An introduction to Coupling

Artur O. Lopes

Abstract In this review paper we describe the use of couplings in several different mathematical problems. We consider the total variation norm, maximal coupling and the $\bar{d}$-distance. We present a detailed proof of a result recently proved: the dual of the Ruelle operator is a contraction with respect to 1-Wasserstein distance. We also show the exponential convergence to equilibrium on the state space for finite state Markov chains when the transition matrix $\mathscr{P}$ has all entries positive.

1 Introduction

This is a review paper presenting some examples which were described in the literature where couplings are used for deriving results in Ergodic Theory and Probability. We present several simple calculations which we believe can help the newcomer to the subject. We are writing for a broad audience and not for the expert.

Our purpose is to present such results in a more direct approach. This will avoid the reader which is interested in the topic to have to look in several different references.

We describe the results in a language which is more familiar to the Dynamical Systems audience.

In section 2 we present some definitions and mention some related results which are required in section 7 where we consider the dual of the Ruelle operator.

In sections 3 and 4 we consider the total variation norm and the maximal coupling (see Theorem 1).

In section 6 we consider the d-bar distance among Bernoulli probabilities (see Theorem 3).

A. O. Lopes
Instituto de Matemática-UFRGS, Avenida Bento Gonçalves 9500 Porto Alegre-RS Brazil, e-mail: alopes@mat.ufrgs.br

We also show the exponential convergence to equilibrium on the state space for finite state Markov chains (see Theorem 2) when the transition matrix $\mathscr{P}$ has all entries positive (see sections 5.1 to 5.2).

The use of coupling and the Wasserstein distance can be useful for estimating decay of correlations (see [BFG2] [GP] and [Sulku]) and also in other different dynamical and ergodic problems (see [Kl] and [Aus]).

We believe is important to describe interesting plans or couplings that can be used in estimations of different nature. Variations of these plans can be helpful to solve other open problems.

According to Frank den Hollander [Hol]: *coupling is an art, not a recipe*.

We would like to thanks Anthony Quas, Diogo Gomes, Adriana Neumann and Rafael Souza for helpful conversations during the period we were writing this review paper.

2 Some definitions and the dual of the Ruelle Operator

Definition 1. Given the Bernoulli space $\Omega = \{1,2,...,d\}^{\mathbb{N}}$ each Γ in the set of probabilities on $\Omega \times \Omega$ is called a plan.

Given the Bernoulli space $\Omega = \{1,2,...,d\}^{\mathbb{N}}$ and two probabilities μ and ν on the natural Borel sigma algebra of Ω, a coupling of μ and ν is a plan Γ on the product space $\Omega \times \Omega$, such that, the first marginal of Γ is μ and the second is ν.

$\mathscr{C}(\mu,\nu)$ by definition is the set of plans Γ on $\Omega \times \Omega$ such that the projection in the first coordinate is μ and in the second is ν.

Definition 2. Given a distance d on Ω we denote

$$W_1(\mu,\nu) = \inf_{\Gamma \in \mathscr{C}(\mu,\nu)} \int d(x,y) d\Gamma(dx,dy).$$

The above expression defines a metric on the space of probabilities over Ω which is compatible with the weak$*$-convergence (see [Vi1]). This value is called the 1-Wasserstein distance of μ and ν.

Each plan (it can exist more than one) which realizes the above infimum is called an optimal plan for d. We denote by $d_1(\mu,\nu) = d_1(\mu,\nu)$ the corresponding metric in the set of probabilities on Ω.

Definition 3. More generally, given a continuous function $c : \Omega \times \Omega \to \mathbb{R}$ and fixed μ and ν, one can be interested in

$$\inf_{\Gamma \in \mathscr{C}(\mu,\nu)} \int c(x,y) d\Gamma(dx,dy).$$

Each plan Γ (it can exist more than one) which realizes the above infimum is called an optimal plan, or optimal coupling, for c and the pair μ and ν.

In general is not so easy to identify exactly the optimal plan Γ. Anyway, if we are lucky to find some plan which is almost optimal then we can get some interesting results. In simple words this is the main issue on Coupling Theory.

The Kantorovich duality theorem (see [Vi1]) is a main result which also helps to get good estimates in problems of different nature.

The total variation norm (to be defined later) is a different way to measure the distance among two probabilities. It is not equivalent to weak-$*$-convergence. This will be also considered in the text.

One of our purposes is to illustrate through several examples the use of coupling in interesting problems.

Suppose $\theta < 1$ is fixed. On the Bernoulli space $\Omega = \{1,2,...,d\}^{\mathbb{N}}$ we consider the metric d_θ. By definition $d_\theta(x,y) = \theta^N$ where $x_1 = y_1,..,x_{N-1} = y_{N-1}$ and $x_N \neq y_N$.

We briefly mention some properties related to Gibbs states of Lipchitz potentials (see [PP] for general results).

Definition 4. Given $A : \Omega \to \mathbb{R}$, the Ruelle operator $\mathscr{L}_A$ acts on functions $\psi : \Omega \to \mathbb{R}$ in the following way

$$\varphi(x) = \mathscr{L}_A(\psi)(x) = \sum_{a=1}^{d} e^{A(ax)} \psi(ax).$$

By this we mean $\mathscr{L}_A(\psi) = \varphi$.

Suppose $A = \log J$ is Lipchitz and normalized, that is, for any $x \in \Omega$ we have $\mathscr{L}_{\log J}(1)(x) = 1$.

All probabilities we consider will be over the Borel sigma-algebra $\mathscr{B}$ of Ω.

Definition 5. Given the continuous potential $\log J : \Omega \to \mathbb{R}$ let $\mathscr{L}^*_{\log J}$ be the operator on the set of Borel Measures on Ω defined so that $\mathscr{L}^*_{\log J}(\nu)$, for each Borel measure ν, satisfies:

$$\int_\Omega \psi \, d\mathscr{L}^*_{\log J}(\nu) = \int_\Omega \mathscr{L}_{\log J}(\psi) \, d\nu,$$

for all continuous functions ψ.

$\mathscr{L}^*_{\log J}$ takes probabilities in probabilities.

A probability which is fixed for such operator $\mathscr{L}^*_{\log J}$ is invariant for the shift and called a g-measure. In the case $\log J$ is Holder such fixed point is unique and is the Gibbs state for $\log J$ (see [PP]).

Is it true that there exist a metric d equivalent to d_θ such that $\mathscr{L}^*_{\log J}$ is a contraction in the 1 Wassertein distance d_1 associated to such d? The answer to this question is yes and we will address the question in the section 7. Before that we will present some more basic material in the next sections.

M. Stadlbauer presented a proof with an affirmative answer to the above question in a more general setting. The present proof is just a simplified version of the one in [Stad]. The proof is based in an adaptation to our case of the more general setting described in [Hair].

This is true indeed. A proof of this fact which applies for the Thermodynamic Formalism setting (the one above) and also for some iterated contraction systems appears in [KLS].

M. Stadlbauer presented a proof with an affirmative answer to the above question in a more general setting [Stad].

3 The total variation norm and the maximal coupling

Suppose ρ is a signed measure in the Borel sigma-algebra $\mathscr{B}$ of $\Omega = \{1,2,..,d\}^{\mathbb{N}}$ such that $\rho(\Omega) = 0$.

Definition 6. The total variation of a signed measure ρ as above is by definition

$$|\rho|_{tv} = 2 \sup_{A\in\mathscr{B}} \rho(A).$$

One can show by duality that

$$|\rho|_{tv} = \sup_{|\phi|_\infty \leq 1} \int \phi \, d\rho.$$

where ϕ are measurable and bounded and $|\phi|_\infty$ is the supremum norm.

Given two probabilites μ_1 and μ_2 in Ω one can consider the distance $|\mu_1 - \mu_2|_{tv}$, called the total variation distance of μ_1 and μ_2. This is a different way to measure the distance among two probabilities.

This distance is also known as the strong distance in opposition to the more well known concept of weak convergence of probabilities.

Remember that we denote $\mathscr{C}(\mu_1,\mu_2)$ the set plans Γ in $\Omega\times\Omega$ such that the projection in the first coordinate is μ_1 and in the second is μ_2.

Proposition 1. : *Given $\Gamma \in \mathscr{C}(\mu_1,\mu_2)$, then,*

$$|\mu_1 - \mu_2|_{tv} \leq 2\,\Gamma\,\{(x,y)\,|\,x \neq y\}.$$

Proof: Given $A \in \mathscr{B}$ we have that

$$\mu_1(A) - \mu_2(A) = \int\int_{x\in A} \Gamma(dx,dy) - \int_{y\in A}\int \Gamma(dx,dy) =$$

$$[\int_{x=y}\int_{x\in A}\Gamma(dx,dy)+\int_{x\neq y}\int_{x\in A}\Gamma(dx,dy)]-$$
$$[\int_{y\in A}\int_{x=y}\Gamma(dx,dy)+\int_{y\in A}\int_{x\neq y}\Gamma(dx,dy)]=$$
$$\int_{x\neq y}\int_{x\in A}\Gamma(dx,dy)-\int_{y\in A}\int_{x\neq y}\Gamma(dx,dy)\leq\int_{x\neq y}\int_{x\in A}\Gamma(dx,dy).$$

Remember that
$$|\mu_1-\mu_2|_{tv}=2\sup_{A\in\mathscr{B}}(\mu_1-\mu_2)(A).$$

Taking supremum in $A\in\mathscr{B}$ we get the claim. □

We follow the general reasoning of [Hol] and [Lin].

Suppose μ_1 and μ_2 are two probabilities on the Bernoulli space Ω.

Put $\lambda=\mu_1+\mu_2$, and
$$g=\frac{d\mu_1}{d\lambda},\, g'=\frac{d\mu_2}{d\lambda}.$$

Now we define Q on Ω by $dQ=(g\wedge g')\,d\lambda$, where $g\wedge g'$ denotes the infimum of g and g'.

Note that by Kantorovich Duality (see [Vi1])
$$|\mu_1-\mu_2|_{tv}=\sup_{|f|\leq 1,\, f \text{ measurable}}|\int f\,d\mu_1-\int f\,d\mu_2|.$$

Therefore,
$$|\mu_1-\mu_2|_{tv}=\sup_{|f|\leq 1,\, f \text{ measurable}}|\int f\,(g-g')d\lambda\,|=$$
$$\int_{g\geq g'}1\,(g-g')d\lambda+\int_{g<g'}(-1)\,(g-g')d\lambda\;=\int|g-g'|d\lambda. \tag{1}$$

Consider $\varphi:\Omega\to\Omega\times\Omega$ by $\varphi(x)=(x,x)$.

Finally, we denote by $\hat{Q}=\varphi^*(Q)$.

Note that the support of $\hat{Q}$ is the diagonal Δ in $\Omega\times\Omega$.

Now, let $\gamma=\hat{Q}(\Delta)=Q(\Omega)$.

We call $\nu_1=\mu_1-Q$ and $\nu_2=\mu_2-Q$, and finally we define a plan π in $\Omega\times\Omega$ by
$$\pi=\hat{Q}+\frac{\nu_1\otimes\nu_2}{1-\gamma}.$$

This plan is sometimes called **maximal coupling**.

Note that $\nu_2(\Omega)=\mu_2(\Omega)-Q(\Omega)=1-\gamma=\nu_1(\Omega)=\mu_1(\Omega)-Q(\Omega)$.

We claim that π projects in the first coordinate on μ_1. Indeed,
$$\pi(A\times\Omega)=\hat{Q}(A\times\Omega)+\frac{\nu_1\otimes\nu_2}{1-\gamma}(A\times\Omega)=$$

$$\hat{Q}(A\times A)+\frac{\nu_1(A)\otimes\nu_2(\Omega)}{1-\gamma}=$$

$$Q(A)+\frac{\nu_1(A)\,(1-\gamma)}{1-\gamma}=Q(A)+(\mu_1(A)-Q(A))=\mu_1(A).$$

The above also shows that π is a probability.
Moreover, π projects in the second coordinate on μ_2. Indeed,

$$\pi(\Omega\times A)=\hat{Q}(\Omega\times A)+\frac{\nu_1\otimes\nu_2}{1-\gamma}(\Omega\times A)=$$

$$\hat{Q}(A\times A)+\frac{\nu_1(\Omega)\otimes\nu_2(A)}{1-\gamma}=$$

$$Q(A)+\frac{\nu_2(A)\,(1-\gamma)}{1-\gamma}=Q(A)+(\mu_2(A)-Q(A))=\mu_2(A).$$

In this way $\pi\in\mathscr{C}(\mu_1,\mu_2)$. Therefore, it follows from a previous result that for such plan it is true the property $|\mu_1-\mu_2|_{tv}\leq 2\,\pi\{(x,y)\,|\,x\neq y\}$. Now we will show:

Theorem 1. *The plan π defined above satisfies*

$$|\mu_1-\mu_2|_{tv}=2\,\pi\{(x,y)\,|\,x\neq y\}.$$

Proof: First note that as $|g-g'|=g+g'-2\,(g\wedge g')$, we have by (1)

$$|\mu_1-\mu_2|_{tv}=\int|g-g'|\,d\lambda=2\,[\,1-\int(\,g\wedge g'\,)\,d\lambda\,]=$$

$$2\,(1-Q(\Omega))=2\,(1-\gamma)\,\geq\,2\,\pi(\Delta^c)=2\,\pi\{(x,y)\,|\,x\neq y\}.$$

The last inequality follows from

$$\pi(\Delta^c)=\hat{Q}(\Delta^c)+\frac{\nu_1\otimes\nu_2(\Delta^c)}{1-\gamma}\leq$$

$$\frac{\nu_1\otimes\nu_2(\Delta^c)}{1-\gamma}\leq\frac{\nu_1\otimes\nu_2(\Omega\times\Omega)}{1-\gamma}=\frac{\nu_1(\Omega)\times\nu_2(\Omega)}{1-\gamma}=\frac{(1-\gamma)^2}{1-\gamma}=1-\gamma.$$

□

Given a probability ν on the Bernoulli space Ω then the probability $\mu=\sigma^*(\nu)$ is by definition the one such that $\mu(A)=\nu(\,\sigma^{-1}(A)\,)$ for any Borel set A on Ω. We say that μ is invariant for the shift if $\mu=\sigma^*(\mu)$.

Proposition 2. *Given μ_1 and μ_2 two probabilities over Ω, then*

$$|\sigma^*(\mu_1)-\sigma^*(\mu_2)|_{tv}\,\leq\,|\mu_1-\mu_2|_{tv}.$$

Proof:
Note that by Kantorovich Duality (see [Vi1])

$$|\sigma^*(\mu_1)-\sigma^*(\mu_2)|_{tv} = \sup_{|f|\leq 1,\, f \text{ measurable}} |\int f\, d\sigma^*\mu_1 - \int f\, d\sigma^*\mu_2\,| =$$

$$\sup_{|f|\leq 1,\, f \text{ measurable}} |\int (f\circ\sigma)\, d\mu_1 - \int (f\circ\sigma)\, d\mu_2\,|.$$

The functions of the form $(f\circ\sigma)$ with $|f|\leq 1$ is a smaller class that the set of functions of the form g such that $|g|\leq 1$.

From this follows the claim. □

In this way the composition with the shift never increase the total variation norm of probabilities.

4 The estimation using T

Remember that points x in $\Omega=\{1,2,...,d\}^{\mathbb{N}}$ are denoted by $x=(x_1,x_2,x_3,...)$.

Definition 7. The coupling time T is **the** measurable function $T:\Omega\times\Omega\to\mathbb{N}$ given by

$$T(x,y)=\inf\{n\,|\,x_m=y_m \,\text{for all}\, m\geq n\},$$

for any $x,y\in\Omega$.

This value can be eventually ∞.

An alternative way to define the coupling time T is given by

$$T(x,y)=\inf\{n\,|\,\sigma^n(x)=\sigma^n(y)\},$$

for any $x,y\in\Omega$.

Note that there is a difference in estimating the total variation of two probabilities in the state space $\{1,2,..,d\}$ and in the Bernoulli space $\{1,2,..,d\}^{\mathbb{N}}$. We consider bellow results for each kind of setting.

We use the notation: for any n we have that X_n is the measurable function $X_n:\{1,2,..,d\}^{\mathbb{N}}\to\{1,2,..,d\}$ such that $X_n(x)=x_n$ if $x=(x_1,x_2,x_3,...,x_n,...)$.

Proposition 3. *Estimation in the state space - Suppose μ_1 and μ_2 are probabilities on $\{1,2,..,d\}$. Suppose also that Γ is a probability on $\{1,2,..,d\}^{\mathbb{N}}\times\{1,2,..,d\}^{\mathbb{N}}$, such that for a fixed n,*

for all $J\subset\{1,2...,d\}$, we have

$$\Gamma(X_n\in J, X_n\in\{1,2,..,d\})=\mu_1(J) \tag{2}$$

and for all $J \subset \{1,2...,d\}$, we have

$$\Gamma(X_n \in \{1,2,..,d\}, X_n \in J) = \mu_2(J). \tag{3}$$

Then,

$$\mu_1\{x\,|x_n \in J\} - \mu_2\{x\,|x_n \in J\} \leq \Gamma\{(x,y)\,|\,T(x,y) > n\}.$$

Therefore,

$$|\mu_1(X_n \in .) - \mu_2(X_n \in .)|_{tv} \leq 2\,\Gamma\{(x,y)\,|\,T(x,y) > n\}. \tag{4}$$

Proof:

$$\mu_1\{x\,|x_n \in J\} - \mu_2\{x\,|x_n \in J\} \leq$$
$$\int\int_{x_n \in J} \Gamma(dx,dy) - \int_{y_n \in J}\int \Gamma(dx,dy) =$$
$$[\int_{x_n=y_n}\int_{x_n \in J} \Gamma(dx,dy) + \int_{x_n \neq y_n}\int_{x_n \in J} \Gamma(dx,dy)] -$$
$$[\int_{y_n \in J}\int_{x_n=y_n} \Gamma(dx,dy) + \int_{y_n \in J}\int_{x_n \neq y_n} \Gamma(dx,dy)] =$$
$$\int_{x_n \neq y_n}\int_{x_n \in J} \Gamma(dx,dy) - \int_{y_n \in J}\int_{x_n \neq y_n} \Gamma(dx,dy) \leq$$
$$\int_{x_n \neq y_n}\int_{x_n \in J} \Gamma(dx,dy) \leq \Gamma\{(x,y)\,|\,T(x,y) > n\},$$

because if x and y are such that $x_n \neq y_n$, then, $T(x,y) > n$. □

We point out that for many plans Γ we have that $T = \infty$ almost everywhere. For some specials ones this is not true.

A more complex result is:

Proposition 4. *Estimation in the Bernoulli space - Suppose μ_1 and μ_2 are probabilities on $\{1,2,..,d\}^{\mathbb{N}}$. Given $\Gamma \in \mathscr{C}(\mu_1,\mu_2)$, then for any n,*

$$|\,(\sigma^n)^*(\mu_1) - (\sigma^n)^*(\mu_2)\,|_{tv} \leq 2\,\Gamma\{(x,y)\,|\,T(x,y) > n\}.$$

Proof:

Remember that if (x,y) is such that for fixed n we have $x_n \neq y_n$, then, $T_\Gamma(x,y) > n$. Given a set $A \subset \Omega$ in $\mathscr{B}$,

$$(\sigma^n)^*(\mu_1)(A) - (\sigma^n)^*(\mu_2)(A) =$$
$$\mu_1\{x\,|(\sigma^n)(x) \in A\} - \mu_2\{y\,|(\sigma^n)(y) \in A\} \leq$$
$$\int\int_{\{x\,|\,(\sigma^n)(x) \in A\}} \Gamma(dx,dy) - \int_{\{y\,|\,(\sigma^n)(y) \in A\}}\int \Gamma(dx,dy) =$$

$$[\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}}\Gamma(dx,dy)+$$
$$\int_{\{(\sigma^n)(x)=(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}}\Gamma(dx,dy)]-$$
$$[\int_{\{y\,|\,(\sigma^n)(y)\in A\}}\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\Gamma(dx,dy)+$$
$$\int_{\{y\,|\,(\sigma^n)(y)\in A\}}\int_{(\sigma^n)(x)=(\sigma^n)(y)}\Gamma(dx,dy)]=$$
$$\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}}\Gamma(dx,dy)-$$
$$\int_{\{y\,|\,(\sigma^n)(y)\in A\}}\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\Gamma(dx,dy)\leq$$
$$\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}}\Gamma(dx,dy)\leq$$
$$\int\int_{\{(x,y)\,|\,T(x,y)>n\}}\Gamma(dx,dy)\leq\Gamma\{(x,y)\,|\,T(x,y)>n\}.$$

Taking supremum among all A we get the claim. □

Suppose X_n, $n\in\mathbb{N}$ is a stochastic process over S and $\Omega=S^{\mathbb{N}}$. We assume that $X_n:\Omega\to S$ is such that $X_n(w)=w_n$, for any $n\in\mathbb{N}$, where $w=(w_1,w_2,...,w_n,..)\in\Omega$. On Ω we consider the sigma-algebra $\mathscr{A}$ generated by the cylinder sets (which is the same as the one generated by the open sets). The stochastic process determines a probability P on the sigma-algebra $\mathscr{A}$ of Ω (see [Walk] or [Lop]).

A **stopping time** on Ω is a measurable function $T:\Omega\to\mathbb{N}$, such that, the set $A=\{w\,|\,T(w)=N\}$ depends only $X_1,X_2,,..,X_N$. In other words to know if $T(w)=N$ we just have to consider the string $w_1,w_2,...,w_N$.

We follow Hollander [Hol].

Given a plan π on $\Omega\times\Omega$, a **generic stopping time** T and a non-decreasing function $\psi:\mathbb{N}\to[0,\infty)$, such that $\lim_{n\to\infty}\psi(n)=\infty$, assume that

$$\int\psi(\,T(x,y)\,)\,\pi(dx,dy)<\infty.$$

Note that for any n

$$\psi(n)\,\pi(T\,>\,n)\leq\int_{T>n}\psi(\,T(x,y)\,)\,\pi(dx,dy).$$

The right hand side tends to zero when $n\to\infty$ by the dominated convergence theorem because $\int\psi(\,T(x,y)\,)\,\pi(dx,dy)<\infty$.

Given ε suppose that N is big enough such that for all $n>N$ we have $\psi(n)\,\pi(T\,>\,n)\leq\varepsilon$.

Suppose the plan π is in $\mathscr{C}(\mu_1,\mu_2)$.

Then, from last proposition we have that for n,

$$|\,(\sigma^n)^*(\mu_1) - (\sigma^n)^*(\mu_2)\,|_{tv} \leq 2\,\pi\{(x,y)\,|\,T(x,y) > n\} \leq 2\,\varepsilon\,\frac{1}{\psi(n)}. \tag{5}$$

Estimates of the form

$$|\mu_1(X_n \in\ .) - \mu_2(X_n \in .)|_{tv} \leq 2\,\pi\{(x,y)\,|\,T(x,y) > n\} \leq 2\,\varepsilon\,\frac{1}{\psi(n)}. \tag{6}$$

are also important and interesting.

In this way one can get an estimation of the speed of convergence of the above difference by means of the coupling time and ψ. This depends on the plan π we pick. The main point is the smart guess for choosing the plan. All of the above also depends on $\psi(n)$ which can be of polynomial type $n^{-\gamma}$, $\gamma > 0$, or exponential type $e^{-\lambda n}$, $\lambda > 0$, depending of the problem.

The main problem in these kind of questions is to estimate $\pi(T > n)$, where $n \in \mathbb{N}$. This of course depends on π and T.

5 Exponential convergence to equilibrium for finite state Markov chains

5.1 The estimation using T^1

We want to consider now the following one: suppose $\mathscr{P}$ is a finite stochastic matrix with all entries positive and λ its unique invariant vector of probability.

Consider a Markov chain X_n with another initial condition ν and the same transition matrix $\mathscr{P}$. We denote the associated probability by P.

We want to show the existence of $0 < \rho < 1$, such that,

$$|\lambda - P(X_n \in .)|_{tv} \leq 2\,(1-\rho)^n.$$

Theorem 2 will describe the speed of convergence (which is exponential) to the equilibrium λ when times n goes to infinity for the chain X_n, $n \in \mathbb{N}$. This result is the main goal of the next subsections.

Definition 8. The coupling time T_1 is **the** measurable function $T_1 : \Omega \times \Omega \to \mathbb{N}$ given by

$$T_1(x,y) = \inf\{n\,|\,x_n = y_n\,\}$$

for any $x, y \in \Omega$.

This value can be eventually ∞.

Note that this couplig time is different from the other one denoted by T. Note also that $T_1(x,y) \leq T(x,y)$ for any (x,y).

We point out that given a plan π on $\Omega \times \Omega$ we have that

$$\pi(T_1(x,y) > n) \leq \pi\{(x,y) \in \Omega \times \Omega \,|\, x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\}.$$

Consider a d dimensional Stochastic Matrix $\mathscr{P} = (P_{i,j})_{i,j=1,2...,d}$ with all entries positive and denote by ρ the minimum value of $P_{i,j}$. Consider two vector of initial probabilities λ and ν. Using the fixed matrix $\mathscr{P}$ they define respectively two different probabilities on Ω which are denoted by μ_1 and μ_2. They generate respectively two different stochastic processes $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ taking values on $\{1,2...,d\}$. We assume the initial time is $n=1$. This is consistent with the notation $x = (x_1, x_2, x_3, ...)$.

Consider the plan Γ on $\Omega \times \Omega$ such that $\Gamma = \mu_1 \otimes \mu_2$. For a fixed n we will estimate $\Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_n \neq y_n\}$.

Note that for any others initial vector of probabilities $\tilde{\lambda}$ and $\tilde{\nu}$ we have

$$\Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_n = y_n\} = \sum_{a_n=1}^{d} \Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_n = y_n = a_n\} =$$

$$\sum_{a_n=1}^{d} [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\lambda}_j P_{j,a_1} P_{a_1,a_2} ... P_{a_{n-1},a_n}] [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\nu}_j P_{j,a_1} P_{a_1,a_2} .. P_{a_{n-1},a_n}] \geq$$

$$\sum_{a_n=1}^{d} [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\lambda}_j P_{j,a_1} P_{a_1,a_2} ... P_{a_{n-1},a_n}] [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\nu}_j P_{j,a_1} P_{a_1,a_2} .. P_{a_{n-2},a_{n-1}} \rho] =$$

$$\sum_{a_n=1}^{d} [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\lambda}_j P_{j,a_1} P_{a_1,a_2} ... P_{a_{n-1},a_n}] \rho = \rho.$$

In this way $\Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_n \neq y_n\} \leq (1-\rho)$. A very important remark is that the above expression does not depend of the initial vector of probability $\tilde{\lambda}$ and $\tilde{\nu}$. Indeed, just depend on the matrix $\mathscr{P}$.

As $\Gamma = \mu_1 \otimes \mu_2$, then the associated Stochastic Process $(X_n, Y_n)_{n\in\mathbb{N}}$ taking values on $\{1,2,...,d\} \times \{1,2,...,d\}$ such that

$$P((X_1,Y_1) \in (A_1,B_1), ..., (X_n,Y_n) \in (A_n,B_n)) =$$

$$\mu_1(A_1 \times ... \times A_n \times \{1,2,..,d\}^{\mathbb{N}})\, \mu_2(B_1 \times ... \times B_n \times \{1,2,..,d\}^{\mathbb{N}}) =$$

$$\Gamma((A_1 \times ... \times A_n \times \{1,2,..,d\}^{\mathbb{N}}) \times (B_1 \times ... \times B_n \times \{1,2,..,d\}^{\mathbb{N}})$$

is a Markov chain with stochastic matrix

$$\begin{pmatrix} \mathscr{P} & 0 \\ 0 & \mathscr{P} \end{pmatrix}. \tag{7}$$

The initial vector of probability is such that

$$P(X_1 = i, Y_1 = j) = \lambda_i \, \nu_j. \tag{8}$$

Γ is Markov probability on the space $\{1,2,..,d\} \times \{1,2,..,d\}^{\mathbb{N}}$.

We point out that given any vector of initial probability (do not have to be as above in (8)) for this Markov Chain with transition matrix (7) and with values on $\{1,2,..,d\} \times \{1,2,..,d\}$ we get that the associated Markov probability P on the space $(\{1,2,..,d\} \times \{1,2,..,d\})^{\mathbb{N}}$ is such that $P(\{(x,y) \in \Omega \times \Omega \,|\, x_n = y_n\}) \geq \rho$.

Note that the event $X_1 \neq Y_1$ is not independent of $X_2 \neq Y_2$.

We want to show that for any n we have

$$\Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\} \leq (1-\rho)^n. \tag{9}$$

From this will follow at once that:

Proposition 5.

$$\Gamma(T_1 > n) \;\leq\; (1-\rho)^n. \tag{10}$$

Proof: Note that

$$\Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\} =$$
$$\Gamma\{X_1 \neq Y_1, X_2 \neq Y_2, ..., X_n \neq Y_n\} =$$
$$\frac{\Gamma\{X_1 \neq Y_1, X_2 \neq Y_2, ..., X_n \neq Y_n\}}{\Gamma\{X_1 \neq Y_1\}} \Gamma\{X_1 \neq Y_1\} =$$
$$\Gamma\{X_1 \neq Y_1, X_2 \neq Y_2, ..., X_n \neq Y_n \,|\, X_1 \neq Y_1\} \,\Gamma\{X_1 \neq Y_1\} \leq$$
$$\Gamma\{X_2 \neq Y_2, ..., X_n \neq Y_n \,|\, X_1 \neq Y_1\} \,(1-\rho).$$

We will need in this moment the following property: suppose Z_n, $n \in \mathbb{N}$, is a Markov chain taking values in a finite set E with transition matrix $\hat{\mathscr{P}}$. Consider also a certain initial condition $\tilde{\pi}$. This defines a Markov Probability P of the space of paths $E^{\mathbb{N}}$.

Then,

$$P(Z_2 \in A_2, Z_3 \in A_3, ...Z_n \in A_n \,|\, Z_1 \in A_1) =$$
$$\frac{P(Z_1 \in A_1, Z_2 \in A_2, Z_3 \in A_3, ..., Z_n \in A_n)}{P(Z_1 \in A_1)} =$$
$$\frac{\sum_{j \in A_1} \; \tilde{\pi}_{a_j} \, [\sum_{a_2 \in A_2} ... \sum_{a_n \in A_n} \, \hat{\mathscr{P}}_{a_j,a_2} \, \hat{\mathscr{P}}_{a_2,a_3} ... \hat{\mathscr{P}}_{a_{n-1},a_n}]}{\sum_{j \in A_1} \tilde{\pi}_{a_j}} =$$
$$P_\gamma(Z_2 \in A_2, Z_3 \in A_3, ...Z_n \in A_n), \tag{11}$$

where γ is an initial vector of probability on E, such that, for $r \in A_1$ we have

$$\gamma_r = \frac{\tilde{\pi}_{a_r}}{\sum_{j \in A_1} \tilde{\pi}_{a_j}}, \tag{12}$$

and for r which is not in A_1 we have and $\gamma_r = 0$.

The above property is a particular case of Prop 1.7 page 78 in [RY] for a Markov Processes Z_n taking values on E which says

$$P_q(Z_{t+s} \in A \,|\, \mathscr{F}_t) = P_{Z_s}(Z_t \in A),$$

where $\mathscr{F}_t$ is the sigma algebra determined by the process from time 1 to t and q is an initial vector of probability on E. This expression is sometimes called the Markov Property.

We will use (11) taking $Z_n = (X_n, Y_n)$, $A_n = \{X_n \neq Y_n\}$ and $E = \{1,2,..,d\} \times \{1,2,..,d\}$.

Therefore,

$$\Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\} =$$

$$\Gamma\{X_2 \neq Y_2, ..., X_n \neq Y_n \,|\, X_1 \neq Y_1\}\,(1-\rho).$$

$$\tilde{\Gamma}\{X_2 \neq Y_2, ..., X_n \neq Y_n\}\,(1-\rho),$$

where $\tilde{\Gamma}$ is another plan (another Markov probability on the space $\{1,2,..,d\} \times \{1,2,..,d\}^{\mathbb{N}}$) due to the fact that we changed the initial condition as described in (11).

Now we use the fact that $\tilde{\Gamma}\{(x,y) \in \Omega \times \Omega \,|\, x_n \neq y_n\} \leq (1-\rho)$ and we get that

$$\Gamma\{(x,y) \in \Omega \times \Omega \,|\, x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\} \leq$$

$$\tilde{\Gamma}\{X_2 \neq Y_2, ..., X_n \neq Y_n\}\,(1-\rho) \leq$$

$$\tilde{\Gamma}\{X_3 \neq Y_3, ..., X_n \neq Y_n \,|\, X_2 \neq Y_2\}(1-\rho)^2.$$

Proceeding by induction we get (9). □

Consider a d by d stochastic matrix $\mathscr{P}$ with all positive entries and we denote by X_n, $n \in \mathbb{N}$, and Y_n, $n \in \mathbb{N}$ two independent Markov Processes with values on $\{1,2...,d\}$ respectively associated to the same matrix $\mathscr{P}$. We denote $S = \{1,2...,d\}$.

Consider an initial probability λ for X_n^λ and ν for Y_n^ν. This will define probabilities on $\{1,2...,d\}^{\mathbb{N}}$ which we will denote respectively by P_1 and P_2. We will denote P_1^j the probability we get from the Markov Process defined by the transition matrix $\mathscr{P}$ and the initial vector of probability δ_j, where $j \in \{1,2,...,d\}$.

We consider first a Stochastic Process V_n, $n \in \mathbb{N}$, with values on $S^2 = \{1,2...,d\}^2$ of the form $V_n = (X_n^\lambda, Y_n^\nu)$. This by assumption is the probability $\tilde{P} = P_1 \otimes P_2$ on $(\{1,2...,d\}^2)^{\mathbb{N}}$. That is the processes X_n^λ and Y_n^ν are independent. By abuse of language $\tilde{P}$ defines a probability on $(\{1,2...,d\}^2)^n$ for each $n \in \mathbb{N}$.

We consider another Stochastic Process U_n, $n \in \mathbb{N}$, with values on $\{1,2...,d\}^2$ of the form $U_n = (X_n, Y_n)$. This will define another probability $\hat{P}$ on $(\{1,2...,d\}^2)^{\mathbb{N}}$. In this case the processes X_n and Y_n are not independent. By abuse of language $\hat{P}$ defines a probability on $(\{1,2...,d\}^2)^n$ for each $n \in \mathbb{N}$.

For a fixed n we have to define $\hat{P}$ is sets of the form

$$\hat{P}[\,(A_1 \times B_1\,) \times (A_2 \times B_2\,) \times ... \times (A_n \times B_n\,) \times (S \times S)^{\mathbb{N}}],$$

where $A_k, B_k \subset S$, $k = 1,2...,n$.

We will define $\hat{P}$ in the following way: suppose $T_1 : \{1,2...,d\}^{\mathbb{N}} \times \{1,2...,d\}^{\mathbb{N}} \to \mathbb{N}$ is given by

$$T_1(x,y) = \inf\{n \,|\, x_n = y_n\,\}$$

for any $x, y \in \{1,2...,d\}^{\mathbb{N}}$.

Consider a fixed value n.

Denote by G_n the set of elements on $\{1,2...,d\}^n \times \{1,2...,d\}^n$ such that $T_1(x,y) > n$.

For any subset $K \subset G_n$ define

$$\hat{P}(K \times (S \times S)^{\mathbb{N}}) = \tilde{P}(K \times (S \times S)^{\mathbb{N}}). \tag{13}$$

This defines $\hat{P}$ on cylinders of the form

$$(\{a_1\} \times \{b_1\}\,) \times (\{a_2\} \times \{b_2\}\,) \times ... \times (\,\{a_n\} \times \{b_n\}\,) \times (S \times S)^{\mathbb{N}} \subset G_n \times (S \times S)^{\mathbb{N}},$$

$n \in \mathbb{N}$, $a_j, b_j \in \{1,2...,d\}$, $j = 1,2...,n$.

For a fixed n the sets of the form $H^n_k = H_k = \{T_1 = k\}$, $k = 1,2,...n$, and the set $G_n \times (S \times S)^{\mathbb{N}}$ define a partition of $\{1,2...,d\}^{\mathbb{N}} \times \{1,2...,d\}^{\mathbb{N}}$.

We denote by Δ the diagonal on $\{1,2...,d\}^2$.

For a fixed n and $k \leq n$, now we will define $\hat{P}$ on subsets of H^n_k.

This means we suppose that k is the smaller one among $\{1,2,...,n\}$, such that, $(A_k \times B_k\,) \cap \Delta \neq \emptyset$.

For defining probabilities on this case we can assume without lost of generality that $A_k = B_k = \{j\}$ for a certain j.

Then,

$$\hat{P}[\,(A_1 \times B_1\,) \times (A_2 \times B_2\,) \times ... \times (A_n \times B_n\,)\,] =$$

$$P_1(A_1 \times ... \times A_{k-1} \times\, j)\; P_2(B_1 \times ... \times B_{k-1} \times\, j)\, P^j_1(A_{k+1} \times A_{k+2} \times ... \times A_n). \tag{14}$$

This procedure defines a probability $\hat{P}$ on $(\{1,2,...,d\}^2)^n$ and subsequently on $(\{1,2,...,d\}^2)^{\mathbb{N}}$.

One can say in an elusive way that $\hat{P}$ describes the following process: two particles located on S evolve in time in an independent way (a Markov Chain in $S \times S$) and when they meet they stay together in the future according to the law of the Markov chain in S.

Note that for fixed n and $j \in S$, we have

$$\hat{P}\{(X_n,Y_n)=(j,j),\text{ and also }(X_m,Y_m)\cap\Delta=\emptyset,\text{ for some } n<m\ \}=0. \tag{15}$$

We claim that:

Proposition 6. *For fixed $j\in\{1,2,..,d\}$ and $n\in\mathbb{N}$*

$$\hat{P}(X_n=j,Y_n\in\{1,2,..,d\})=P_1(X_n=j). \tag{16}$$

Proof: We denote $S=\{1,2,...,d\}$ and we consider a fixed $n\in\mathbb{N}$ and a fixed $j\in S$.

Note that

$$\tilde{P}(X_n=j,Y_n\in\{1,2,..,d\})=\sum_{r=1}^{d}\tilde{P}(X_n=j,Y_n=r)=$$

$$P_1(X_n=j)\sum_{r=1}^{d}P_2(Y_n=r)=P_1(X_n=j). \tag{17}$$

Given $(x,y)\in(S\times S)^{\mathbb{N}}$ we have that either $T_1(x,y)>n$ or $T_1(x,y)\leq n$.

For a fixed n we use the notation described above for the sets of the form $H_k^n=H_k=\{T_1=k\}$, $k=1,2,...n$, and the set $G_n\times(S\times S)^{\mathbb{N}}$ which define a partition of $\{1,2...,d\}^{\mathbb{N}}\times\{1,2...,d\}^{\mathbb{N}}$.

Note also that

$$\hat{P}(X_n=j,Y_n\in\{1,2,..,d\})=\sum_{r\neq j}^{d}\hat{P}(X_n=j,Y_n=r)+\hat{P}(X_n=j,Y_n=j)=$$

$$\sum_{r\neq j}^{d}\tilde{P}(X_n=j,Y_n=r)+\hat{P}(X_n=j,Y_n=j). \tag{18}$$

From (15) and also (13) we get that $\sum_{r\neq j}^{d}\tilde{P}(X_n=j,Y_n=r)=\sum_{r\neq j}^{d}\hat{P}(X_n=j,Y_n=r)$. This part corresponds to the subset $G_n\times(S\times S)^{\mathbb{N}}$.

We claim that $\hat{P}(X_n=j,Y_n=j)=\tilde{P}(X_n=j,Y_n=j)$.

Indeed,

$$\hat{P}(X_n=j,Y_n=j)=$$

$$\sum_{k=1}^{n-1}\hat{P}(\{T_1=k\}\cap(S\times S)^n)+\hat{P}(\{T_1=n\}\cap\{(X_n,Y_n)=(j,j)\})=$$

$$\sum_{r=1}^{d}\sum_{k=1}^{n-1}\hat{P}((S\times S)^{k-1}\times(\{r\}\times\{r\})\times(S\times S)^{n-k})+$$

$$\hat{P}(\{T_1=n\}\cap\{(X_n,Y_n)=(j,j)\})=$$

$$\sum_{r=1}^{d}\sum_{k=1}^{n-1}\hat{P}((S\times S-\Delta)^{k-1}\times(\{r\}\times\{r\})\times(S\times S)^{n-k})+$$

$$\hat{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\sum_{r=1}^{d} \sum_{k=1}^{n-1} \tilde{P}((S \times S - \Delta)^{k-1} \times (\{r\} \times \{r\}) \times (S \times S)^{n-k}) +$$

$$\tilde{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\sum_{k=1}^{n-1} \tilde{P}(\{T_1 = k\} \cap (S \times S)^n) + \tilde{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\tilde{P}(X_n = j, Y_n = j).$$

Now, from (18) we get that

$$\hat{P}(X_n = j, Y_n \in \{1, 2, .., d\}) = \tilde{P}(X_n = j, Y_n \in \{1, 2, .., d\}) = P_1(X_n = j).$$

□

In a similar way we get:

Proposition 7. *For fixed $j \in \{1, 2, .., d\}$ and $n \in \mathbb{N}$*

$$\hat{P}(X_n \in \{1, 2, .., d\}, Y_n = j,) = P_2(Y_n = j). \tag{19}$$

We will need later the following result:

Proposition 8. *The probability $\hat{P}$ on $(\{1, 2..., d\}^2)^{\mathbb{N}}$ which is a plan on the product space $\{1, 2..., d\}^{\mathbb{N}} \times \{1, 2..., d\}^{\mathbb{N}}$ satisfies*

$$\hat{P}(T > n) = \hat{P}(T_1 > n) = \tilde{P}(T_1 > n), \tag{20}$$

where $\tilde{P}$ is the independent process.

Proof: We have that $\hat{P}(T > n) = \hat{P}(T_1 > n)$ because of (15).

Finally, we claim that $\hat{P}(T_1 > n) = \tilde{P}(T_1 > n)$. Indeed, if $m > n$ we have from (13) that $\hat{P}(\{T_1 = m\} \cap (S \times S)^m) = \tilde{P}(\{T_1 = m\} \cap (S \times S)^m)$. □

5.2 Speed estimation on total variation

The probabilities P_1 and P_2 on $\{1, 2, ..., d\}^{\mathbb{N}}$ and X_n, Y_n, $n \in \mathbb{N}$, were described above. We want to take advantage of (4), (10) and (16).

We get before estimates of the form (6) for T satisfying expression (4), that is, for X_n and Y_n independent

$$|P_1(X_n \in .) - P_2(Y_n \in .)|_{tv} \leq 2\Gamma\{(x, y) \,|\, T(x, y) > n\},$$

when

$$T(x,y) = \inf\{n \,|\, x_m = y_m \,\text{for all}\, m \geq n\},$$

and where Γ satisfies the hypothesis (2) and (3) of Proposition 3.

This of course depends of a good choice of Γ. We take $\Gamma = \hat{P}$ and we know from (16) and (19) that these hypothesis are true.

Then, we get from (20) that

$$|P_1(X_n \in .) - P_2(Y_n \in .)|_{tv} \leq 2\hat{P}\{(x,y) \,|\, T > n\} = 2\tilde{P}\{(x,y) \,|\, T_1 > n\}.$$

Now, from (10) we get that $\tilde{P}(T_1 > n) \leq (1-\rho)^n$.

From this follows that

$$|P_1(X_n \in .) - P_2(Y_n \in .)|_{tv} \leq 2\,(1-\rho)^n. \tag{21}$$

Note that $P_1(X_n \in .)$ and $P_2(Y_n \in .)$ are probabilities on the state space S.

The bottom line is: suppose λ is the stationary vector for the d by d matrix $\mathscr{P}$ (which have all entries positive) and ν is another initial vector of probability on $\{1,2,...,d\}$. This defines respectively two probabilities on $\{1,2,...,d\}^{\mathbb{N}}$ which we denote P_1 and P_2. We also denote, respectively, X_n, $n \in \mathbb{N}$, the first Markov Process and Y_n, $n \in \mathbb{N}$, the second.

Note that for any n and $J \subset S$ we have $P_1(X_n \in J) = P_1(X_0 \in J)$ because λ is stationary. In other words $P_1(X_n \in .)$ is λ

In this way we are analyzing the time evolution of two probabilities on the space state S and from (21), we get:

Theorem 2. *Suppose the transition matrix $\mathscr{P}$ has all entries positive and $\lambda\,\mathscr{P} = \lambda$, where λ is the initial stationary vector of probability on S. Then, for any n we have*

$$|\lambda - P_2(Y_n \in .)|_{tv} \leq 2\,(1-\rho)^n,$$

and this describes the speed of convergence to the equilibrium λ when times goes to infinity for a chain Y_n with initial condition ν and matrix transition $\mathscr{P}$.

6 The $\bar{d}$ distance

Given two probabilities on μ, ν in the set $\Omega = \{1,2\}^{\mathbb{N}}$, consider the set $\mathscr{C}(\mu,\nu)$ of plans λ in $\Omega \times \Omega$ such the projection in the first coordinate is μ and in the second is ν.

Definition 9. - A joining is a probability on $\mathscr{C}(\mu,\nu)$ which is invariant for the dynamical system $T : \Omega \times \Omega \to \Omega \times \Omega$ given by $T(x,y) = (\sigma(x), \sigma(y))$. The set of such joinings is denoted by $J(\mu,\nu)$.

A general reference for joinings is [GLAS].

Definition 10. Denote $\{1,2\}^n = Q_n$ and consider the d_n-Hamming metric in Q_n defined by

$$d_n(x,y) = \frac{1}{n}\sum_{j=1}^{n} I_{\{x_j \neq y_j\}}.$$

Points x in Ω are denoted by $x = (x_0, x_1, x_2, ...)$.

Definition 11. For two σ-invariant probabilities μ and ν we define the distance

$$\bar{d}(\mu,\nu) = \inf_{\lambda\in J(\mu,\nu)} \int I_{\{x_0\neq y_0\}} \lambda(dx,dy) =$$

$$\inf_{\lambda\in J(\mu,\nu)} \lim_{n\to\infty} \int d_n(x,y)\lambda(dx,dy) =$$

$$\lim_{n\to\infty} \inf_{\lambda\in J(\mu,\nu)} \int d_n(x,y)\lambda(dx,dy).$$

In the above equalities we used the ergodic theorem (see [Walk]).

The above value try to minimize the asymptotic mean disagreement between the symbols of the paths for a given plan.

We point out that the $\bar{d}$ has a dynamical content. Several dynamical properties are preserves under limit over the $\bar{d}$ distance (see [CQ]). In this paper properties related to Bernoullicity, g-measures and Gibbs states are considered. For instance the map that takes a potential to its equilibrium state is continuous with respect to the $\bar{d}$ distance (see Theorem 3 in [CQ]).

Theorem 3. *Suppose μ is the independent Bernoulli probability associated to p_1, p_2 and ν is the independent Bernoulli probability associated to q_1, q_2.*

Suppose $p_1 \leq q_1$. Then,

$$\bar{d}(\mu,\nu) = q_1 - p_1 = \frac{1}{2}(|q_1 - p_1| + |q_2 - p_2|).$$

Proof:

Given $\lambda \in J(\mu,\nu)$ we have

$$\int I_{\{x_0\neq y_0\}} \lambda(dx,dy) =$$

$$\int_{[1]\times[1]} I_{\{x_0\neq y_0\}} \lambda(dx,dy) + \int_{[1]\times[2]} I_{\{x_0\neq y_0\}} \lambda(dx,dy) +$$

$$\int_{[2]\times[1]} I_{\{x_0\neq y_0\}} \lambda(dx,dy) + \int_{[2]\times[2]} I_{\{x_0\neq y_0\}} \lambda(dx,dy) =$$

$$\int_{[1]\times[2]} I_{\{x_0\neq y_0\}}\lambda(dx,dy)+\int_{[2]\times[1]} I_{\{x_0\neq y_0\}}\lambda(dx,dy).$$

We denote $a_{i,j}=\int_{[i]\times[j]} I_{\{x_0\neq y_0\}}\lambda(dx,dy)$.

We have the relations

$$\lambda_{1,1}+\lambda_{2,1}=\nu[1]=q_1\,,\ \lambda_{1,2}+\lambda_{2,2}=\nu[2]=q_2\,,$$

$$\lambda_{1,1}+\lambda_{1,2}=\mu[1]=p_1\,,\ \lambda_{2,1}+\lambda_{2,2}=\mu[2]=p_2\,.$$

Denote by $t=\lambda_{1,1}$, then

$$\lambda_{1,2}=p_1-t,$$

$$\lambda_{2,1}=q_1-t,$$

and

$$\lambda_{2,2}=p_2-q_1+t,$$

We have to minimize $a_{12}+a_{21}\ =\ p_1+q_1-2t$ given the above linear constrains. This is a linear maximization problem and the largest possible value of t will be the solution.

From the fact that the $\lambda_{i,j}$ are non negative we get that $t\leq q_1$ and $t\leq p_1$. But as we assume that $q_1\geq p_1$, we get the only restriction $0\leq t\leq p_1$.

Taking $\lambda_{11}=p_1$ we get the minimal value for the sum $a_{12}+a_{21}$ which is $p_1+q_1-2p_1=q_1-p_1$.

This shows that $\bar{d}(\mu,\nu)\geq q_1\,-\,p_1$.

Now we will show that there exists a plan λ that realizes the value q_1-p_1.

Consider a new Bernoulli independent process (X_n,Y_n) with value on $\{1,2\}\times\{1,2\}$.

We set

$$P(X_0=1,Y_0=1)=p_1,P(X_0=1,Y_0=2)=\ 0,$$

$$P(X_0=2,Y_0=1)=q_1-p_1,P(X_0=2,Y_0=2)=q_2=p_2-q_1+p_1,$$

It is easy to se that such plan λ is in $J(\mu,\nu)$ and $\int I_{\{x_0\neq y_0\}}\lambda(dx,dy)=q_1-p_1$. $\square$

A natural question is what can be said for the $\bar{d}$ distance of two finite state Markov Processes (taking values in the same state space S) obtained from two different stochastic matrices, or more generally, for Gibbs states. This is a not so easy problem and partial results can be found for instance in [Ellis2] [GLT] [BFG] and [BFG2]. In most of them couplings are used in an essential way.

We refer the reader to [FG] [Perez] [To] for more details on couplings from a probabilistic point of view.

7 Contraction for the dual of the Ruelle Operator

In this section we study properties related to Gibbs states of Lipchitz potentials (see [PP] for general results).

Suppose $A = \log J$ is Lipchitz and normalized, that is, for any $x \in \Omega$ we have $\mathscr{L}_{\log J}(1)(x) = 1$.

If $x = (x_1, x_2, ..) \in \{1, 2, ..., d\}^{\mathbb{N}}$ and $t \in \mathbb{N}$, we denote by $z_j^t(x)$, $j = 1, 2, ..., d^t$, the d^t solutions of $\sigma^t(z) = x$.

We denote $J^t(z_j^t(x))$ the expression $e^{\sum_{k=0}^{t-1} \log J(\sigma^k(z_j^t(x))}$.

Therefore, given x and y we have two probabilites

$$(P^t)^*(\delta_x) = \sum_{j=1}^{d^t} \delta_{z_j^t(x)} J^t(z_j^t(x))$$

and

$$(P^t)^*(\delta_y) = \sum_{j=1}^{d^t} \delta_{z_j^t(y)} J^t(z_j^t(y)).$$

For each k we have that the points $z_j^k(x)$ and $z_j^k(y)$, $j = 1, 2, 3, .., 2^k$, are all different but the distance $d_\theta(z_j^k(x), z_j^k(y))$ is at most θ^k (j pair by pair).

Suppose $A = \log J$ has Lipchitz constant M. Then, for t and $j = 1, 2, 3, .., 2^t$ fixed

$$| \sum_{k=0}^{t-1} \log J(\sigma^k(z_j^t(x)) - \sum_{k=0}^{t-1} \log J(\sigma^k(z_j^t(y)) \,| \leq$$

$$\sum_{k=0}^{t-1} M \, d_\theta(\sigma^k(z_j^t(x)) - \sigma^k(z_j^t(y))) \,| \leq$$

$$M \, [\sum_{k=0}^{t-1} \theta^k] \, d_\theta(x, y) \leq M \frac{1}{1-\theta} \, d_\theta(x, y)$$

In this way for any x and y we have for any t, $j = 1, 2, 3, .., 2^t$,

$$\frac{J^t(z_j^t(x))}{J^t(z_j^t(y))} \leq e^{M \frac{1}{1-\theta} d_\theta(x,y)}. \tag{22}$$

The above kind of estimation is known as the bounded distortion property. It is key element in the subsequent developments.

Lemma 1. *For every $\delta > 0$, there is T such that for any $k \geq T$ there exists an $a > 0$, so that*

$$\sup_{\Gamma \in \mathscr{C}((P^k)^*(\delta_x), (P^k)^*(\delta_y))} \Gamma\{(x', y')) \in \Omega \times \Omega \, : \, d_\theta(x', y') \leq \delta\} \geq a.$$

Proof.
In order to proof the Lemma, given x and y we construct explicitly an element in

$$\mathscr{C}\big((P^k)^*(\delta_x),\,(P^k)^*(\delta_y)\big).$$

By the orbit structure of the Bernoulli shift, the preimages of x and y come in pairs, and as stated above, the distance of a pair satisfies $d_\theta(z^k_j(x),z^k_j(y)) < \theta^k$. Hence, if $T > \log\delta/\log\theta$, then $d_\theta(z^T_j(x),z^T_j(y)) < \delta$. The other basic observation for the construction stems from bounded distortion in (22). Namely, there exists $\Lambda \in (0,1]$, independent from T, j and x such that[1] for $j = 1,2,3,..,d^T$

$$\Lambda J^T(z^T_j(x)) \le \alpha^T_j := \inf_{z\in\Omega} J^T(z^T_j(z)) \le J^T(z^T_j(x)).$$

For $\beta^T_j(x) := J^T(z^T_j(x)) - \alpha^T_j$, $j = 1,2,3,..,d^T$, we hence obtained a decomposition into a strictly positive part $\beta^T_j(x)$ and a strictly positive part α^T_j, which is independent from x (and y) but comparable to $J^T(z^T_j(x))$ by Λ. Hence, we obtain a decomposition of a probability measure into two sub-probability measures by

$$(P^T)^*(\delta_x) = \sum_{j=1}^{d^T} \delta_{z^T_j(x)}\alpha^T_j + \sum_{j=1}^{d^T} \delta_{z^T_j(x)}\beta^T_j(x) =: \mu_T + \nu^x_T =$$

$$\sum_{j=1}^{d^T} \delta_{z^T_j(x)} J^T(z^T_j(x)).$$

We claim that the probability Γ on $\Omega\times\Omega$ defined by

$$\Gamma := \sum_{j=1}^{d^T} \delta_{(z^T_j(x),z^T_j(y))}\alpha^T_j + \frac{1}{\nu^x_T(\Omega)}\nu^x_T\otimes\nu^y_T.$$

is in $\Gamma \in \mathscr{C}\big((P^k)^*(\delta_x),\,(P^k)^*(\delta_y)\big)$.
Indeed, consider a Borel set A then

$$\Gamma(\Omega\times A) = \sum_{j=1}^{d^T} \delta_{(z^T_j(x),z^T_j(y))}\alpha^T_j(\Omega\times A) + \frac{1}{\nu^x_T(\Omega)}\nu^x_T\otimes\nu^y_T(\Omega\times A) =$$

$$\sum_{j=1}^{d^T} \delta_{z^T_j(y)}\alpha^T_j(A) + \nu^y_T(A) = \mu_T(A) + \nu^y_T(A) =$$

$$\sum_{j=1}^{d^T} \delta_{z^T_j(y)} J^T(z^T_j(y))(A) = (P^T)^*(\delta_y)(A).$$

Note that $\nu^x_T(\Omega) = \nu^y_T(\Omega) = 1 - \mu_T(\Omega)$.

[1] $\Lambda = 1$ if and only if J is constant on cylinders of length 1

In the same way given A we have

$$\Gamma(A\times\Omega)=\sum_{j=1}^{d^T}\delta_{(z_j^T(x),z_j^T(y))}\,\alpha_j^T(A\times\Omega)+\frac{1}{\nu_T^x(\Omega)}\,\nu_T^x\otimes\nu_T^y(A\times\Omega)=$$

$$\sum_{j=1}^{d^T}\delta_{z_j^T(x)}\,\alpha_j^T(A)+\nu_T^x(A)\,\frac{\nu_T^y(\Omega)}{\nu_T^x(\Omega)}=\mu_T(A)+\nu_T^x(A)=$$

$$\sum_{j=1}^{d^T}\delta_{z_j^T(x)}J^T(z_j^T(x))\,(A)\,=\,(P^T)^*(\delta_x)\,(A).$$

The analogous result for sets of the for $\Omega\times A$ is true.
This shows that $\Gamma\in\mathscr{C}(\,(P^k)^*(\delta_x),\,(P^k)^*(\delta_y)).$
We claim that Γ also satisfies

$$\Gamma\,\{(x',y'))\in\,\Omega\times\Omega\,:\,d_\theta(x',y')\le\delta\,\}\ge\Lambda.$$

Indeed,

$$\Gamma\,\{(x',y'))\in\,\Omega\times\Omega\;d_\theta(x',y')\le\delta\,\}\ge$$

$$\sum_{j=1}^{d^T}\delta_{(z_j^T(x),z_j^T(y))}\;\alpha_j^T\,[\,\{(x',y'))\in\,\Omega\times\Omega\,:\,d_\theta(x',y')\le\delta\,\}\,]=$$

$$\sum_{j=1}^{d^T}\delta_{z_j^T(x)}\;\alpha_j^T(x)\ge\sum_{j=1}^{d^T}\Lambda J^T(z_j^T(x))\,=\,\Lambda$$

This proves the Lemma for $a:=\Lambda$. □

We define

$$\mathrm{var}_n f=\sup\{|f(x)-f(y)|\,:x_i=y_i,\,0\le i<n\},$$

and the pseudo norm

$$|f|_\theta=\sup\{\frac{\mathrm{var}_n f}{\theta^n}\;:n\ge 0\}.$$

We know that if $\log J$ is Lipchitz, then, there exist $C>0$ and $\alpha_1\in(0,1)$ such that for any $t\ge 0$, and any Lipchitz function ϕ we have (prop 2.1 in [PP])

$$|\mathscr{L}^t(\phi)\,|_\theta\le C\sup_x|\phi(x)|+\theta^t|\phi|_\theta.$$

We wrote this in the notation of [Hair] as

$$\sup\{\frac{\mathrm{var}_n \mathscr{L}^t(\phi)}{\theta^n}\ : n \geq 0\} \leq C \sup_x |\phi(x)| + \alpha_1 \sup\{\frac{\mathrm{var}_n \phi}{\theta^n}\ : n \geq 0\}.$$

The above expression is known as the Lasota-Yorke inequality.

Take $\delta < \frac{1-\alpha_1}{2C}$. Now, we define a metric $d(x,y) = \min\{1, \delta^{-1} d_\theta(x,y)\}$. The two metrics d and d_θ are equivalent.

Remember that we denote $\mathscr{C}(\mu_1, \mu_2)$ the set plans in $\Omega \times \Omega$ such that the projection in the first coordinate is μ_1 and in the second is μ_2.

Remember also that, we define the 1-Wasserstein metric associated to d

$$d_1(\mu_1, \mu_2) = \inf\{\int \int d(x,y)\, d\,\Gamma(dx,dy) \mid \Gamma \in \mathscr{C}(\mu_1, \mu_2)\}.$$

Proposition 9. *There exist* $\alpha < 1$, *where* $\alpha = \max\{1 - \frac{a}{2}, \frac{1}{2}(1+\alpha_1)\}$, *and* $t > 0$, *such that, for any* x, y

$$d_1((P^t)^*(\delta_x), (P^t)^*(\delta_y)) \leq \alpha\, d(x,y).$$

The proof will be done later.

Suppose that this result is proved then we get:

Theorem 4. *There exist* $\alpha < 1$, *where* $\alpha = \max\{1 - \frac{a}{2}, \frac{1}{2}(1+\alpha_1)\}$, *and* $t > 0$, *such that, for any* μ_1, μ_2

$$d_1((P^t)^*(\mu_1), (P^t)^*(\mu_2)) \leq \alpha\, d_1(\mu_1, \mu_2).$$

Proof: The main idea is to prove first the following claim: suppose Q is the d_1-optimal plan for μ_1 and μ_2, then,

$$d_1\left((P^t)^*(\mu_1), (P^t)^*(\mu_2)\right) \leq \int d_1((P^t)^*(\delta_x), (P^t)^*(\delta_y))\, dQ(dx,dy).$$

Suppose the plan in $\Omega \times \Omega$ denoted by $Q(dx,dy)$ has marginals μ_1 and μ_2 in respectively the first and second coordinates.

We will prove the result for a more general continuous potential c. Then, you just have to take $c = d$ in order to get the claim.

Given a continuous cost $c(z_1,z_2)$, $c : X \times X \to \mathbb{R}$, we assume that Q is c-optimal for μ_1 and μ_2. Now, given two points x,y suppose $R^{x,y}(dz_1,dz_2)$ is the c-optimal probability plan for $P^*(\delta_x)$ and $P^*(\delta_y)$.

We denote $S(dz_1,dz_2)$ the plan

$$S(dz_1,dz_2) = \int \int R^{x,y}(dz_1,dz_2)\, Q(dx,dy).$$

We are going to show that the marginals of this plan are μ_1 and μ_2.

Indeed,

$$\int\int \varphi(z_1) S(dz_1, dz_2) =$$

$$\int\int\int\int \varphi(z_1) R^{x,y}(dz_1, dz_2)\, Q(dx, dy)\, dx\, dy\, dz_1\, dz_2 =$$

$$\int\int\int \varphi(z_1) P^*(\delta_x)\, Q(dx, dy)\, dxdydz_1 =$$

$$\int [P^*(\int\int \varphi(z_1)\, Q(\,dx\,, dy) dy dz_1\,)]\,(\delta_x) =$$

$$\int [P^*(\int \varphi(.)\, Q(\,.\,, dy) dy\,)\,(\delta_x)]\, dx = \int \varphi(x)\, dP^*(\mu_1)\ (dx)$$

because P^* is linear on measures.

In this way the first marginal of $S(dz_1, dz_2)$ is $P^*\,(\mu_1)$.

In the same way one can prove that the second marginal of $S(dz_1, dz_2)$ is $P^*\,\mu_2$. Now we consider $c(x,y) = d(x,y)$.

From the above we get

$$d_1(P^*\,(\mu_1), P^*\,(\mu_2)) \leq \int d(x,y)\, S(dx, dy),$$

where S was defined from Q which is the d-optimal plan for μ_1 and μ_2.

Therefore, from the above

$$d_1\,(P^*(\mu_1), P^*(\mu_2)) \leq \int d_1(P^*(\delta_x), P^*(\delta_y))\, dQ(dx, dy),$$

where Q is the d_1-optimal plan for μ_1 and μ_2.

In a similar way given $t > 0$ on can show the analogous result

$$d_1\,((P^t)^*(\mu_1), (P^t)^*(\mu_2)) \leq \int d_1((P^t)^*(\delta_x), (P^t)^*(\delta_y))\, dQ(dx, dy),$$

where Q is the d_1-optimal plan for μ_1 and μ_2.

Therefore, if there exists a $t > 0$ and $\alpha < 1$, such that, for any x and y we have

$$d_1((P^t)^*(\delta_x), (P^t)^*(\delta_y)) \,\leq\, \alpha\, d(x,y),$$

then

$$d_1((P^t)^*(\mu_1), (P^t)^*(\mu_2) \,\leq\, \alpha\, d_1(\mu_1, \mu_2).$$

This is so because

$$d_1\,((P^t)^*(\mu_1), (P^t)^*(\mu_2)) \leq \int d_1((P^t)^*(\delta_x), (P^t)^*(\delta_y))\, dQ(dx, dy) \leq$$

$$\alpha \int d(x,y)\, dQ(dx, dy) = \alpha\, d_1(\mu_1, \mu_2).$$

□

Now we prove that for any x,y

$$d_1((P^t)^*(\delta_x),(P^t)^*(\delta_y)) \leq \alpha\, d(x,y).$$

Suppose $d_\theta(x,y) \leq \delta$, where $\delta < \frac{1-\alpha_1}{2C}$.
Remember that $d(x,y) = \min\{1, \delta^{-1} d_\theta(x,y)\}$. In this case $d(x,y) = \delta^{-1} d_\theta(x,y)$.
By Kantorovich duality (see [Vi1] and [Vi2])

$$d_1(\mu_1,\mu_2) = \sup_{\phi: X\to\mathbb{R} \text{ has } d \text{ Lipchitz constant} \leq 1} \{ \int \phi\, d\mu_1 - \int \phi\, d\mu_2 \}.$$

We have to show that: if ϕ has d-Lipchitz constant smaller than 1, then, for such pair of x,y

$$|\mathscr{L}^t_{\log J}\phi(x) - \mathscr{L}^t_{\log J}\phi(y)| \leq \alpha d(x,y) = \alpha \delta^{-1}\, d_\theta(x,y).$$

We can assume without lost of generality that ϕ attains the value 0.
In this case $\sup_x |\phi(x)| \leq 1$.
Moreover,

$$\sup \frac{\delta\, |\phi(x) - \phi(y)|}{d_\theta(x,y)} \leq \sup \frac{|\phi(x)-\phi(y)|}{d(x,y)} \leq 1.$$

Then,

$$\frac{|\mathscr{L}^t_{\log J}(\phi)(x) - \mathscr{L}^t_{\log J}(\phi)(y)|}{d_\theta(x,y)} \leq$$

$$C \sup_x |\phi(x)| + \alpha_1 \sup \frac{|\phi(x)-\phi(y)|}{d_\theta(x,y)} \leq C + \alpha_1\, \delta^{-1}.$$

As $\delta < \frac{1-\alpha_1}{2C}$, then $C < \frac{1-\alpha_1}{2}\, \delta^{-1}$.
Therefore, from the above, we get that for any x,y such that $d_\theta(x,y) \leq \delta$, we have

$$|\mathscr{L}^t_{\log J}(\phi)(x) - \mathscr{L}^t_{\log J}(\phi)(y)| \leq d_\theta(x,y)\,(C + \alpha_1\, \delta^{-1}) \leq$$

$$d_\theta(x,y)\,(\frac{1-\alpha_1}{2}\, \delta^{-1} + \alpha_1\, \delta^{-1}) =$$

$$d_\theta(x,y)\, \delta^{-1}\, (\frac{1+\alpha_1}{2}) = d(x,y)\, (\frac{1+\alpha_1}{2}) \leq d(x,y)\, \alpha,$$

because $\alpha = \max\{1 - \frac{a}{2}, \frac{1}{2}(1+\alpha_1)\}$.

Now we suppose that x,y such that $d_\theta(x,y) > \delta$. This implies that $d(x,y) = 1$.
We denote

$$\Delta_\delta = \{(x',y') \in \Omega \times \Omega \,:\, d_\theta(x',y') \leq \frac{1}{2}\delta\}.$$

For such δ there exists $a > 0$ and $T > 0$, such that, for $k > T$, there exists a plan $\Gamma = \Gamma_k$ which satisfies $\Gamma \in \mathscr{C}(\,(P^k)^*(\delta_x),\,(P^k)^*(\delta_y))$ and $\Gamma\,(\Delta_\delta) \geq a$.

Note that if $d_\theta(x',y') \leq \frac{1}{2}\delta$, then $d(x',y') \leq \frac{1}{2}$.

Therefore,

$$\int d(x',y')\,\Gamma(dx',dy') \;\leq \frac{1}{2}\Gamma\,(\Delta_\delta) + 1 - \Gamma\,(\Delta_\delta) = 1 - \frac{1}{2}\Gamma\,(\Delta_\delta) \leq 1 - \frac{a}{2},$$

because $d(x',y') \leq 1$ in the complement of Δ_δ.

Then, if $d_\theta(x,y) > \delta$ we get

$$d_1((P^t)^*(\delta_x),(P^t)^*(\delta_y)) \leq \int d(x',y')\,\Gamma(dx',dy') \;\leq$$

$$(1-\frac{a}{2}) = (1-\frac{a}{2})\,d\,(x,y) \;\leq\; \alpha\; d(x,y)$$

because $\alpha = \max\{1-\frac{a}{2}, \frac{1}{2}(1+\alpha_1)\}$.

From all this it follows the main result.

References

[Aus] T. Austin, Egodic Theory - Notes 11: Entropy, couplings and joinings, Lecture Notes - Courant Institute, NYU (2013)

[BFG] X. Bressaud, R. Fernandez and A. Galves, Decay of correlations for non-Holderian dynamics. A coupling approach. Electron. J. Probab. 4, no. 3, Electron. J. Probab., 4:no. 3, 19 pp. (electronic) (1999).

[BFG2] X. Bressaud, R. Fernandez and Antonio Galves, Speed of $\bar{d}$-convergence for Markov approximations of chains with complete connections. A coupling approach, Stochastic Processes and their Applications 83 (1999) 127–138

[CQ] Z. Coelho and A. Quas. A criteria for $\overline{d}$-continuity, Trans. AMS, Volume 350, Number 8, 3257-3268 (1998).

[Ellis2] M. Ellis, Distances between two-state Markov Processes attainble by Markov Joinings, TAMS, Vol 241, (1978)

[FFG] R. Fernandez, P. Ferrari and A. Galvez, Coupling, renewal and Perfect Simulation of Chains of infinite order, Univ of Rouen (2001)

[FG] P. Ferrari and A. Galves, Construction of Stochastic Proceses, Coupling and Regeneration, XIII Escuela Venezolana de Matematica,

[GP] S. Galatolo and M. J. Pacifico, Lorenz-like flows: exponential decay of correlations for the Poincar map, logarithm law, quantitative recurrence, ETDS 30, no. 6, 17031737. (2010)

[GLT] S. Gallo, M. Lerasle and D. Takahashi, Markov approximation of chains of infinte order in the $\bar{d}$ metric, Markov Process. Related Fields 19 (2013), no. 1, 51-82.

[GLAS] E. Glasner, Ergodic Theory via Joinings, AMS (2003)

[Hair] M. Hairer and J. Mattingly, Spectral gaps in Wasserstein distance and the 2-D Stochastic Navier-Stokes Equations, The Annals of Probability 2008, Vol. 36, No. 6, 2050-2091

[Hol] F. den Hollander, Probability Theory: The Coupling Method, Lectures Notes - Mathematical Institute, Leiden University (2012)

[Kl] B. Kloeckner, Optimal transport and dynamics of expanding circle maps acting on measures, Ergodic Theory Dynam. Systems 33 (2013), no. 2, 529-548.

[KLS] B. Kloeckner, A. O. Lopes and M. Stadlbauer, Contraction in the Wassertein metric for some Markov Chains and applications for the dynamics of expanding maps, Nonlinearity, 28, Number 11, 4117-4137 (2015)

[Lin] T. Lindvall, Lectures on the Coupling method, Dover (1992)

[Lop] A. Lopes and S. Lopes, Introdução aos Processos Estocásticos para estudantes de Matemática, UFRGS (2015)

[PP] W. Parry and M. Pollicott. Zeta functions and the periodic orbit structure of hyperbolic dynamics, *Astérisque* Vol 187-188 (1990)

[Perez] D. Levin, Y. Perez and E. Wilmer, Markov Chains and Mixing Times, AMS (2008)

[RY] D. Revuz and M. Yor, Continuous Martingales and Brownian motion, Springer Verlag (1991)

[Stad] M. Stadlbauer, Coupling methods for random topological Markov chains, to appear in Ergodic Theory and Dynamical Systems

[Sulku] H. Sulku, Explicit correlation bounds for expanding maps using the couplig method, preprint ArXiv (2013)

[To] H. Torrison, Coupling, Stationary and Regeneration, Springer (2000).

[Walk] C. Walkden, Ergodic Theory, Lecture Notes University of Manchester (2014)

[Vi1] C. Villani, *Topics in optimal transportation*, AMS, Providence (2003)

[Vi2] C. Villani, *Optimal transport: old and new*, Springer-Verlag, Berlin, (2009)