice of d enables to include uniformly ergodic Markov chains in this framework, simply by taking $d = \mathbf{1}_{\neq}$, i.e. $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ otherwise.

Given a multivariate function $\Phi : \Omega^n \to \mathbb{R}^n$, we define as usual the coordinate-wise Lipschitz constants of Φ by

$$\Lambda_i(\Phi) = \sup_{x_1,\dots,x_n \in \Omega, x'_i \neq x_i} \frac{|\Phi(x_1, \dots, x_i, \dots, x_n) - \Phi(x_1, \dots, x'_i, \dots, x_n)|}{d(x_i, x'_i)}$$

and we say that Φ is separately Lipschitz if $\Lambda_i(\Phi) < \infty$ for all i (when $d = \mathbf{1}_{/=}$, the coordinate-wise Lipschitz constant become the coordinate-wise oscillations).

Theorem 5.2. *Let $(X_k)_{k\geq 1}$ be a Markov chain whose kernel is exponentially contracting with constant $D \geq 1$ and rate $\theta \in (0, 1)$, with respect to a lower-semi-continuous distance d on Ω giving it finite diameter* $\operatorname{diam}(\Omega)$.

Let $n \in \mathbb{N}$ and $\Phi : \Omega^n \to \mathbb{R}$ be separately Lipschitz with constants $\Lambda_i(\Phi) \leq \Lambda$. Then

$$\mathbb{P}\left[\Phi(X_1, \dots, X_n) \geq \mathbb{E}[\Phi(X_1, \dots, X_n)] + t\right] \leq \exp\left(-\frac{(1-\theta)^2 t^2}{2nD^2 \operatorname{diam}(\Omega)^2 \Lambda^2}\right)$$

Proof. We set $X = (X_1, \dots, X_n)$ and $X_{i:j} = (X_i, \dots, X_j)$ (meaning the empty family whenever $j < i$).

We shall apply the Azuma-Hoeffding inequality with the filtration $\mathscr{B}_k = \sigma(X_1^k)$, leaving us with the task of bounding the oscillations c_k of the random variable

$$\Delta_k = \mathbb{E}[\Phi(X)|X_{1:k}] - \mathbb{E}[\Phi(X)|X_{1:k-1}].$$

Given an arbitrary $x_{1:k} = (x_1, \dots, x_k) \in \Omega^k$ and $x'_k \in \Omega$ we set

$$V_k(x_{1:k}, x'_k) = \mathbb{E}[\Phi(X)|X_{1:k} = x_{1:k}] - \mathbb{E}[\Phi(X)|X_{1:k-1} = x_{1:k-1}, X_k = x'_k]$$

so that $c_k = \sup V_k - \inf V_k \leq 2\|V_k\|_\infty$. Let $(X'_i)_{i\geq k}$ be a copy of $(X_i)_{i\geq k}$ as in the definition of exponential contraction; then

$$\begin{aligned}
V_k(x_{1:k}, x'_k) &= \mathbb{E}\left[\Phi(x_{1:k-1}, X_{k:n})\middle|X_k = x_k\right] - \mathbb{E}\left[\Phi(x_{1:k-1}, X'_{k:n})\middle|X'_k = x'_k\right] \\
&= \sum_{i=k}^{n} \mathbb{E}\left[\Phi(x_{1:k-1}, X_{k:i}, X'_{i+1:n}) - \Phi(x_{1:k-1}, X_{k:i-1}, X'_{i:n})\middle|X_k = x_k, X'_k = x'_k\right] \\
|V_k(x_1^k, x'_k)| &\leq \sum_{i=k}^{n} \mathbb{E}\left[\Lambda d(X_i, X'_i)\middle|X_k = x_k, X'_k = x'_k\right] \\
&\leq D\Lambda d(x_k, x'_k) \sum_{i=k}^{\infty} \theta^{i-k} \\
c_k &\leq 2C\Lambda \operatorname{diam}(\Omega)/(1-\theta).
\end{aligned}$$

Applying the Azuma-Hoeffding inequality finishes the proof. □

Remark 5.3. The above inequality is probably not optimal; one can expect to improve the rate, either by moving the constant 2 from the denominator to the numerator, or by replacing $(1-\theta)^2$ by $(1-\theta)$ (probably with another constant).

As soon as $\mathscr{F} \subset \mathrm{Lip}_1(\Omega)$ (e.g. $\mathscr{F} = \mathcal{C}_1^{\mathrm{s}}$), Theorem 5.2 applies to

$$\Phi(X) = \|\hat{\mu}_n - \mu\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{k=1}^{n} f(X_k) - \mu(f)$$

with $\Lambda = \frac{1}{n}$, yielding

$$\mathbb{P}\left[\|\hat{\mu}_n - \mu\|_{\mathscr{F}} \geq \mathbb{E}\left[\|\hat{\mu}_n - \mu\|_{\mathscr{F}}\right] + t\right] \leq \exp\left(-\frac{(1-\theta)^2}{2D^2 \operatorname{diam}(\Omega)^2} \cdot nt^2\right) \tag{30}$$

i.e., as in the independent case, subgaussian concentration. Corollary D follows.

References

[AKT84] Miklós Ajtai, János Komlós, and Gábor Tusnády, *On optimal matchings*, Combinatorica **4** (1984), no. 4, 259–264. 1.1

[AST16] Luigi Ambrosio, Federico Stra, and Dario Trevisan, *A pde approach to a 2-dimensional matching problem*, arXiv preprint arXiv:1611.04960 (2016). 1.1

[BLG14] Emmanuel Boissard and Thibaut Le Gouic, *On the mean speed of convergence of empirical and occupation measures in wasserstein distance*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques **50** (2014), no. 2, 539–563. 1.1, 1.3, 2.1

[CG12] Jean-René Chazottes and Sébastien Gouëzel, *Optimal concentration inequalities for dynamical systems*, Communications in Mathematical Physics **316** (2012), no. 3, 843–889. 5

[CR09] Jean Rene Chazottes and Frank Redig, *Concentration inequalities for markov processes via coupling*, Electronic Journal of Probability **14** (2009), 1162–1180. 5

[Dau88] Ingrid Daubechies, *Orthonormal bases of compactly supported wavelets*, Communications on pure and applied mathematics **41** (1988), no. 7, 909–996. 3.1

[DSS13] Steffen Dereich, Michael Scheutzow, and Reik Schottstedt, *Constructive quantization: Approximation by empirical measures*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques **49** (2013), no. 4, 1183–1203. 1.1, 1.3, 2.1

[FG15] Nicolas Fournier and Arnaud Guillin, *On the rate of convergence in wasserstein distance of the empirical measure*, Probability Theory and Related Fields **162** (2015), no. 3-4, 707–738. 1.1, 1.3, 2.1

[JO10] Aldéric Joulin and Yann Ollivier, *Curvature, concentration and error estimates for Markov chain Monte Carlo*, Ann. Probab. **38** (2010), no. 6, 2418–2442. MR 2683634 1.2

[Klo17a] Benoît Kloeckner, *An optimal transportation approach to the decay of correlations for non-uniformly expanding maps*, arXiv:1711.08052, 2017. 1.2

[Klo17b] Benoît R. Kloeckner, *Effective limit theorems for markov chains with a spectral gap*, arXiv:1703.09623, 2017. 1.2

[KLS15] Benoît R Kloeckner, Artur O Lopes, and Manuel Stadlbauer, *Contraction in the wasserstein metric for some markov chains, and applications to the dynamics of expanding maps*, Nonlinearity **28** (2015), no. 11, 4117, arXiv:1412.0848. 1.2

[Mas80] J. C. Mason, *Near-best multivariate approximation by Fourier series, Chebyshev series and Chebyshev interpolation*, J. Approx. Theory **28** (1980), no. 4, 349–358. MR 589990 4

[Mey92] Yves Meyer, *Wavelets and operators*, vol. 1, Cambridge university press, 1992. 3.1

[Oll09] Yann Ollivier, *Ricci curvature of markov chains on metric spaces*, Journal of Functional Analysis **256** (2009), no. 3, 810–864. 1.2, 1.2

[Sch69] Martin H Schultz, *l^∞-multivariate approximation theory*, SIAM Journal on Numerical Analysis **6** (1969), no. 2, 161–183. 4

[Tal92] Michel Talagrand, *Matching random samples in many dimensions*, The Annals of Applied Probability (1992), 846–856. 1.1, 1.3, 2.2, 5

[Tal94] ———, *Sharper bounds for gaussian and empirical processes*, The Annals of Probability (1994), 28–76. 1.1

[VdVW96] AW Van der Vaart and JA Wellner, *Weak convergence and empirical processes*, Springer, New York, 1996. 1.4

[vH96] Ramon van Handel, *Probability in high dimension*, 1996, APC 550 Lecture Notes, Princeton University, http://www.princeton.edu/~rvan/APC550.pdf. 1.4

[WB17] Jonathan Weed and Francis Bach, *Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance*, arXiv preprint arXiv:1707.00087 (2017). 1.1, 5