On the convergence to equilibrium of Kac's random walk on matrices

Roberto I. Oliveira*

May 18, 2007

Abstract

We consider Kac's random walk on n-dimensional rotation matrices, where each step is a random rotation in the plane generated by two randomly picked coordinates. We show that this process converges to the uniform (Haar) measure in the (Wasserstein) transportation cost metric in $O\left(n^2 \ln n\right)$ steps. This improves on previous results of Diaconis/Saloff Coste and Pak/Sidenko and is a $\ln n$ factor away from being optimal.

Our proof method includes a general result akin to the path coupling method of Bubley and Dyer. Suppose that P is a Markov chain on a Polish length space (M, d) and that for all $x, y \in M$ with $d(x, y) \ll 1$ there is a coupling (X, Y) of one step from P from x and y (respectively) that is $(c + o(1))$-contracting on average. Then the map $\mu \mapsto \mu P$ is c-contracting in the transportation cost metric. Other applications of this result are also presented.

1 Introduction

Around 50 years ago Kac [6] introduced a one-dimensional toy model of a Boltzmann gas. It is a discrete-time Markov process whose state at a time $t \in \{0, 1, 2, 3, \dots\}$ is a vector $v(t) = (v_1(t), \dots, v_n(t)) \in \mathbb{R}^n$, corresponding to the velocities of of n interacting particles of equal mass. At each time t, a uniformly distributed pair $1 \leq i_t < j_t \leq n$ and a uniform angle $\theta_t \in (0, 2\pi]$ are chosen independently. This choice corresponds to a collision between particles i_t, j_t whose velocities are changed to new values $v_i(t+1) = \cos\theta_t v_{i_t}(t) + \sin\theta_t v_{j_t}(t)$, $v_{j_t}(t+1) = \cos\theta_t v_{j_t}(t) - \sin\theta_t v_{i_t}(t)$, whereas the other velocities are kept the same. This prescription for the new velocities implies that the *total kinetic energy* $E(t) \equiv \sum_{k=1}^{n} v_k(t)^2$ is conserved.

For each time step t,

$$v(t+1) = R(i_t, j_t, \theta_t)v(t),$$

*IMPA, Rio de Janeiro, RJ, Brazil, 22430-040. rimfo@impa.br

where $R(i_t, j_t, \theta_t)$ is a rotation by θ_t of the plane generated by the coordinates i_t and j_t in n-dimensional space. Two related processes have been studied in the literature under the heading of "Kac's random walk":

- On the one hand, one might focus on the evolution of $v(0), v(1), v(2), v(3), \ldots$. If one sets $E(0) = 1$ (so that $E(t) = 1$ for all t), one obtains an ergodic Markov chain over the $n-1$-dimensional sphere $S^{n-1} \subset \mathbb{R}^n$ with uniform invariant distribution;

- On the other hand, one might consider the random walk on rotation matrices determined by choosing some $X(0)$ and then setting $X(t+1) = R(i_t, j_t, \alpha_t)X(t)$ $(t \geq 0)$. This corresponds to a discrete-time ergodic random walk on the group $SO(n)$ of $n \times n$ rotation matrices (as long as $X(0) \in SO(n)$) whose stationary distribution is Haar (uniform) measure.

Kac deals with different aspects of the first process, such as entropy production and "propagation of chaos" (i.e. approximate factorizability of two-particle density functions); see the original [6] and the more recent [4, 3] for discussions and references. The second process is also mathematically natural and has been considered as a Monte Carlo method to sample approximately from Haar measure [5] and do dimensionality reduction [1].

The natural question arises of how fast Kac's random walk on $SO(n)$ converges to equilibrium. This question may be posed different forms. L^2 convergence to equilibrium is well understood since Carlen, Carvalho and Loss [3] obtained the exact spectral gap and Maslin [8] computed the entire spectrum of the two processes. Convergence in total variation also occurs, as shown by Diaconis and Saloff-Coste [4] who obtained a very poor $e^{O(n^2)}$ mixing time bound for convergence in total variation of the matrix process. We cannot improve on this bound, but note that total variation is too stringent a notion of convergence for many applications (as it is sensitive to errors at arbitrarily small scales), whereas L^2 convergence seems to be too weak (e.g. when one starts from a discrete distribution).

We consider an intermediate notion of convergence to equilibrium based on *transportation cost*, for which we prove near-optimal convergence estimates. Given a metric space (M, d) and two probability measures μ, ν over the Borel σ-field of M, the *p-transportation cost (or Wasserstein) distance between μ and ν*, denoted by $W_{d,p}(\mu, \nu)$ is the infimum of $(\mathbb{E}\left[d(X,Y)^p\right])^{1/p}$ over all couplings (X, Y) of (μ, ν) (see Section 2.2 for a formal definition). Diaconis and Saloff Coste [4] and Pak and Sidenko [10] use the dual characterization of $W_{d,1}$ [13, Remark 6.5] that is especially relevant for simulations:

$$W_{d,1}(\mu, \nu) = \sup \left\{ \int_M f \, d(\mu - \nu) \ : \ f : M \to \mathbb{R} \text{ is } 1 - \text{Lipschitz under } d \right\}. \qquad (1)$$

That is, if one can sample from μ, we can estimate $\int_M f \, d\nu$ for any Lipschitz f up to a $W_{d,1}(\mu, \nu)$ intrinsic bias. One scenario where transportation cost seems to be the ideal

metric is described by Ailon and Chazelle [1]. They start from the fact that one can "reduce the dimension" of a point set $S \subset \mathbb{R}^n$ by first applying a random linear transformation X drawn from Haar measure on $SO(n)$ and then projecting onto the first k coordinates. A result known as the Johnson Lindenstrauss lemma says that if one chooses $k = O(\ln |S|/\epsilon^2)$ (which does not depend on the ambient dimension n), with high probability the ratios of pairwise distances in S are all preserved up to $(1 \pm \epsilon)$-factors. One can easily check that a similar result holds when X is $W_{d,1}$-close to being Haar distributed (for an appropriate metric d; see below). In that case, bounds for the convergence of Kac's random walk imply good approximations to random projections. Moreover, as noted in [1], for $X = X(t)$ coming from Kac's random walk the products $s_t = X(t)s$ ($s \in S$) can be computed with just a constant amount of extra memory, as the map $s_t \mapsto s_{t+1}$ affects only two coordinates of s_t. This is a very important feature for large datasets.

Our main result is a bound on the convergence to equilibrium of Kac's random walk on matrices. Our setting corresponds to metric spaces (M, d) with $M = SO(n)$ and three choices of metric d. For $a, b \in SO(n)$ we define:

$$\begin{aligned}
\rho(a,b) &\equiv \sup_{\psi \in \mathbb{R}^n,\, |\psi|=1} |(a-b)\psi|, \text{ with } |\cdot| = \text{Euclidean norm.;} \\
\mathrm{hs}(a,b) &\equiv \|a-b\|_{\mathrm{hs}} = \sqrt{\mathrm{Tr}((a-b)^\dagger (a-b))}, \text{ the Hilbert-Schmidt norm;} \\
D(a,b) &\equiv \text{the Riemannian metric on } SO(n) \text{ induced by the Hilbert-Schmidt} \\
&\quad \text{inner product } \langle u, v\rangle_{\mathrm{hs}} \equiv \mathrm{Tr}(u^\dagger v).
\end{aligned}$$

Clearly $\rho \le \mathrm{hs} \le D$ always. Define the *p-transportation-cost mixing times*:

$$\tau_{d,p}(\epsilon) \equiv \inf\{t \in \mathbb{N} : W_{SO(n),d,p}(\mu P^t, \mathcal{H}) \le \epsilon \text{ for all prob. measures } \mu \text{ on } SO(n)\},$$

where $d = D$, hs or ρ; $\mathcal{H}$ is Haar measure on $SO(n)$; and μP^t is the time-t distribution of a walk started from distribution μ. Note that $\tau_{\rho,p}(\cdot) \le \tau_{\mathrm{hs},p}(\cdot) \le \tau_{D,p}(\cdot)$ (it is also the case that $\tau_{d,p}$ is increasing in p). We will show that:

Theorem 1 *For $d = D$,* hs *or ρ and $1 \le p \le 2$, Kac's random walk on matrices satisfies the following mixing time estimate:*

$$\tau_{d,p}(\epsilon) \le \left\lceil n^2 \ln\left(\frac{4\sqrt{2}\pi n^2}{\epsilon}\right)\right\rceil.$$

Thus $O\left(n^2 \ln n\right)$ steps of the Markov chain suffice to bring μP^t ϵ-close to Haar measure $\mathcal{H}$ for any $\epsilon = n^{-O(1)}$. This improves on a $O\left(n^4 \ln n\right)$ by Diaconis and Saloff Coste and a very recent preprint by Pak and Sidenko [10] that lowered the estimate to $O\left(n^{2.5} \ln n\right)$ steps (we only learned about that result after proving the main results in the present paper). As

noted by the latter authors, $\Omega(n^2)$ steps are necessary for $p = 1$-convergence under the Hilbert-Schmidt distance hs (which they call the Frobenious norm). This implies that the dependence on n in our result optimal up to a logarithmic factor. We note in passing that the $d = \rho$ bound already suffices for dimensionality reduction.

The key to proving Theorem 1 is a contraction property of the Markov transition kernel of the random walk under consideration. Fix again a metric space (M, d). For $0 \leq c < 1$, say that a Markov transition kernel P on M is *c-contracting for the $W_{d,p}$ metric* if for all probability measures $\mu, \nu \in \Pr_{d,p}(M)$:

$$W_{d,p}(\mu P, \nu P) \leq c\, W_{d,p}(\mu, \nu). \tag{2}$$

We will prove the following estimate:

Lemma 1 *In the same setting as Theorem 1, Kac's random walk on matrices is*

$$\sqrt{1 - \frac{1}{\binom{n}{2}}}\text{-}\textit{contracting}$$

in the $W_{D,p}$ metric for any $1 \leq p \leq 2$.

The proof of Lemma 1 follows a strategy related to the *path coupling method* for discrete Markov chains introduced by Bubley and Dyer [2]. Suppose P is now a Markov chain on the set of vertices V of a connected graph G. The graph induces a natural shortest-path metric d on G. It is sometimes possible to prove a "local contraction" estimate of the following form: for any $x, y \in V$ that are adjacent in G, there is a coupling of X (distributed according to one step of P from x) and Y (distributed according to one step of P from y) such that:

$$\mathbb{E}\left[d(X, Y)\right] \leq c = c\, d(x, y) < 1.$$

If that is the case, Bubley and Dyer proved that the local couplings extend to "globally contracting" couplings for all random pairs $(x, y) = (X_0, Y_0) \in V^2$, with:

$$\mathbb{E}\left[d(X, Y)\right] \leq c\, \mathbb{E}\left[d(X_0, Y_0)\right].$$

This implies in particular that $W_{d,1}(\mu P^t, \nu P^t) \leq \operatorname{diam}(G) c^t$ for all distributions μ, ν, where $\operatorname{diam}(G)$ is the diameter of the graph G. In the discrete setting such results easily extend to total variation bounds.

Our adaptation of their technique is based on the fact that $SO(n)$ is a *geodesic space* with the metric D: that is, $D(a, b)$ is the length of the shortest curve connecting a and b. We will show that whenever (M, d) is a geodesic space (or more generally a *length space*; see Section 2.1) and P is such that, for all deterministic $x, y \in M$ with $d(x, y) \ll 1$

$$\mathbb{E}\left[d(X, Y)^p\right] \leq (c + o\,(1)) d(x, y)^p,$$

then P is c-contractive and $W_{d,p}(\mu P^t, \eta P^t) \leq c^t \mathrm{diam}(M)$ for all $\mu, \eta \in \mathrm{Pr}_{d,p}(M)$, where $\mathrm{diam}(M)$ is the diameter of M. That is, we show that if (M, d) is a Polish length space and P satisfies some reasonable assumptions, one can always extend "local contracting couplings" of random walks started from deterministic states near each other to "globally contracting couplings" for arbitrary initial distributions. This result is stated as Theorem 2 below.

We note that proving local contraction is the problem-specific part of our technique. In our case we can use the *local geometry* of $SO(n)$ as a Riemmann manifold to do calculations in the *tangent space*, which greatly simplifies our proof. Pak and Sidenko [10] use a similar coupling construction, but neither do they use the local structure of $SO(n)$ as effectively, nor do they state any general result on local-to-global couplings. On the other hand, Diaconis and Saloff Coste [4] use the analytic technique known as the comparison method which seems intrinsically sub-optimal as well as more complex to apply.

We should also point out that our general coupling idea for continuous-state-space Markov chains has appeared in other works. In particular, while this paper was being prepared Ollivier released a preprint containing a result very similar to our Theorem 2 in his study of positive Ricci curvature for Markov chains on metric spaces [9, Proposition 17] (albeit with a less detailed proof). In fact, what he calls "positive Ricci curvature" is precisely what we call c-contractivity above; from that one can deduce many properties, such as concentration for the stationary distribution and some log-Sobolev-like inequalities (see [9] for details and other references where contractivity of the Markov chain has been used recently). We present our own version of the coupling result both to make the paper self-contained and in order to provide a fully detailed proof (see Remark 1). This is especially important since there have been several important recent results involving analytic, geometric and probabilistic applications of transportation cost [13, 7, 11, 12] and we believe that our technique might be applicable to that field. One sample application is discussed in the last section.

The remainder of the paper is as follows. Section 2 reviews some important concepts from probability, metric geometry and optimal transport. Section 3 proves our general result on local-to-global couplings, Theorem 2. We formally define Kac's random walk on matrices and then prove Lemma 1 and Theorem 1 in Section 4. Section 5 sketches proofs of mixing time estimates for other random walks on matrices, discusses some geometric applications of our results and presents some open problems.

2 Preliminaries

2.1 Metric spaces, length spaces, σ-fields

Whenever we discuss metric spaces (M, d), saying that $A \subset M$ is *measurable* will mean that A belongs the σ-field generated by open sets in M, i.e. the Borel σ-field $\mathcal{B}(M)$. Moreover,

all measures on metric spaces will be implicitly defined over Borel sets. We will always assume that the metric spaces under consideration are *Polish*, i.e. complete and separable.

Following [13, page 123], we say that metric space (M, d) is a *length space* if for all $x, y \in M$ and every $\epsilon > 0$ there exists an ϵ-approximate midpoint $z \in M$ with $|d(x, z) - d(x, y)/2| \leq \epsilon$ and $|d(y, z) - d(x, y)/2| \leq \epsilon$. All complete Riemannian manifolds and their Gromov-Haussdorf limits are length spaces. Non-locally-compact examples of Polish length spaces include separable Hilbert spaces.

2.2 Distributions, couplings and mass transportation

All facts stated below can be found in [13, Chapter 6].

Let (M, d) be a metric space and $\Pr(M)$ be the space of probability measures on (the Borel σ-field of) M. Given $\mu, \nu \in \Pr(M)$, a measure $\nu \in \Pr(M \times M)$ (with the product Borel σ-field) is a *coupling* of (μ, ν) if for all Borel-measurable $A \subset M$:

$$\eta(A \times M) = \mu(A),\ \eta(M \times A) = \nu(A).$$

The set of couplings of (μ, ν) is denoted by $\mathrm{Cp}(\mu, \nu)$. This is always a non-empty set since the product measure $\mu \times \nu$ is in it.

Given $p > 0$, $\Pr_{d,p}(M) \subset \Pr(M)$ is the set of all probability measures μ such that for some (and hence all) $o \in M$

$$\int_M d(o, x)^p \, d\mu(x) < +\infty.$$

One can define the *p-transportation cost (or p-Wasserstein) metric* $W_{d,p}$ on $\Pr_{d,p}(M)$ by the formula:

$$W_{d,p}(\mu, \nu)^p \equiv \inf \left\{ \int_{M \times M} d(x, y)^p \, d\eta(x, y) \ : \ \eta \in \mathrm{Cp}(\mu, \nu) \right\}, \quad \mu, \nu \in \Pr_{d,p}(M). \tag{3}$$

Such metrics are related to the "mass transportation problem" where one attempts to minimize the average distance traveled by grains of sand when a sandpile is moved from one configuration to another.

It is known that $(\Pr_{d,p}(M), W_{d,p})$ is Polish iff (M, d) is Polish. If (M, d) is Polish, the infimum above is always achieved by some $\eta = \eta^{\mathrm{opt}}(\mu, \nu)$, which we will refer to as a *p-optimal coupling* of μ and ν.

For $x \in M$, $\delta_x \in \Pr(M)$ is the *point mass at* x, the distribution that assigns measure 1 to the set $\{x\}$. A basic property of mass transportation is that if $x, y \in M$, then $W_{d,p}(\delta_x, \delta_y) = d(x, y)$.

It is often convenient to deal with random variables rather than measures. If X is a M-valued random variable, $\mathcal{L}_X \in \Pr(M)$ is the distribution (or law) of X. Notice that $\mathcal{L}_X \in \Pr_{d,p}(M)$ iff $\mathbb{E}\,[d(o, X)^p] < +\infty$ for some (hence for all) $o \in M$. We will

write $X =_d \mu$ whenever X is a random variable with $\mathcal{L}_X = \mu$ and $X =_d Y$ if X, Y are random variables with $\mathcal{L}_X = \mathcal{L}_Y$. Call a random pair (X, Y) a *coupling* of (μ, ν) if $\mathcal{L}_{(X,Y)} \in \mathrm{Cp}(\mu, \nu)$. $W_{d,p}(\mu, \nu)$ can be equivalently viewed as the infimum of $\mathbb{E}\left[d(X, Y)^p\right]^{1/p}$ over all such couplings.

Finally, we note that if M is compact (as it is in our application) then for any $p \geq 1$ $\mathrm{Pr}_{d,p}(M) = \mathrm{Pr}(M)$ and $W_{d,p}$ metrizes weak convergence.

2.3 Markov transition kernels

In this section we assume (M, d) is Polish. A *Markov transition kernel* on M is a map $P : M \times \mathcal{B}(M) \to [0, 1]$ such that for all $x \in M$ $P_x(\cdot) \equiv P(x, \cdot)$ is a probability measure and for all $A \in \mathcal{B}(M)$ $P_x(A)$ is a measurable function of x. A Markov transition kernel defines a M-valued Markov chain; for each $\mu \in \mathrm{Pr}(M)$, there exists a unique distribution on sequences of random variables $\{X(t)\}_{t=0}^{+\infty}$ such that $X(0) =_d \mu$ and for all $t \in \{1, 2, 3, \dots\}$, the distribution of $X(t)$ conditioned on $\{X(s)\}_{s=0}^{t-1}$ is $P_{X(t-1)}$.

For $\mu \in \mathrm{Pr}(M)$ and $t \in \mathbb{N}$, μP^t is the measure of $X(t)$ defined as above; one can check that $\mu P^{t+1} = (\mu P^t)P$ for all $t \geq 0$.

3 From Local to Global Couplings

In this section we will discuss our method for moving from local to global bounds on the contraction/expansion properties of a Markov kernel. In our application we have a Markov kernel P on a Polish space (M, d). Using explicit couplings, we will show that for some $C > 0$ and all $x, y \in M$, $W_{d,p}(P_x, P_y) \leq (C + o(1))d(x, y)$, where $o(1) \to 0$ when $y \to x$. The main result in this section implies that under some natural conditions, it follows that $W_{d,p}(\mu P, \nu P) \leq Cr$ whenever $\mu, \nu \in \mathrm{Pr}_{d,p}(M)$ are r-close.

Theorem 2 (Local-to-Global Coupling) *Suppose (M, d) is a Polish length space, $p \geq 1$ is given and P is a Markov transition kernel on (M, d) satisfying the following characteristics.*

1. P_x has finite p-th moments for all x: *that is, $P_x \in \mathrm{Pr}_{d,p}(M)$ for all for all $x \in M$;*
2. Markov steps according to P have uniformly bounded p-th moments: *there exists some $\Delta > 0$ such that for all $x \in M$ $W_{d,p}(\delta_x, P_x) \leq \Delta$, or equivalently*

$$\forall x \in M, \int_M d(x, x')^p \, dP_x(x') \leq \Delta^p.$$

3. P is locally C-Lipschitz on M. *That is, the map*

$$x \mapsto P_x$$

from (M,d) *to* $(\mathrm{Pr}_{d,p}(M), W_{d,p})$ *is* locally C-Lipschitz *in the following sense: for all bounded* $S \subset M$ *there exists a map* $\alpha_S : [0,+\infty) \to [0,+\infty]$ *such that* $\lim_{r\to 0} \alpha_S(r) = 0$ *and*

$$W_{d,p}(P_x, P_y) \leq (C + \alpha(d(x,y)))\, d(x,y).$$

Then for all $\mu \in \mathrm{Pr}_{d,p}(M)$ *we also have* $\mu P \in \mathrm{Pr}_{d,p}(M)$ *and moreover the map* $\mu \mapsto \mu P$ *is* C*-Lipschitz, that is:*

$$\forall \mu, \nu \in \mathrm{Pr}_{d,p}(M),\ W_{d,p}(\mu P, \nu P) \leq C W_{d,p}(\mu, \nu).$$

Before we prove this result we discuss its application to the setting where $C = (1-\kappa)$ for some $\kappa > 0$ and the other assumptions are satisfied. First we recall the well-known fact that a $(1-\kappa)$-Lipschitz map between complete metric spaces has a unique fixed point. Since (M,d) and $(\mathrm{Pr}_{d,p}(M), W_{d,p})$ are Polish, this immediately implies that there exists a unique element $\pi \in \mathrm{Pr}_{d,p}(M)$ with $\pi P = \pi$.

Secondly, the $(1-\kappa)$-Lipschitz property also implies that (M,d) has finite diameter. This was noted by Ollivier [9, Lemma 2.1] and we reproduce his argument here: for all $x, y \in M$ we have on the one hand

$$W_{d,p}(P_x, P_y) = W_{d,p}(\delta_x P, \delta_y P) \leq (1-\kappa) W_{d,p}(\delta_x, \delta_y) = (1-\kappa) d(x,y)$$

and on the other

$$W_{d,p}(P_x, P_y) \geq W_{d,p}(\delta_x, \delta_y) - 2 \sup_{z \in M} W_{d,p}(\delta_z, P_z) \geq d(x,y) - 2\Delta.$$

Hence

$$\forall x, y \in M,\ d(x,y) \leq \frac{2\Delta}{\kappa}.$$

In particular, $\mathrm{Pr}(M) = \mathrm{Pr}_{d,p}(M)$. It follows that π is the unique P-invariant distribution on M. Moreover for all $t \in \mathbb{N}$ and $\mu \in \mathrm{Pr}(M)$,

$$W_{d,p}(\mu P^t, \pi) = W_{d,p}(\mu P^t, \pi P^t) \leq (1-\kappa)^t W_{d,p}(\mu, \pi) \leq (\mathrm{diam}_d(M))\, e^{-\kappa t},$$

where $\mathrm{diam}_d(M) \leq 2\Delta/\kappa$ is the diameter of (M,d).

We collect those facts in the following corollary.

Corollary 1 *Assume* (M,d) *and* P *satisfy the assumptions of Theorem 2 for some* $p \geq 1$ *and* $C = (1-\kappa) < 1$ *(i.e.* $\kappa > 0$*). Then the diameter of* (M,d) *is at most* $2\Delta/\kappa$ *and there exists a unique* P*-invariant measure* π *on* M*. Moreover, the* p-transportation-cost mixing times*:*

$$\tau_{d,p}(\epsilon) \equiv \min\{t \in \mathbb{N} : \forall \mu \in \mathrm{Pr}(M),\ W_{d,p}(\mu P^t, \pi) \leq \epsilon\}$$

satisfy

$$\tau_{d,p}(\epsilon) \leq \left\lceil \kappa^{-1} \ln\left(\frac{\mathrm{diam}_d(M)}{\epsilon}\right)\right\rceil \leq \left\lceil \kappa^{-1} \ln\left(\frac{2\Delta}{\kappa\epsilon}\right)\right\rceil.$$

We now proceed to prove the Theorem.

Proof: [of Theorem 2] First note that the local C-Lipschitz condition can be formulated in a more general setting.

Definition 1 *A map $f : M \to N$ between metric spaces (M, d) and (N, d') is said to be* locally C-Lipschitz *(for some $C > 0$) if for all bounded subsets $S \subset M$ there exists a function $\alpha_S : \mathbb{R}^+ \to \mathbb{R}^+$ such that:*

$$\forall x, y \in S, d'(f(x), f(y)) \leq (C + \alpha_S(d(x, y)))\, d(x, y)$$

and $\lim_{r \searrow 0} \alpha_S(r) = 0$.

Then the following analytic lemma holds.

Lemma 2 *With the notation of Definition 1, assume that M is a length space. Then any $f : M \to N$ that is locally C-Lipschitz is C-Lipschitz according to the standard definition.*

This lemma has a very simple proof in the case where M is e.g. a Riemannian manifold. For in that case we can find a unit-speed geodesic $\gamma : [0, d(x, y)] \to M$ connecting x to y. The local C-Lipschitz condition then implies that $f \circ \gamma$ is Lipschitz and $\|(f \circ \gamma)'\|_\infty \leq C$, so that:

$$|f(y) - f(x)| \leq |f \circ \gamma(d(x, y)) - f \circ \gamma(0)| \leq C\, d(x, y).$$

For general length spaces the proof is only slightly more complicated, but it still uses the intuitive idea that x and y are "close" to being connected by a geodesic; see Section 3.1 for details. For our proof we only need the following consequence.

Corollary 2 *If P is a Markov transition kernel on a length space (M, d) satisfying condition* 3. *of Theorem 2, then $W_{d,p}(P_x, P_y) \leq C\, d(x, y)$ for all $x, y \in M$.*

The bounding of $W_{d,p}(P_x, P_y)$ can be thought of as an implicit construction of a coupling along a geodesic path; this is precisely where the name "path coupling" comes from.

To continue, we simply note that $x \mapsto P_x$ is uniformly continuous. The second lemma (proven in Section 3.2) shows that $\mu P \in \Pr_{d,p}(M)$ whenever $\mu \in \Pr_{d,p}(M)$ and shows that we will only need to compare μP and νP, for μ, ν with countable support.

Lemma 3 *Let (M, d) be Polish. Suppose P is a Markov transition kernel on M such that:*

1. *$P_x \in \Pr_{d,p}(M)$ for all $x \in M$;*
2. *$\sup_{x \in M} W_{d,p}(\delta_x, P_x) \leq \Delta < +\infty$, or equivalently*

$$\sup_{x \in M} \int_M d(x, x')^p \, dP_x(x') \leq \Delta^p < +\infty; \text{ and}$$

3. $x \mapsto P_x$ is a uniformly continuous map from M to $\mathrm{Pr}_{d,p}(M)$.

Then for all $\mu \in \mathrm{Pr}_{d,p}(M)$ we have $\mu P \in \mathrm{Pr}_{d,p}(M)$. Moreover, there exists a sequence $\{\mu_j\}_j \subset \mathrm{Pr}_{d,p}(M)$ of measures with countable support such that $W_{d,p}(\mu_j, \mu) \to 0$ and $W_{d,p}(\mu_j P, \mu P) \to 0$.

The proposition implies the following statement: if $W_{d,p}(\mu P, \nu P) \leq C W_{d,p}(\mu, \nu)$ for all μ, ν in $\mathrm{Pr}_{d,p}(M)$ that have countable support, then the same holds for all μ, ν in $\mathrm{Pr}_{d,p}(M)$. Our final goal is to prove the Lipschitz estimate for measures with countable support.

Thus let $\mu = \sum_{j\in\mathbb{N}} p_j \delta_{x_j}$ be a convex combination of a countable number of point masses ($x_j \in M$ for all j); similarly, let $\nu = \sum_{k\in\mathbb{N}} q_k \delta_{y_k}$. The p-optimal coupling η of μ and ν is of the form

$$\eta = \sum_{j,k\in\mathbb{N}} r_{j,k} \delta_{(x_j, y_k)}$$

for some convex weights $r_{j,k}$. Now define for each pair j, k a p-optimal coupling $\xi_{j,k}$ of P_{x_j}, P_{y_k}. Then

$$\eta' = \sum_{j,k\in\mathbb{N}} r_{j,k} \xi_{j,k} \in \mathrm{Cp}(\mu P, \nu P).$$

Moreover, since $x \mapsto P_x$ is C-Lipschitz,

$$\int_{M\times M} d(u,v)^p \, d\xi_{j,k}(u,v) = W_{d,p}(P_{x_j}, P_{y_k})^p \leq C^p \, d(x_j, y_k)^p,$$

which implies

$$\begin{aligned} W_{d,p}(\mu P, \nu P)^p &\leq \int_{M\times M} d(u,v)^p \, d\eta'(u,v) \\ &= \sum_{j,k\in\mathbb{N}} r_{j,k} \int_{M\times M} d(u,v)^p \, d\xi_{j,k}(u,v) \\ &\leq C^p \sum_{j,k\in\mathbb{N}} r_{j,k} d(x_j, y_k)^p \\ &= C^p \int_{M\times M} d(u,v)^p \, d\eta(u,v). \end{aligned}$$

The RHS is simply $C^p W_{d,p}(\mu, \nu)^p$. □

Remark 1 *As pointed out in the Introduction, Ollivier stated a very similar result in [9, Proposition 17]. In his proof he assumes that there exists a Markov transition kernel ξ on M^2 such that for all $(x, y) \in M^2$, $\xi_{(x,y)}$ is a 1-optimal coupling of (P_x, P_y). This greatly simplifies his argument, but we could not find there or elsewhere a detailed proof of the existence of ξ. At any rate, our own argument provides an alternative approach to the same result and makes our paper self-contained.*

3.1 Proof of Lemma 2

Proof: Fix $x, y \in M$, $S = B_{(M,d)}(x, d(x,y)+1)$ and α_S be as in Definition 1 (we may assume it to be increasing wlog). We will show that

$$\forall j \in \mathbb{N}\, \forall \delta \in (0,1),\ d'(f(x), f(y)) \leq \left(C + \alpha_S \left(\frac{d(x,y)+\delta}{2^j} \right) \right) (d(x,y)+\delta). \tag{4}$$

which implies via $j \to +\infty$ and $\delta \to 0$ that

$$d'(f(x), f(y)) \leq C d(x,y).$$

Iterating the definition of length space implies that for any $\delta \in (0,1)$ and $j \in \mathbb{N}$ one can find $z_0 = x, z_1, \ldots, z_{2^j-1}, z_{2^j} = y$, all in M, with

$$\forall i \in \{1, 2, \ldots, 2^j\},\ \left| d(z_i, z_{i-1}) - \frac{d(x,y)}{2^j} \right| \leq \frac{\delta}{2^j}.$$

We then have

$$d'(f(x), f(y)) \leq \sum_{i=1}^{2^j} d'(f(z_i), f(z_{i-1})) \leq \sum_{i=1}^{2^j} (C + \alpha_S(d(z_i, z_{i-1})))d(z_{i-1}, z_i).$$

For the last inequality we used the fact that f is locally C-Lipschitz and that for each i

$$d(z_i, x) \leq \sum_{\ell=1}^{2^j} d(z_\ell, z_{\ell-1}) \leq d(x,y) + \delta \leq d(x,y) + 1 \Rightarrow z_i \in S.$$

To deduce (4) we simply note that for each i

$$d(z_i, z_{i-1}) \leq \frac{d(x,y)+\delta}{2^j}$$

and sum the corresponding bounds. $\square$

3.2 Proof of Lemma 3

Proof: Let $\Delta > 0$ be as in the assumptions. For the first statement we note that, for any $\mu \in \Pr_{d,p}(M)$ and any $o \in M$,

$$\begin{aligned}
\int_M d(o,x')^p \, d\mu P(x') &= \int_M \left(\int_M d(o,x')^p \, dP_x(x') \right) d\mu(x) && (5)\\
(\text{use } d(o,x') \le d(o,x) + d(x,x')) &= \int_M \left(\int_M [d(o,x) + d(x,x')]^p \, dP_x(x') \right) && (6)\\
(\text{use } |a+b|^p \le 2^p(|a|^p + |b|^p)) &\le 2^p \int_M \int_M d(o,x)^p \, dP_x(x') \, d\mu(x) && (7)\\
&\quad + 2^p \int_M \int_M d(x,x')^p \, dP_x(x') \, d\mu(x) && (8)\\
&\le 2^p \int_M d(o,x)^p \, d\mu(x) + (2\Delta)^p < +\infty. && (9)
\end{aligned}$$

Thus μP is in $\Pr_{d,p}(M)$ whenever μ is.

We now present a discrete approximation scheme for μ. Since M is separable, there exists a sequence of partitions $\{\mathcal{P}_j\}_{j \in \mathbb{N}}$ of M such that:

- each partition contains countably many measurable sets;
- for all $j \in \mathbb{N}$, $\mathcal{P}_{j+1}$ refines $\mathcal{P}_j$; and
- for all $j \in \mathbb{N}$ the sets in $\mathcal{P}_j$ have diameter at most ϵ_j for some sequence $\epsilon_j \to 0$.

Let us also assume that for each $j \in \mathbb{N}$ and $A \in \mathcal{P}_j$ we have picked some $x_A^{(j)} \in A$. Consider the measures:

$$\mu_j \equiv \sum_{A \in \mathcal{P}_j} \mu(A) \delta_{x_A^{(j)}}. \tag{10}$$

Clearly $\mu_j \in \Pr_{d,p}(M)$ for all j and $W_{d,p}(\mu_j, \mu) \to 0$ when $\mu_j \to 0$. Our goal will be to show that $W_{d,p}(\mu_j P, \mu P) \to 0$. First recall that $x \mapsto P_x$ is uniformly continuous, hence there exists a sequence $\delta_j \to 0$ such that for all $j \in \mathbb{N}$, all $A \in \mathcal{P}_j$ and all $x \in A$:

$$W_{d,p}(P_{x_A^{(j)}}, P_x) \le \delta_j.$$

We will use this to show that

$$\forall j < k, \; W_{d,p}(\mu_j, \mu_k) \le \delta_j \text{ (in particular, } \{\mu_j\}_j \text{ is Cauchy)}. \tag{11}$$

Recall that if $j < k$ $\mathcal{P}_k$ is a refinement of $\mathcal{P}_j$, hence for all $B \in \mathcal{P}_k$ there exists a set $A_B \in \mathcal{P}_j$ with $B \subset A_B$. For each such B, we have $x_B^{(k)} \in A_B$, hence there exists a coupling $\eta_{B,k,j}$ of

$P_{x_B^{(k)}}$ and $P_{x_{A_B}^{(j)}}$ with

$$\int_{M\times M} d(u,v)^p \, d\eta_{B,k,j}(u,v) \le \delta_j^p.$$

Extend this to a coupling of $\mu_k P$ and $\mu_j P$ by:

$$\eta_{k,j} \equiv \sum_{B\in\mathcal{P}_k} \mu(B)\, \eta_{B,k,j}.$$

To prove that $\eta_{k,j} \in \mathrm{Cp}(\mu_j P, \mu_k P)$, notice that the first marginal of this measure is

$$\sum_{B\in\mathcal{P}_k} \mu(B) P_{x_B^{(k)}} = \mu_k P.$$

Moreover, for any $A \in \mathcal{P}_j$, the set of all $B \in \mathcal{P}_k$ with $A_B = A$ is a partition of A, hence the second marginal is also right:

$$\sum_{B\in\mathcal{P}_k} \mu(B) P_{x_A^{(k)}} = \sum_{A\in\mathcal{P}_j} \left(\sum_{B\in\mathcal{P}_k \,:\, A_B = A} \mu(B) \right) P_{x_A^{(j)}} = \sum_{A\in\mathcal{P}_j} \mu(A) P_{x_A^{(j)}} = \mu_j P.$$

It follows that $\eta_{k,j} \in \mathrm{Cp}(\mu_j P, \mu_k P)$ and moreover one can check that

$$\int_{M\times M} d(u,v)^p \, d\eta_{k,j}(u,v) \le \delta_j^p,$$

which implies (11).

$(\mathrm{Pr}_{d,p}(M), W_{d,p})$ is Polish since (M,d) is. By the above, we know that there exists a measure $\xi \in \mathrm{Pr}_{d,p}(M)$ such that $W_{d,p}(\mu_j P, \xi) \le \delta_j$. This also implies [13, Theorem 6.8] that $\mu_j P \Rightarrow \xi$ in the weak topology. However, it is an exercise to show that $\mu_j P \Rightarrow \mu P$ weakly, hence $\xi = \mu P$ and $W_{d,p}(\mu_j P, \mu P) \to 0$, as desired. □

4 Analysis of Kac's random walk

4.1 Definitions

Let $M(n,\mathbb{R})$ be the set of all $n \times n$ matrices with complex entries. These are the linear operators from $\mathbb{R}^n$ to itself and we equip $\mathbb{R}^n$ with a canonical basis $e_1, \ldots, e_n$ of orthonormal vectors. For $a \in M(n,\mathbb{R})$, $a^\dagger$ is the transpose of a. Using it, one can define the Hilbert-Schmidt inner product $\langle a, b\rangle_{\mathrm{hs}} \equiv \mathrm{Tr}(a^\dagger b)$ on $M(n,\mathbb{R})$, under which it is isomorphic to $\mathbb{R}^{n^2}$ with the standard Euclidean inner product. We let $\|\cdot\|_{\mathrm{hs}}$ be the corresponding norm.

An element $a \in M(n,\mathbb{R})$ is *orthogonal* if $a^\dagger a = aa^\dagger = \mathrm{id}$, the identity matrix. The subset of $M(n,\mathbb{R})$ given by:

$$SO(n) \equiv \{a \in M(n,\mathbb{R}) \,:\, aa^\dagger = \mathrm{id}, \det(a) = 1\}$$

is a compact, connected submanifold of $M(n,\mathbb{R})$. It is also a Lie group since it is closed under matrix multiplication and matrix inverse. Therefore $SO(n)$ has a Haar measure $\mathcal{H}$, which we may define as the unique probability measure on that group such that for all measurable $S \subset SO(n)$ and $a \in SO(n)$ we have $\mathcal{H}(S) = \mathcal{H}(Sa) = \mathcal{H}(aS)$ where $Sa = \{sa \,:\, s \in S\}$ and $aS = \{as \,:\, s \in S\}$.

We now define *Kac's random walk on matrices.* For $1 \le i < j \le n$ and $\theta \in [0, 2\pi]$ define $R(i,j,\theta)$ as a rotation by θ of the plane generated by e_i, e_j. This is equivalent to setting:

$$R(i,j,\theta)e_k = \begin{cases} \cos\theta e_i + \sin\theta e_j, & k = i; \\ \cos\theta e_j - \sin\theta e_i, & k = j; \\ e_k & k \in \{1,\ldots,n\}\backslash\{i,j\} \end{cases} \tag{12}$$

and extending $R(i,j,\theta)$ to all $\psi \in \mathbb{R}^n$ by linearity. Kac's random walk on matrices corresponds to the following Markov transition kernel:

$$P_x(S) \equiv \frac{1}{2\pi\binom{n}{2}} \sum_{1 \le i < j \le n} \int_0^{2\pi} \delta_{R(i,j,\theta)x}(S)\, d\theta \;\; (x \in SO(n), S \subset SO(n) \text{ measurable}).$$

Thus to generate $X =_d P_x$ from x one chooses $1 \le i < j \le n$ uniformly at random from all $\binom{n}{2}$ possible choices, pick and $\theta \in [0, 2\pi]$ also uniformly at random and then sets $X = R(i,j,\theta)x$. The required measurability conditions are easily established. One can also check that Haar measure $\mathcal{H}$ is P-invariant.

4.2 The geometry of $SO(n)$

We collect some standard facts that will be used in our proofs.

The tangent space at the identity matrix id is the set of all anti-self-adjoint operators:

$$T \equiv T_{\mathrm{id}} SO(n) = \{h \in M(n,\mathbb{R}) \,:\, h^\dagger = -h\}. \tag{13}$$

We let D be the Riemannian metric induced on $SO(n)$ by $\langle \cdot, \cdot \rangle_{\mathrm{hs}}$. Since $SO(n)$ is compact, one can show the following.

$$\forall z, w \in SO(n),\; \|z - w\|_{\mathrm{hs}} \le D(z,w) \le \|z - w\|_{\mathrm{hs}} + O\left(\|z - w\|_{\mathrm{hs}}^2\right), \tag{14}$$

where $O(r^\alpha)$ is just some term whose absolute value is uniformly bounded by cr^α, $c > 0$ a constant not depending on r (we will use this notation from now on). Moreover, if we let

Π_T be the orthogonal projector onto T (according to the Hilbert-Schmidt inner product), then

$$\forall z \in SO(n),\ \|z - \mathrm{id} - \Pi_T(z - \mathrm{id})\|_{\mathrm{hs}} \leq O\left(D(z, \mathrm{id})^2\right). \tag{15}$$

This is so because if $\|z - \mathrm{id}\| = r \ll 1$, then $\|z - \mathrm{id} - \tilde{h}\|_{\mathrm{hs}} = O\left(r^2\right)$ for some $\tilde{h} \in T$, and $\tilde{h} = h = \Pi_T(z - \mathrm{id})$ is the best choice of approximation one may make. Notice that the two equations together imply:

$$|D(z, \mathrm{id}) - \|\Pi_T(z - \mathrm{id})\|_{\mathrm{hs}}| = O\left(\|\Pi_T(z - \mathrm{id})\|_{\mathrm{hs}}^2\right). \tag{16}$$

One can define another distance on $M(n, \mathbb{R})$ via:

$$\rho(a, b) \equiv \sup\{|(a - b)\psi| \,:\, \psi \in \mathbb{R}^n,\ |\psi| = 1\},\ \text{where } |\cdot| \text{ is the Euclidean norm.}$$

Then for all $a, b \in SO(n)$:

$$\rho(a, b) \leq \|a - b\|_{\mathrm{hs}} \leq D(a, b).$$

We notice that these distances are all invariant under multiplication: if $a, b, c \in SO(n)$,

$$\rho(ca, cb) = \rho(ac, bc) = \rho(a, b)$$

and similarly for $\mathrm{hs}(a, b) = \|a - b\|_{\mathrm{hs}}$ and $D(a, b)$.

4.3 The contraction coefficient

In this section we prove Lemma 1.

Proof: Consider $x, y \in SO(n)$ and let $D(x, y) = r$. We will show that there exists a coupling (X, Y) of (P_x, P_y) with

$$\mathbb{E}\left[D(X, Y)^2\right] \leq (1 - \delta)\, r^2 + O\left(r^3\right)$$

for $\delta = 1/\binom{n}{2}$. As in the previous section, $O\left(r^3\right)$ is some term that is uniformly bounded by a multiple of r^3. The existence of such a coupling implies that

$$W_{D,2}(P_x, P_y) \leq (1 - \delta) D(x, y)^2 + O\left(D(x, y)^3\right),$$

which shows that P is locally $\sqrt{(1 - \delta)}$-Lipschitz for $p = 2$.

Our coupling will be as follows. Suppose we set $X = R(i, j, \theta)x$ with i, j, θ randomly picked as prescribed in the definition of the random walk. We will set $Y = R(i, j, \theta')$ with the same i, j and some $\theta' = (\theta - \alpha) \mod 2\pi$, where $\alpha = \alpha(i, j, x, y)$ depends on i, j, x, y but *not* on θ. In that case θ' is uniform on $[0, 2\pi]$ independently of i, j, x, y, hence (X, Y) is a valid coupling of (P_x, P_y). Also notice that, using the invariance of D under multiplication,

$$D(X, Y) = D(R(i, j, \theta)x, R(i, j, \theta')y) = D(R(i, j, \theta), R(i, j, \theta')yx^\dagger) = D(R(i, j, \alpha), yx^\dagger), \tag{17}$$

as

$$R(i,j,\theta')^\dagger R(i,j,\theta) = R(i,j,\theta-\theta') = R(i,j,\alpha).$$

We will use (14), (15) and (16) to bound the RHS of (17): this will allow us to do all calculations we need in the tangent space $T = T_{\mathrm{id}}SO(n)$. First, however, we need an orthonormal basis for that space. For each $1 \le k < \ell \le n$ let $a_{k\ell} \in T$ be the operator that is uniquely defined by:

$$a_{k\ell}\, e_t \equiv \begin{cases} \frac{e_\ell}{\sqrt{2}}, & t = k; \\ -\frac{e_k}{\sqrt{2}}, & t = \ell; \\ 0, & t \in \{1,\ldots,n\}\backslash\{k,\ell\}. \end{cases}.$$

One can check that $\{a_{k\ell}\}_{1\le k<\ell\le n}$ is indeed an orthonormal basis for $T = T_{\mathrm{id}}SO(n)$ with the Hilbert Schmidt inner product. For $1 \le t \le n$ we also define $d_t \in M(n,\mathbb{R})$ as the matrix that has a 1 at the (t,t)-th entry and zeroes elsewhere. Then $\langle d_t, d_s\rangle_{\mathrm{hs}} = 1$ if $t = s$ and 0 otherwise and also $\langle d_t, a_{k\ell}\rangle_{\mathrm{hs}} = 0$ for any t,k,ℓ. With these definitions one can write:

$$R(i,j,\alpha) = \mathrm{id} + (\cos\alpha - 1)d_i + (\cos\alpha - 1)d_j + \sqrt{2}\sin\alpha a_{ij}. \tag{18}$$

Now set $h = \Pi_T(yx^\dagger - \mathrm{id})$. Since $D(yx^\dagger, \mathrm{id}) = D(x,y) = r$, $\|h\|_{\mathrm{hs}} = r + O\left(r^2\right)$ and $\|yx^\dagger - \mathrm{id} - h\|_{\mathrm{hs}} = O\left(r^2\right)$. Suppose we commit ourselves to making a choice of $\alpha = O(r)$ (i.e. $|\alpha| \le cr$ for a constant c independent of r). Expanding sin and cos we get:

$$R(i,j,\alpha) = \mathrm{id} + \sqrt{2}\alpha a_{ij} + O\left(r^2\right).$$

Moreover, we also have

$$\begin{aligned} D(yx^\dagger, R(i,j,\alpha)) &= \|yx^\dagger - R(i,j,\alpha)\|_{\mathrm{hs}} + O\left(\|yx^\dagger - R(i,j,\alpha)\|_{\mathrm{hs}}^2\right) && (19)\\ &= \|yx^\dagger - \mathrm{id} - \sqrt{2}\,\alpha a_{ij}\|_{\mathrm{hs}} + O\left(\|yx^\dagger - \mathrm{id} - \sqrt{2}\,\alpha a_{ij}\|_{\mathrm{hs}}^2 + r^2\right) && (20)\\ &= \|h - \sqrt{2}\,\alpha a_{ij}\|_{\mathrm{hs}} + O\left(r^2\right). && (21)\end{aligned}$$

Thus we choose $\alpha = \langle h, a_{ij}\rangle_{\mathrm{hs}}/\sqrt{2}$, which minimizes $\|h - \sqrt{2}\,\alpha a_{ij}\|_{\mathrm{hs}}$ and only depends on i,j and $h = \Pi_T(yx^\dagger - \mathrm{id})$. Since the $a_{k\ell}$ form an orthonormal basis of $T \ni h$, we have

$$h = \sum_{1\le k<\ell\le n} \langle h, a_{k\ell}\rangle_{\mathrm{hs}} a_{k\ell} \Rightarrow \sum_{1\le k<\ell\le n} \langle h, a_{k\ell}\rangle_{\mathrm{hs}}^2 = \|h\|_{\mathrm{hs}}^2 = r^2 + O\left(r^3\right).$$

This shows that $|\alpha| = O(r)$ as desired and moreover

$$\begin{aligned} D(X,Y)^2 &= D(yx^\dagger, R(i,j,\alpha))^2 \quad (\text{by (17)})\\ &= \|h - \langle h, a_{ij}\rangle_{\mathrm{hs}} a_{ij}\|_{\mathrm{hs}}^2 + O\left(r^3\right)\\ (\text{expand } h) &= \|h\|_{\mathrm{hs}}^2 - \langle h, a_{ij}\rangle_{\mathrm{hs}}^2 + O\left(r^3\right).\end{aligned}$$

If we now average over i, j, θ we obtain

$$\begin{aligned}\mathbb{E}\left[D(X,Y)^2\right] &= \|h\|^2 - \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} \langle h, a_{ij}\rangle_{\text{hs}}^2 + O\left(r^3\right) \\ &= \left(1 - \frac{1}{\binom{n}{2}}\right) \|h\|_{\text{hs}}^2 + O\left(r^3\right) \\ &= \left(1 - \frac{1}{\binom{n}{2}}\right) r^2 + O\left(r^3\right),\end{aligned}$$

which is the desired bound.

To finish the proof we apply our result on local-to-global couplings, Theorem 2. We have shown that the Markov transition kernel P for Kac's random walk is locally C-Lipschitz for

$$C = \left(1 - \frac{1}{\binom{n}{2}}\right)^{1/2}, 1 \le p \le 2.$$

The remaining assumptions of Theorem 2 and Corollary 2 are trivially verified since $SO(n)$ has bounded diameter. We conclude that:

$$\forall \mu, \eta \in \Pr(M),\ W_{d,p}(\mu P, \nu P) \le \sqrt{1 - \frac{1}{\binom{n}{2}}}\ W_{d,p}(\mu, \nu).$$

□

4.4 Mixing time estimates

We now prove Theorem 1.

Proof: [of Theorem 1] For the upper bound it suffices to prove the estimate for $\tau_{D,2}(\epsilon)$. We will apply Corollary 2 with $M = SO(n)$, $d = D$ and P the transition kernel for Kac's random walk. According to Lemma 1 we can take $C = \sqrt{1 - 1/\binom{n}{2}} \le (1 - \kappa)$ for $\kappa = 1/n^2$.

We need an estimate for $\Delta = \sup_{x \in SO(n)} W_{d,2}(\delta_x, P_x)$. Let $x \in SO(n)$, $X =_d P_x$. We see that $X = R(i, j, \theta)x$ for $\theta \in [0, 2\pi]$ uniform. The curve

$$\gamma(s) \equiv R(i, j, s\theta)x,\ 0 \le s \le 1$$

connects x to X inside $SO(n)$. It speed is given by

$$\|\gamma'(s)\|_{\text{hs}} = \left\| \frac{d}{ds} R(i, j, s\theta) \right\|_{\text{hs}}.$$

Using formula (18) above one can check that

$$\frac{d}{du}R(i,j,u) = \sin u(d_i + d_j) - \sqrt{2}\cos u a_{ij} \Rightarrow \left\| \frac{d}{du}R(i,j,u) \right\|_{\text{hs}} = \sqrt{2}.$$

Thus

$$D(x,X) \leq \int_0^1 \|\gamma'(s)\|_{\text{hs}}\, ds = \sqrt{2}\theta \leq 2\sqrt{2}\pi.$$

Since $x \in M$ was arbitrary, we see that we may take $\Delta = 2\sqrt{2}\pi$. We deduce from the Corollary that

$$\tau_{D,2}(\epsilon) \leq \left\lceil n^2 \ln\left(\frac{4\sqrt{2}\pi n^2}{\epsilon}\right)\right\rceil.$$

□

5 Final remarks

- The same ideas used in proving Theorem 1 can be employed to analyze a related random walk on the group $U(n)$ of $n \times n$ unitary matrices. In that case the evolution from state $X(t)$ to $X(t+1) = S(i_t, j_t)X(t)$ consists of multiplication by $S(i_t, j_t)$, where S is drawn from Haar measure on $U(2)$, $1 \leq i_t < j_t \leq n$ are uniform and $S(i_t, j_t)$ acts as S within the plane spanned by coordinates i_t, j_t and as identity in the orthogonal complement. The key step in proving such a result is adapting Lemma 1. Let a^* be the conjugate transpose of the matrix a. One can show (using tangent spaces) that if $x, y \in U(n)$ are close, $yx^* = e^h \approx \mathrm{id} + h$ for some anti-Hermitian matrix h with $h^* = -h$. One can couple $X = S(i,j)x$ to $Y = S(i,j)e^{h_{ij}}y$, where h_{ij} is the projection of h onto the space of anti-Hermitian matrices acting on the (i,j)-plane ($S(i,j)e^{h_{ij}}$ has Haar distribution on $U(2)$ because $S(i,j)$ does and $e^{h_{ij}}$ is unitary and independent of $S(i,j)$). The same bounds obtained for Kac's random walk are available in this setting.

- One can also consider a variant of Kac's random walk where θ_t is chosen from a density f on $[0, 2\pi]$ with $\inf_{x\in[0,2\pi]} f(x) = c > 0$. In this case, one can show that the corresponding Markov transition kernel is $(1 - 2\pi c/\binom{n}{2})^{1/2}$-contractive and from that obtain bounds on the mixing time. To prove contractivity, one notices that f can be written as a mixture of densities $f = w(1/2\pi) + (1-w)g$, where $w = 2\pi c$ and g is another density. One can then modify the proof of Lemma 1 and couple $X = R(i, j\theta)x$, $Y = R(i,j,\theta')y$ as follows: with probability w, choose θ uniformly on $[0, 2\pi]$ and $\theta' = (\theta - \alpha) \mod 2\pi$ as in the original proof; with probability $1-w$, sample θ from g and set $\theta' = \theta$. In the first case, one achieves contraction; in the second, $D(X,Y) = D(x,y)$. Putting those facts together implies the desired result.

- The dimensionality reduction application of [1] discussed in the Introduction does not require that Haar measure on $SO(n)$ is well approximated, but only that certain projections behave as they should. It is thus natural to ask whether better bounds are available in that case. More precisely, Kac's random walk induces a random walk on the *Grassmanian* $G(k,n)$, the set of k-dimensional subspaces of $\mathbb{R}^n$. We conjecture that this random walk mixes in $O(nk \ln n)$ time. Recall that for dimension reduction we need $k = O(\ln |S|)$, which might be much smaller than $O\left(n^2 \ln n\right)$; our conjecture would imply great time savings for $n \gg \ln |S|$.

- Theorem 2 on local-to-global coupling can be used to reprove some known results. Consider for instance a Riemannian manifold M with dimension n, distance d and Ricci curvature lower bounded by $K \in \mathbb{R}$. Let $P = P^{(\epsilon)}$ correspond to the ball walk on M where a step from x consists of choosing X uniformly from the ball $B(x,\epsilon)$ (under the natural Riemannian volume). Using [14, Theorem 3] and Theorem 2 one can show that $\mu \mapsto \mu P^{(\epsilon)}$ is $(1 - K/2(n+2) + O(\epsilon))$-Lipschitz (thus contractive when $K > 0$ and ϵ is small enough).

Acknowledgement: We thank Yann Ollivier for useful discussions on Ricci curvature for Markov chains.

References

[1] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC '06: Proceedings of the thirty-eighth annual ACM Symposium on Theory of Computing*, pages 557–563, New York, NY, USA, 2006. ACM Press.

[2] Russell Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in Markov Chains. In *FOCS'97: Proceedings of the 38th annual Symposium on the Foundations of Computer Science*, pages 223–231, 1997.

[3] Eric Carlen, Maria da Conceição Carvalho, and Michael Loss. Determination of the spectral gap for Kac's master equation and related stochastic evolution. *Acta Mathematica*, 191(1):1–54, 2003.

[4] P. Diaconis and L. Saloff-Coste. Bounds for Kac's Master Equation. *Communications in Mathematical Physics*, 209:729–755, 2000.

[5] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

[6] Mark Kac. *Probability and Related Topics in Physical Science*. Interscience, 1959.

[7] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, to appear.

[8] D. Maslin. The eigenvalues of Kac's master equation. *Math. Z.*, 243:291–331, 2003.

[9] Yann Ollivier. Ricci curvature of Markov chains on metric spaces. Arxiv preprint `math/0701886`, 2007.

[10] Igor Pak and Sergyi Sidenko. Convergence of Kac's Random Walk. Preprint available from `http://www-math.mit.edu/~pak/research.html`, 2007.

[11] Karl Theodor Sturm. On the geometry of metric measure spaces. *Acta Mathematica*, 196(1):65–131, 2006.

[12] Karl Theodor Sturm. On the geometry of metric measure spaces II. *Acta Mathematica*, 196(1):133–177, 2006.

[13] Cédric Villani. *Optimal Transport, old and new.* Book draft available from author's website `http://www.umpa.ens-lyon.fr/~cvillani/`.

[14] Max von Renesse and Karl Theodor Sturm. Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on Pure and Applied Mathematics*, 58(7):923–940, 2005.