Robustness of Markov processes on large networks

R.S.MacKay

Mathematics Institute and Centre for Complexity Science,
University of Warwick, Coventry CV4 7AL, UK
R.S.MacKay@warwick.ac.uk

Summary. A metric on the space of probability measures on the state of a large network is introduced, with respect to which the stationary measure of a Markov process on the network is proved under suitable hypotheses to vary uniformly smoothly with parameters and the rate of relaxation to equilibrium to never suddenly decrease.

1 Introduction

A variety of stochastic dynamic systems can be modelled as Markov processes on networks. Examples include the system of ion channels through a cell membrane [2], spread of disease in an orchard [12], the system of checkout queues at a supermarket, and probabilistic cellular automata models of "subgrid" weather [21]. In each case one wants to know how the behaviour depends on parameters. This can aid understanding, prediction, control and design.

The first aspect to investigate for such a system is its stationary probability measures (existence and uniqueness), the second is whether and how fast they attract other initial probability measures, and the third is how the answers depend on parameters. These questions are well studied for general Markov processes, e.g. [13], but the specifics that arise for processes on large networks seem not to have been addressed adequately.

For this paper, a *network* is a countable (finite or countably infinite) set S with a local state space X_s for each "site" $s \in S$. Each local state space carries a metric d_s; if X_s is discrete and no metric is specified, take the *discrete metric* $d_s(x_s, x'_s) = 1$ for all $x_s \neq x'_s$. A *Markov process on a network* is a discrete- or continuous-time Markov process on the product space $X = \prod_{s\in S} X_s$. The discrete-time case includes probabilistic cellular automata (PCA, [23, 15]), and the continuous-time case includes interacting particle systems (IPS, e.g. [16]).

States of X are denoted $x = (x_s)_{s\in S}$. The set of probability measures on X is denoted $\mathcal{P}(X)$ (more precisely, the Borel probability measures with respect to product topology on X). For questions about spatial correlations (cf. [8]), S should be endowed with a metric and the transition rates for a given site chosen to depend weakly on the states of distant sites, but that topic is deferred to a later paper.

Recall (e.g. [22]) that for a Markov process with discrete-time transition matrix P (or operator if the X_s are not discrete) the *stationary probability measures* are those $\rho \in \mathcal{P}(X)$ satisfying $\rho P = \rho$. For a continuous-time Markov process with transition rate matrix (or operator) Q, they are the solutions of $\rho Q = 0$. A stationary probability measure exists under very general conditions. Uniqueness holds under more restricted conditions, for example if the state space is finite and has a single communicating component (the *communicating components* of a Markov process are the maximal subsets of the state space such that it is possible to go from any state of the subset to any other one). Note in particular that uniqueness does not hold for systems with non-trivial conservation laws, like diffusion processes; although there might be a unique stationary measure on a subset of fixed conserved quantities, such a subset does not have the form of a product over the network.

It might appear that the existence, uniqueness and attraction rate of the stationary probability measure ρ is just a question of linear algebra, but when one asks how fast ρ moves with parameters of the model, it becomes a question of analysis, i.e. it depends on metrics on $\mathcal{P}(X)$.

There are many metrics on spaces of probability measures, e.g. [11]. Most are unsuitable for large networks, however, because they assign a relatively large distance between probability distributions that one would like to regard as close. For example, "total variation convergence essentially never occurs for particle systems" (p.70 of [16]). The reader who is not aware there is a problem is invited to jump ahead to Section 2 and return here when convinced.

If the X_s are complete separable metric spaces ("Polish spaces") and

$$\Omega := \sup_{s\in S} \text{ diam } X_s$$

is finite, I suggest that a useful metric on $\mathcal{P}(X)$ for large networks is defined by taking the worst case of the difference of expectations of a class of functions relative to a semi-norm of Dobrushin's [6] to be recalled below. These ingredients have been used widely, e.g. [15, 16, 17], but that they make a useful metric seems not to have been remarked. Specifically, for $\rho, \sigma \in \mathcal{P}(X)$, let

$$D(\rho,\sigma) = \sup_{f\in F\setminus C} \frac{\rho(f)-\sigma(f)}{\|f\|_F}, \tag{1}$$

where F is the space of continuous (with respect to product topology on X) functions $f : X \to \mathbb{R}$ with finite

$$\|f\|_F = \sum_{s\in S} \Delta_s(f), \tag{2}$$

$$\Delta_s(f) = \sup\{\frac{f(x)-f(x')}{d_s(x_s,x'_s)} : x_r = x'_r \ \forall r \neq s, x_s \neq x'_s\}, \tag{3}$$

C denotes the constant functions, and for a measure μ, $\mu(f) = \int f \, d\mu$.

The supremum in (1) is finite, indeed at most Ω, because

$$\rho(f)-\sigma(f) \le \sup_x f(x) - \inf_y f(y) \le \sum_{s\in S} \Delta_s(f) \text{ diam } X_s \le \|f\|_F \Omega.$$

The bounds can be approached arbitrarily closely by choosing $s \in S$ such that diam X_s is close to Ω, points $0_s, 1_s \in X_s$ for which $d_s(0_s, 1_s)$ is close to diam X_s,

probability measures ρ, σ concentrated on states with $x_s = 1_s, 0_s$, respectively, and $f(x) = d_s(x_s, 0_s)$; so $\mathcal{P}(X)$ has diameter Ω. Because $\| \ \|_F$ is a semi-norm, vanishing only on constants, and the functions $f \in F$ are continuous with respect to product topology, it follows that D is a metric on $\mathcal{P}(X)$.

Equation (2) defines a norm on the space F mod C of equivalence classes of functions in F modulo addition of constants. F mod C can be thought of as a space of "local" Lipschitz functions (no proximity of the sites s on which $\Delta_s(f)$ is non-small is imposed, but continuity of f in product topology forces it to depend negligibly on sites outside some bounded region). Equation (1) induces the dual norm on the space $\mathcal{Z}(X)$ of "zero-charge" measures on X, i.e. measures μ such that $\mu(c) = 0$ for constant functions c, by:

$$\|\mu\|_{\mathcal{Z}} = \sup_{f \in F \setminus C} \frac{\mu(f)}{\|f\|_F}. \tag{4}$$

The space $\mathcal{Z}(X)$ is complete with respect to $\| \ \|_{\mathcal{Z}}$ because the dual of a normed space is always complete. $\mathcal{P}(X)$ is complete with respect to D, because homeomorphic to a closed subset of $\mathcal{Z}(X)$, using that the X_s are Polish. Often I'll drop (X).

For linear maps L on $\mathcal{Z}$, like the restriction of P to $\mathcal{Z}$, write

$$\|L\|_{\mathcal{Z}} = \sup_{\mu \in \mathcal{Z} \setminus 0} \frac{\|\mu L\|_{\mathcal{Z}}}{\|\mu\|_{\mathcal{Z}}}.$$

Often I will drop the subscript $\mathcal{Z}$, but remember that the norm refers to the restriction of the operator to $\mathcal{Z}$. Such a map L can be considered as acting on $f \in F$ mod C: Lf is the unique $g \in F$ mod C such that $\mu(g) = (\mu L)(f)$ for all $\mu \in \mathcal{Z}$. This gives a convenient way to estimate the size of L:

$$\|L\|_{\mathcal{Z}} \leq \|L\|_F := \sup_{f \in F \setminus C} \|Lf\|/\|f\|, \tag{5}$$

because for all $\mu \in \mathcal{Z}$ and $f \in F$ with $Lf \notin C$,

$$\frac{\mu L f}{\|f\|} \leq \frac{\mu L f}{\|Lf\|} \frac{\|Lf\|}{\|f\|} \leq \|\mu\| \|L\|_F.$$

Actually, $\|L\|_{\mathcal{Z}} = \|L\|_F$, by choosing a sequence of μ to approach equality in $\mu Lf \leq \|\mu\| \|Lf\|$, but we will not need this.

The metric (1) on $\mathcal{P}(X)$ permits the following two theorems, which are the main results of this paper. The first says that under suitable assumptions the stationary probability measure of a family of Markov processes on networks depends smoothly on parameters, uniformly in the size of the network. To save space, attention is restricted here to the discrete-time case, but there are directly analogous statements and proofs for the continuous-time case.

Theorem 1. *(a) If discrete-time Markov transition matrix P_0 has unique stationary probability measure ρ_0, the restriction of $I - P_0$ to $\mathcal{Z}$ is invertible, $K := \|(I - P_0)^{-1}\|_{\mathcal{Z}} < \infty$, $\delta := \|P - P_0\|_{\mathcal{Z}} < 1/K$ and $\beta := \|\rho_0(P - P_0)\|_{\mathcal{Z}} < \infty$, then P has unique stationary probability measure ρ,*

$$\|\rho - \rho_0\| \leq \frac{\beta}{K^{-1} - \delta},$$

and $(I-P)$ is invertible on $\mathcal{Z}$ with $\|(I-P)^{-1}\|_{\mathcal{Z}} \leq (K^{-1}-\delta)^{-1}$.
(b) If P_λ is a family of Markov transition matrices depending continuously on parameters λ in a manifold M, and satisfies the conditions of (a) at λ_0, then there is a unique stationary probability measure ρ_λ for all λ in an open set $\Lambda \subset M$ containing all λ with $\|P_\lambda - P_0\| < 1/K$, ρ_λ varies continuously with λ on Λ, and $I-P_\lambda$ does not have bounded inverse on $\mathcal{Z}$ for $\lambda \in \partial\Lambda$.
(c) If P_λ varies C^1 with λ then so does ρ_λ on Λ, and

$$\frac{d\rho}{d\lambda} = \rho \frac{dP}{d\lambda}(I-P)^{-1}_{\mathcal{Z}}.$$

Definitions of continuous and C^1 dependence of a Markov process on parameters will be given at the appropriate points in Section 3.

The second main result of this paper is that if a stationary probability measure attracts exponentially then the rate of attraction does not decrease significantly on small change of parameters.

Definition 1. *For a discrete-time Markov process with transition matrix P, a stationary probability measure ρ* attracts exponentially with factor $r<1$ *if there is a* prefactor $C \in \mathbb{R}$ *such that $D(\sigma P^n, \rho) \leq C r^n D(\sigma,\rho)$ $\forall n \geq 0$ for all σ close enough to ρ in $\mathcal{P}(X)$.*

Theorem 2. *If ρ_0 is stationary for discrete-time transition matrix P_0 and attracts exponentially with factor $r<1$ and prefactor C then for all P with $\delta := \|P-P_0\|_{\mathcal{Z}} < \frac{1-r}{C}$ its stationary measure ρ attracts exponentially with factor at most $r+C\delta$ and the same prefactor C.*

Again there is an analogous result in continuous time.

Section 2 shows why various standard metrics on $\mathcal{P}(X)$ are not suitable. Section 3 proves Theorem 1. Section 4 proves Theorem 2. Section 5 gives some examples of application of the results, and Section 6 summarises and poses some problems for the future.

2 Standard metrics on spaces of probabilities

To see why there is a need for a metric like (1), let $S = \{1,\dots,N\}$, $X_s = \{0,1\}$ for each $s \in S$, and for $\lambda \in [0,1]$ let p_λ be Bernoulli$(1-\lambda,\lambda)$, i.e. the product of identical independent distributions with $p_\lambda(x_s = 1) = \lambda$ for all $s \in S$. One can think of p_λ as the stationary probability measure for a discrete-time Markov process with independent transition matrices $P_s = \begin{bmatrix} 1-\lambda & \lambda \\ 1-\lambda & \lambda \end{bmatrix}$ or continuous-time Markov process with independent transition rate matrices $Q_s = \begin{bmatrix} -\lambda & \lambda \\ 1-\lambda & \lambda-1 \end{bmatrix}$. Let us evaluate the speed v of change of p_λ with respect to λ for some standard metrics D on $\mathcal{P}(X)$, i.e. $v = \lim_{\varepsilon\to 0} D(p_\lambda, p_{\lambda+\varepsilon})/\varepsilon$.

The *total variation metric* is defined by

$$D_V(\rho,\sigma) = \sup_{A\subset X} \rho(A) - \sigma(A)$$

over measurable subsets A. If ρ and σ are absolutely continuous with respect to some reference measure μ this can equivalently be written as $\frac{1}{2}\int |\frac{d\rho}{d\mu} - \frac{d\sigma}{d\mu}|\, d\mu$, where $\frac{d\rho}{d\mu}$ is the Radon-Nikodym derivative, so the total variation metric is half the L^1 distance between the densities (some authors move the factor of 2 into D_V). For $\varepsilon > 0$, $D_V(p_{\lambda+\varepsilon}, p_\lambda)$ is attained by the event

$$\frac{n_1}{n_0} > \frac{-\log(1-\varepsilon/(1-\lambda))}{\log(1+\varepsilon/\lambda)},$$

where n_j is the number of sites with state $j = 0, 1$. In the limit $\varepsilon \to 0$ this event is $n_1 > N\lambda$, so the central limit theorem on $p_{\lambda'}(n_1 > N\lambda)$ yields

$$v \sim \sqrt{\frac{N}{2\pi\lambda(1-\lambda)}},$$

which grows like $\sqrt{N}$, whereas diam $\mathcal{P}(X)$ in total variation metric is only 1.

The *relative entropy* (Kullback-Leibler divergence) of ρ from σ is

$$h(\rho|\sigma) = \sum_{x\in X} \rho(x) \log \frac{\rho(x)}{\sigma(x)}$$

in the discrete case (assuming $\rho(x) = 0$ whenever $\sigma(x) = 0$, with interpretation $0 \log \frac{0}{0} = 0$), or $\int d\sigma(x)\phi(\frac{d\rho}{d\sigma}(x))$ with $\phi(t) = t\log t, t \geq 0$ for ρ absolutely continuous with respect to σ. It is non-negative with equality iff $\rho = \sigma$, but is not symmetric, and even its symmetrisation $h(\rho|\sigma) + h(\sigma|\rho)$ does not satisfy the triangle inequality in general. Nevertheless,

$$D(\rho,\sigma) = \sqrt{h\left(\rho|\frac{\rho+\sigma}{2}\right) + h\left(\sigma|\frac{\rho+\sigma}{2}\right)} \tag{6}$$

is a metric on $\mathcal{P}(X)$ [10] (the argument of the square root is known as the Jeffreys divergence or twice the Jensen-Shannon divergence). Now $h(p_\lambda|(p_\lambda + p_{\lambda+\varepsilon})/2)$ is the expectation of $-\log(\frac{1}{2} + \frac{1}{2}(1+\frac{\varepsilon}{\lambda})^n(1-\frac{\varepsilon}{1-\lambda})^{N-n})$ with respect to the binomial distribution Binomial(N,λ) for n. Expanding to second order and taking the expectation gives $\frac{N\varepsilon^2}{8\lambda(1-\lambda)} + O(\varepsilon^3)$. Thus

$$v = \sqrt{\frac{N}{4\lambda(1-\lambda)}},$$

so grows like $\sqrt{N}$ again, whereas diam $\mathcal{P}(X)$ in this metric is $\sqrt{2\log 2}$.

The *Hellinger metric* for a discrete space is

$$D(\rho,\sigma) = \sqrt{\sum_{x\in X} (\sqrt{\rho(x)} - \sqrt{\sigma(x)})^2}, \tag{7}$$

or more generally for probabilities absolutely continuous with respect to a reference measure μ,

$$D(\rho,\sigma) = \sqrt{\int d\mu \left(\sqrt{\frac{d\rho}{d\mu}} - \sqrt{\frac{d\sigma}{d\mu}}\right)^2}.$$

For the family p_λ, this gives speed $v = \sqrt{\frac{N}{\lambda(1-\lambda)}}$, whereas diam $\mathcal{P}(X) = \sqrt{2}$.

Up to factors of $\sqrt{2}$, (6) and (7) for ρ close to σ agree to first order with *Fisher's information metric*, given by the Riemannian metric

$$ds^2 = \sum_{x\in X} \rho(x)(d(\log\rho(x)))^2,$$

and similar for the absolutely continuous generalisation.

The *projective metric* [3] is defined on $\mathcal{P}_+(X)$ only, the set of positive probabilities. For discrete state spaces it is

$$D(\rho,\sigma) = \sup_{x\in X} \log\frac{\rho(x)}{\sigma(x)} - \inf_{y\in X} \log\frac{\rho(y)}{\sigma(y)},$$

and for mutually absolutely continuous probabilities on infinite state spaces replace $\frac{\rho(x)}{\sigma(x)}$ by $\frac{d\rho}{d\sigma}(x)$. For $\varepsilon > 0$, $D(p_{\lambda+\varepsilon}, p_\lambda)$ is attained by $x_s = 1$, $y_s = 0$ for all $s \in S$, and evaluates to $N(\log(1+\frac{\varepsilon}{\lambda}) - \log(1-\frac{\varepsilon}{1-\lambda}))$, so

$$v = \frac{N}{\lambda(1-\lambda)}.$$

With respect to this metric, $\mathcal{P}_+(X)$ has infinite diameter, so one might propose dividing the metric by N to obtain a speed uniform in N. This scaled metric, however, would give distance of order only $1/N$ for a change at a single site in a product distribution, and it does not extend to an infinite network.

Finally, I consider the *transportation metric* [24] (due to Monge in Euclidean space, shown to be a metric in general by Kantorovich, rediscovered successively by Vasserstein and Hutchinson). Although not itself suitable, it is a necessary concept for making the general class of examples in Section 5. It requires the underlying space to be a metric space (X,d) and is defined on $\mathcal{P}(X)$ by

$$D_T(\rho,\sigma) = \inf_\tau \int d(x,x')\ \tau(dx,dx')$$

over probability measures τ on $X\times X$ with marginals ρ and σ (there are variants with other "cost" functions than $d(x,x')$). Such τ are called "joinings" of the two probability distributions [20] (most probabilists call them "couplings", but "coupling" already has many other meanings); the infimum is attained and those which realise it are called "optimal joinings". D_T can also be written as

$$D_T(\rho,\sigma) = \sup_{f\in L\setminus C} \frac{\rho(f)-\sigma(f)}{\|f\|_L}, \tag{8}$$

where L is the space of Lipschitz functions $f : X \to \mathbb{R}$, C is the constant functions, and $\|f\|_L = \sup_{x\neq x'} \frac{f(x)-f(x')}{d(x,x')}$, the best Lipschitz constant. For the discrete metric, $D_T = D_V$, so D_T can be seen as a generalisation of the total variation metric to take into account the distance one measure has to be moved to turn it into the

other. The transportation metric is essentially the dual norm on the space of zero-charge measures $\mathcal{Z}(X)$, to the Lipschitz norm on F mod C. The diameter of $\mathcal{P}(X)$ is precisely the diameter of X. Specialising to a product of metric spaces $X = \prod_{s\in S} X_s$, one can endow it (or at least subsets for which the following are finite) with the ℓ_1 metric $d(x,x') = \sum_{s\in S} d_s(x_s,x'_s)$ or the ℓ_∞ metric $d(x,x') = \sup_{s\in S} d(x_s,x'_s)$ (or other choices). Each choice induces a transportation metric on $\mathcal{P}(X)$, equivalently norm $\|\ \|_{K_p}$ on $\mathcal{Z}$ for $p = 1,\infty$. For the ℓ_∞ metric on X, $D_T = D_V$, so $v \sim \sqrt{\frac{N}{2\pi\lambda(1-\lambda)}}$ again and diam $\mathcal{P}(X) = \Omega$. With respect to the ℓ_1 metric on X, optimal transport between two Bernoulli distributions can be achieved by optimal transport of the associated binomial distributions with respect to the standard metric on $\{0,\dots,N\}$; this is achieved by moving the difference between their cumulative distributions up by 1 (as in a formula of Kolmogorov), giving exact answer $D_T(p_\lambda, p_{\lambda+\varepsilon}) = N\varepsilon$, so $v = N$. The diameter of $\mathcal{P}(X)$ is $\sum_{s\in S}$ diam X_s, which grows like N if all X_s have similar diameter, so one could divide the ℓ_1 norm by N (cf. Ornstein's $\bar{d}$ metric [20]) to make $v = 1$ and diam $\mathcal{P}(X)$ roughly independent of N, but this metric would not give enough weight to localised changes, nor extend to one for infinite networks.

So we see that already for the simple example of a one-parameter family of Bernoulli distributions, none of the above metrics give a speed of parameter dependence uniform in the network size, unless one scales them, but then either the diameter of the space of probabilities goes to 0 as the network size goes to infinity or the metric does not give much weight to localised changes or fails to extend to infinite networks.

In contrast, for the metric (1), the supremum for $D(p_\lambda, p_{\lambda+\varepsilon})$ is realised by $f = n_1$, the number of 1s; $p_\lambda(n_1) = N\lambda$, $\Delta_s(n_1) = 1$, so $\|n_1\| = N$ and $D(p_\lambda, p_{\lambda+\varepsilon}) = |\varepsilon|$; thus $v = 1$ independently of N. The metric gives diam $\mathcal{P}(X) = \Omega$, it pays attention to localised changes, and it extends to a metric on infinite networks.

An alternative is to use metric $D(\rho,\sigma) = \sup_\Lambda D_V(\rho_\Lambda,\sigma_\Lambda)/|\Lambda|$ over non-empty finite subsets Λ, where $|\Lambda|$ is its size and ρ_Λ and σ_Λ are the marginals of ρ,σ on Λ. I proposed this in 1999, but to obtain results like Theorems 1 and 2 with it I had to introduce a second norm using a reference Markov process, which complicated the picture, so I switched to (1) when I came across the semi-norm (2) in [15] and realised that its dual would work nicely.

3 Proof of Theorem 1

Proof (Theorem 1(a)). Given a solution ρ_0 for P_0, the equation $\rho P = \rho$ for $\rho \in \mathcal{P}(X)$ can be rewritten as

$$(\rho - \rho_0)(I - P_0) = \rho(P - P_0), \tag{9}$$

with both $\rho - \rho_0$ and $\rho(P - P_0)$ in $\mathcal{Z}$. By hypothesis, $Z := (I - P_0)^{-1}$ exists on $\mathcal{Z}$ and is bounded, so the solutions of (9) are precisely the fixed points of the map $T : \rho \mapsto \rho_0 + \rho\Delta Z$ on $\mathcal{P}$, where $\Delta = P - P_0$. For $\|\Delta Z\|_{\mathcal{Z}} < 1$, T is a contraction on $\mathcal{P}$. Since $\|\Delta Z\| \le \|\Delta\|\|Z\| = K\delta$, $\delta < 1/K$ suffices. Then T has the unique fixed point

$$\rho = \rho_0 \sum_{n\ge 0} (\Delta Z)^n \tag{10}$$

(the sum converges because the partial sums are a Cauchy sequence and $\mathcal{P}(X)$ is complete). The change in ρ is

$$\rho - \rho_0 = \rho_0 \Delta Z \sum_{n\geq 0} (\Delta Z)^n, \tag{11}$$

which is bounded by

$$\|\rho - \rho_0\| \leq \frac{\|\rho_0 \Delta\| \|Z\|}{1 - \|\Delta Z\|} \leq \frac{\beta}{K^{-1} - \delta}, \tag{12}$$

as claimed. The formula $(I-P)^{-1} = Z\sum_{n\geq 0}(\Delta Z)^n$ provides an inverse for $I-P$ on $\mathcal{Z}$ whenever $\|\Delta Z\| < 1$. In particular $I-P$ is invertible for $\delta < K^{-1}$, and then $\|(I-P)^{-1}\| \leq (K^{-1}-\delta)^{-1}$. □

Before proving parts (b) and (c), precise definitions of continuous and C^1 dependence of a discrete-time Markov process on parameters are introduced.

Definition 2. *P_λ depends continuously on λ at $\lambda_0 \in M$ if $\|P_\lambda - P_{\lambda_0}\|_{\mathcal{Z}} \to 0$ as $\lambda \to \lambda_0$, and for all $\rho \in \mathcal{P}(X)$, $\|\rho(P_\lambda - P_{\lambda_0})\|_{\mathcal{Z}} \to 0$ as $\lambda \to \lambda_0$.*

Note that it is enough to check the second condition at a single ρ_0 because

$$\|\rho(P-P_0)\| \leq \|\rho_0(P-P_0)\| + \|(\rho-\rho_0)(P-P_0)\|,$$

of which the second term is bounded by $\|\rho - \rho_0\| \|P - P_0\|$.

Proof (Theorem 1(b)). If P_λ depends continuously on λ then (12) establishes continuity of ρ_λ at λ_0. The same is true at any λ for which $(I-P)^{-1}$ has bounded inverse. Since this is an open property, we obtain the open set Λ with unique and continuously dependent stationary measure, and the absence of a bounded inverse for any λ on its boundary. □

Definition 3. *P_λ depends C^1 on λ in a differentiable manifold M if (i) there exists $P'_\lambda : \mathcal{P}(X) \times T_\lambda M \to \mathcal{Z}$ such that $\|P_{\lambda+\varepsilon} - P_\lambda - P'_\lambda \varepsilon\|_{\mathcal{Z}} = o(|\varepsilon|)$ as $|\varepsilon| \to 0$ for tangent vectors ε to M, in a local chart for M, (ii) for all $\rho \in \mathcal{P}(X)$ then $\|\rho P_{\lambda+\varepsilon} - \rho P_\lambda - \rho P'_\lambda \varepsilon\|_{\mathcal{Z}} = o(|\varepsilon|)$, and (iii) P' depends continuously on λ in the sense of Definition 2 extended to maps from $\mathcal{P}(X) \times TM \to \mathcal{Z}$.*

Proof (Theorem 1(c)). If P_λ depends C^1 on λ then

$$\Delta := P_{\lambda+\varepsilon} - P_\lambda = P'\varepsilon + o(\varepsilon)$$

and (9) shows that ρ_λ is differentiable at λ_0 with derivative

$$\frac{d\rho}{d\lambda} = \rho \frac{dP}{d\lambda} (I-P)^{-1}_{\mathcal{Z}}, \tag{13}$$

and hence at any λ for which $I-P$ on $\mathcal{Z}$ is invertible, as claimed. To prove the derivative is continuous, first note that

$$(I-P)^{-1} - (I-P_0)^{-1} = Z \sum_{n\geq 1} (\Delta Z)^n,$$

so is bounded by $\frac{K\delta}{K^{-1}-\delta}$, which proves continuity of $(I-P)^{-1}_{\mathcal{Z}}$. Then continuity of $\frac{d\rho}{d\lambda}$ follows from continuity of the terms out of which (13) is composed. □

4 Proof of Theorem 2

Proof (Theorem 2). Given $r < 1$, $C \geq 1$ such that $\|P_0^n\| \leq Cr^n \ \forall n \geq 0$, introduce an *adapted norm* on $\mathcal{Z}$:

$$\|\mu\|_r = \sup_{n \geq 0} \|\mu P_0^n\| r^{-n}.$$

It is equivalent to the original norm:

$$\|\mu\| \leq \|\mu\|_r \leq C\|\mu\|, \tag{14}$$

and it is contracted by P_0:

$$\|\mu P_0\|_r = \sup_{n \geq 0} \|\mu P_0^{n+1}\| r^{-n} \leq r\|\mu\|_r.$$

From (14), for any linear operator P on $\mathcal{Z}$ we have $\|P\| \leq C\|P\|_r$ and $\|P\|_r \leq C\|P\|$. Then $\|P - P_0\| = \delta$ implies $\|P - P_0\|_r \leq C\delta$, so $\|P\|_r \leq r + C\delta$. So $\|P^n\|_r \leq (r + C\delta)^n$ and hence $\|P^n\| \leq C(r + C\delta)^n$. The factor $r + C\delta$ is close to r for δ small enough, and the prefactor has not changed. □

A similar proof can be done in continuous time, with respect to the following definition.

Definition 4. *For a continuous-time Markov process with rate matrix Q, a stationary measure ρ* attracts exponentially with rate $\kappa > 0$ *if there is a* prefactor $C \in \mathbb{R}$ *such that $D(e^{Qt}\sigma, \rho) \leq Ce^{-\kappa t}$ for all close enough σ to ρ.*

Note that in both discrete and continuous time the prefactor $C \geq 1$ because the case $n = 0$ or $t = 0$ is included. To make the factor r or rate κ unique, one could take the infimum of such r or supremum of such κ, but the prefactor C might go to infinity there, so I prefer not to. Also if D comes from a norm on $\mathcal{Z}$ there is no need to restrict σ to be close to ρ.

Existence of an exponentially attracting stationary measure (with some technical extra conditions) is known as "geometric ergodicity" by probabilists. Ergodic theorists would say an exponentially attracting stationary measure is "exponentially mixing".

For example, for finite state Markov processes, exponentially attracting stationary measure holds in continuous-time as soon as there is a unique communicating component. In discrete-time it holds if in addition the unique communicating component is aperiodic, i.e. there exists $T \in \mathbb{N}$ such that it is possible to go from any state to any state in the same time T. These results are often proved by showing that P or e^Q is a contraction in total variation metric (an approach of Dobrushin, e.g. [4]), or projective metric [3].

5 Examples

The theory applies to several relevant classes of example. They include families of exponentially ergodic PCA such as kinetic Ising models, Toom's voter model, and directed percolation, in the regimes of unique stationary measure.

Definition 5. *A PCA with independent transition probability measures $p_r(x)$ for the state at site $r \in S$ at time $n+1$ as functions of the state x of the whole network at time n is* weakly dependent *if there is a* dependency matrix $k = (k_{rs})_{r,s\in S}$ *of non-negative reals such that (i) for all $r, s \in S$ and states x, x' with $x_t = x'_t$ for all $t \neq s$,*

$$D_{T_r}(p_r(x), p_r(x')) \leq k_{rs} d_s(x_s, x'_s), \tag{15}$$

where D_{T_r} is transportation metric on $\mathcal{P}(X_r)$, and (ii) there is $C \geq 1, \gamma \in [0,1)$ such that for all $n \geq 0$, $\|k^n\|_\infty \leq C\gamma^n$ (where $\|k\|_\infty = \sup_{r\in S} \sum_{s\in S} k_{rs}$).

If the X_s carry discrete metric, then (15) reduces to $D_{V_r}(p_r(x), p_r(x')) \leq k_{rs}$. The usual definition (e.g. [17]) requires only $\|k\|_\infty < 1$, but extending to (ii) allows more cases (somewhat analogous to [18]). In particular, one can start from independent Markov chains with strong dependence on initial state provided they mix with uniform bounds, and then allow them to interact weakly. A more or less equivalent alternative to (ii) is to assume there is some matrix norm such that $\|k\| < 1$ [7], but factors of network size can enter this approach unless it is chosen to be an operator norm, uniform in network size.

Definition 6. *A family of weakly dependent PCA with transition probability measures $p_r(x, \lambda)$ for updating the state on site r given state x and parameter value λ* depends continuously on parameters *if there is a function g such that*

$$D_{T_r}(p_r(x,\lambda), p_r(x,\lambda')) \leq g(\lambda, \lambda') \to 0 \text{ as } \lambda' \to \lambda$$

and for all $r, s \in S$, states x, x' which differ only on site s and functions $f : X_r \to \mathbb{R}$,

$$(p_r(x',\lambda') - p_r(x,\lambda') - p_r(x',\lambda) + p_r(x,\lambda))(f) \leq g(\lambda,\lambda') k_{rs} d_s(x_s, x'_s) \|f\|_{L_r}.$$

To save space, I don't formulate the C^1 or continuous-time cases here.

Theorem 3. *If P is the transition matrix for a weakly dependent PCA then it is an eventual contraction in metric (1).*

Proof. As in [17], for x, x' agreeing off $s \in S$, let τ_r, $r \in S$, be an optimal joining for $p_r(x)$ and $p_r(x')$. Given $f \in F$, $Pf(x) - Pf(x')$

$$= \int (f(\xi) - f(\xi')) \prod_{r\in S} \tau_r(d\xi_r, d\xi'_r) \leq \sum_{r\in S} \int \Delta_r(f) d_r(\xi_r, \xi'_r) \tau_r(d\xi_r, d\xi'_r)$$
$$= \sum_r \Delta_r(f) D_{T_r}(p_r(x), p_r(x')) \leq \sum_r \Delta_r(f) k_{rs} d_s(x_s, x'_s).$$

So $\Delta_s(Pf) \leq \sum_{r\in S} \Delta_r(f) k_{rs}$. Let $\Delta(f) = (\Delta_s(f))_{s\in S}$ and $\leq$ the component-wise partial order on $\mathbb{R}^S$, then $\Delta(Pf) \leq \Delta(f)k$, and iteration yields $\Delta(P^n f) \leq \Delta(f)k^n$ for all $n \geq 0$. Summing over components gives $\|P^n f\| \leq \|k^n\|_\infty \|f\| \leq C\gamma^n \|f\|$. So, using (5), $\|P^n\|_{\mathcal{Z}} \leq C\gamma^n$. Thus P is an eventual contraction of $\mathcal{P}$, so has a unique fixed point ρ_0 and ρ_0 attracts exponentially. $\sum_{n\geq 0} P^n < \infty$ provides an inverse for $I - P$ and shows that $\|(I-P)^{-1}\| \leq 1 + \frac{C\gamma}{1-\gamma}$. □

Theorem 4. *For a family P_λ of weakly dependent PCA depending continuously on λ, and any λ_0 (sometimes shortened to 0), then $\|\rho_0(P_\lambda - P_0)\|$ and $\|P_\lambda - P_0\|$ go to zero as $\lambda \to \lambda_0$.*

Proof. Given $f \in F$, $\rho_0(P_\lambda - P_0)f \le \sup_{x\in X}((P_\lambda - P_0)f)(x)$. Given $x \in X$, take optimal joinings τ_r of $p_r(x,\lambda)$ with $p_r(x,\lambda_0)$ for $r \in S$. Then

$$(P_\lambda - P_0)f(x) = \int (f(\xi) - f(\xi'))\prod_r \tau_r(d\xi, d\xi') \tag{16}$$
$$\le \sum_r \Delta_r(f) D_{T_r}(p_r(x,\lambda), p_r(x,\lambda_0)) \le g(\lambda,\lambda_0)\|f\|.$$

Next, for x, x' agreeing off site s, by the last assumption of Definition 6,

$$(P_\lambda - P_0)f(x) - (P_\lambda - P_0)f(x') \le \sum_r g(\lambda,\lambda_0)k_{rs}d_s(x_s, x'_s)\Delta_r(f).$$

So $\Delta_s((P_\lambda - P_0)f) \le \sum_r g\Delta_r(f)k_{rs}$, and $\|(P_\lambda - P_0)f\| \le g\|k\|_\infty\|f\|$. □

Thus Theorems 1 and 2 apply to any continuous family P_λ of weakly dependent PCAs.

Theorem 2 leads to bounds for mixing time which are uniform in system size. Given a Markov process on X with exponentially attracting stationary probability measure ρ for a metric D on $\mathcal{P}$, and $\varepsilon > 0$, the *mixing time* $N(\varepsilon)$ is the minimum N such that $D(\sigma P^n, \rho) \le \varepsilon$ for all $n \ge N$ and $\sigma \in \mathcal{P}$ (beware of an additional factor of system size that is sometimes quoted in mixing time estimates, e.g. [7]). Then $\|P^n\| \le Cr^n$ implies that $N \le \frac{\log \varepsilon/C\Omega'}{\log r}$, where $\Omega' = \operatorname{diam} \mathcal{P}$. In metric (1), $\Omega' = \Omega$, so if Ω is uniform in N I obtain mixing time uniform in the size of the network. Contrast [7], which uses total variation metric and thus the best achieved there is logarithmic in the size.

The above notion of mixing time is sometimes used to make statements about the accuracy of Monte Carlo simulations, but such simulations do not push forward measures. Instead they evaluate empirical measures $\mu_T = \frac{1}{T}\sum_{n=0}^{T-1}\delta_{x^n}$ along simulations $x^n, n \in Z_+$. Thus one needs to ask how far μ_T is from the stationary measure ρ. Typically one obtains an answer proportional to $T^{-\frac{1}{2}}$, which suggests that the convergence is very slow. The more relevant notion of mixing time, however, is the minimum time T for the probability that μ_T deviates more than a prescribed tolerance ε from the stationary measure ρ to be less than a prescribed amount η. For processes satisfying the above hypotheses, large deviation theory [5, 9] can be used to show that

$$P\{\|\mu_T - \rho\| \ge \varepsilon\} \simeq \exp(-\frac{T\varepsilon^2}{\Omega^2(K+\frac{1}{2})}). \tag{17}$$

Thus this is of order η for $T \sim \frac{\Omega^2(K+\frac{1}{2})}{\varepsilon^2}\log\frac{1}{\eta}$, which is uniform in N and only logarithmic in η. I suggest this explains the surprising accuracy of Monte Carlo methods.

In both discrete and continuous time, the results apply to more than just the standard type of PCA or IPS with independent updates of different sites. For example, the Markov transitions can involve simultaneous correlated change in state at a group of sites, as occurs in reaction-diffusion models.

6 Conclusion and Problems for the future

In summary, metric (1) provides a good way to measure the rates at which a stationary probability measure for a Markov process on a large network changes with parameters and attracts other initial probability measures.

Here is a list of some problems for the future:

- Obtain bounds on the spatial and spatio-temporal correlations of the stationary measure (these will follow from spatial hypotheses on the dependency matrix, via [1]).
- Determine how to design a PCA to achieve desired spatiotemporal statistics.
- Find the generic ways that $\|(I-P)^{-1}\|_{\mathcal{Z}}$ can go to infinity.
- Study the case of a level set of conserved quantities for a system with conservation laws. There is no obvious analogue of the results of this paper, because for example for particles diffusing on a lattice the relaxation to equilibrium is a power law in time, not exponential.
- Study the continuation of the set of stationary probability measures, or more generally space-time Gibbs measures, when there is more than one.
- Study the continuation of other spectral projections for Markov processes.
- Develop good computational methods for stationary measures of large networks.
- Study the control of Markov processes on large networks.
- Refine the convergence estimate (17) for Markov chain Monte Carlo methods for multivariate probabilities by computing a rigorous upper bound (cf. [14]).

Acknowledgements

This paper is an extended version of [19], which was published on CD in the proceedings of the European Complex Systems Society Conference 2007. I am grateful to Wolfgang Loehr for pointing out some errors in the proceedings version, which have been corrected here.

References

1. C Baesens, RS MacKay, Exponential localization of linear response in networks with exponentially decaying coupling, Nonlin 10 (1997) 931–940.
2. FG Ball, RK Milne, GF Yeo, Stochastic models for systems of interacting ion channels, IMA J Med Biol 17 (2000) 263–293.
3. G Birkhoff, Extensions of Jentzch's theorem, Am Math Soc 85 (1957) 219–227.
4. P Brémaud, Markov chains (Springer, 1999).
5. A Dembo, O Zeitouni, Large deviation techniques and applications (Springer, 1993, 1998).
6. RL Dobrushin, Prescribing a system of random variables by conditional distributions, Theory Prob Applns 15 (1970) 458–486.
7. M Dyer, LA Goldberg, M Jerrum, Matrix norms and rapid mixing for spin systems, arXiv: math/0702744
8. M Dyer, A Sinclair, E Vigoda, D Weitz, Mixing in time and space for lattice spin systems: a combinatorial view, Rand Structs Algos 24 (2004) 461–479.

9. A Eizenberg, Y Kifer, Large deviations for PCA II, J Stat Phys 117 (2004) 845–889.
10. DM Endres, JE Schindelin, A new metric for probability distributions, IEEE Trans Info Theory 49 (2003) 1858–60.
11. AL Gibbs, FE Su, On choosing and bounding probability metrics, Int Stat Rev 70 (2002) 419–435.
12. GJ Gibson, Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology, Appl Statist 46 (1997) 215–233.
13. O Hernandez-Lerma, JB Lasserre, Markov chains and invariant probabilities (Birkhäuser, 2003).
14. I Kontoyiannis, LA Lastras-Montaño, SP Meyn, Exponential bounds and stopping rules for MCMC and general Markov chains, ACM Internat Conf Proc 180 (2006) article 45.
15. JL Lebowitz, C Maes, ER Speer, Statistical mechanics of probabilistic cellular automata, J Stat Phys 59 (1990) 117–170.
16. TM Liggett, Interacting particle systems (Springer, 1985).
17. C Maes, Coupling interacting particle systems, Rev Math Phys 5 (1993) 457–75.
18. C Maes, SB Shlosman, Ergodocity of probabilitistic cellular automata: a constructive criterion, Commun Math Phys 135 (1991) 233–251; When is an interacting particle system ergodic? Commun Math Phys 151 (1993) 447–466.
19. MacKay RS, Parameter-dependence of Markov processes on large networks, in: ECCS'07 Proceedings (CD, 2007), eds J Jost et al, p41.pdf (12 pages)
20. DS Ornstein, B Weiss, Statistical properties of chaotic systems, Bull Am Math Soc 24 (1991) 11–116.
21. TJ Palmer, A nonlinear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models, Q J Roy Meteor Soc 127 (2001) 279–304.
22. DW Stroock, An introduction to Markov processes (Springer, 2005).
23. AL Toom, NB Vasilyev, ON Stavskaya, LG Mityushin, GL Kurdyumov, SA Pirogov, Discrete local Markov systems, in: Stochastic cellular systems, ergodicity, memory and morphogenesis, eds RL Dobrushin, VI Kryukov, AL Toom (Manchester Univ Press, 1990) 1–182.
24. Vershik AM, Kantorovich metric: initial history and little-known applications, J Math Sci 133 (2006) 1410–7.