Mixing Time of Markov Chains, Dynamical Systems and Evolution

Ioannis Panageas
Georgia Institute of Technology
ioannis@gatech.edu

Nisheeth K. Vishnoi
École Polytechnique Fédérale de Lausanne (EPFL)
nisheeth.vishnoi@epfl.ch

Abstract

In this paper we study the mixing time of evolutionary Markov chains over populations of a fixed size (N) in which each individual can be one of m types. These Markov chains have the property that they are guided by a dynamical system from the m-dimensional probability simplex to itself. Roughly, given the current state of the Markov chain, which can be viewed as a probability distribution over the m types, the next state is generated by applying this dynamical system to this distribution, and then sampling from it N times. Many processes in nature, from biology to sociology, are evolutionary and such chains can be used to model them. In this study, the mixing time is of particular interest as it determines the speed of evolution and whether the statistics of the steady state can be efficiently computed. In a recent result [Panageas, Srivastava, Vishnoi, Soda, 2016], it was suggested that the mixing time of such Markov chains is connected to the geometry of this guiding dynamical system. In particular, when the dynamical system has a fixed point which is a global attractor, then the mixing is fast. The limit sets of dynamical systems, however, can exhibit more complex behavior: they could have multiple fixed points that are not necessarily stable, periodic orbits, or even chaos. Such behavior arises in important evolutionary settings such as the dynamics of sexual evolution and that of grammar acquisition. In this paper we prove that the geometry of the dynamical system can also give tight mixing time bounds when the dynamical system has multiple fixed points and periodic orbits. We show that the mixing time continues to remain small in the presence of several unstable fixed points and is exponential in N when there are two or more stable fixed points. As a consequence of our results, we obtain a phase transition result for the mixing time of the sexual/grammar model mentioned above. We arrive at the conclusion that in the interesting parameter regime for these models, i.e., when there are multiple stable fixed points, the mixing is slow. Our techniques strengthen the connections between Markov chains and dynamical systems and we expect that the tools developed in this paper should have a wider applicability.

1 Introduction

Evolutionary Markov chains and mixing time

In this paper we study Markov chains that arise in the context of evolution and which have also been used to model a wide variety of social, economical and cultural phenomena, see [17]. Typically, in such Markov chains, each state consists of a population of size N where each individual is of one of m *types*. Thus, the state space Ω has size $\binom{N+m-1}{m-1}$. At a very high level, in each iteration, the different types in the current generation reproduce according to their *fitnesses*, the reproduction could be *asexual* or *sexual* and have *mutations* that transform one type into another. This gives rise to an intermediate population that is subjected to the force of *selection*; a sample of size N is selected giving us the new generation. The specific way in which each of the *reproduction*, *mutation* and *selection* steps happen determine the transition matrix of the

corresponding Markov chain. The size of the population (N), the number of types (m), the fitness of each type ($\{a_i \geq 0 : i \in [m]\}$), and the probabilities of mutation of one type to another ($\{Q_{ij} \geq 0 : i, j \in [m]\}$) are the parameters of the model. If we make the natural assumption that all the fitnesses are strictly positive and there is a non-zero probability of mutating from any type to the other, $Q_{ij} > 0$ for all $i, j \in [m]$, then the underlying chain is ergodic and has a unique steady state.

Most questions in evolution reduce to understanding the statistical properties of the steady state of an evolutionary Markov chain and how it changes with its parameters. However, in general, there seems to be no way to compute the desired statistical properties other than to sample from (close to) the steady state distribution by running the Markov chain for sufficiently long [5]. In the chains of interest, while there is an efficient way to sample the next state given the current state, typically, the state space is huge[1] and the efficiency of such a sampling algorithm rests on the number of iterations required for the chain to be close to its steady state. This is captured by the notion of its *mixing time*. The mixing time of a Markov chain, t_{mix}, is defined to be the smallest time t such that for all $x \in \Omega$, the distribution of the Markov chain starting at x after t-time steps is within an ℓ_1-distance of $1/4$ of the steady state.[2] Apart from dictating the computational feasibility of sampling procedures, the mixing time also gives us the number of generations required to reach a steady state; an important consideration for validating evolutionary models [5, 23]. However, despite the importance of understanding when an evolutionary Markov chain mixes fast (i.e., is significantly smaller than the size of the state space), until recently, there has been a lack of rigorous mixing time bounds for the full range of evolutionary parameters in even in the simplest of stochastic evolutionary models; see [7–9] for results under restricted assumptions and [5, 24] for an extended discussion on mixing time bounds in evolutionary Markov chains.

The expected motion of a Markov chain

In a recent result [19], a new approach for bounding the mixing time of such Markov chains was suggested. Towards this, it is convenient to think of each state of an evolutionary Markov chain as a vector which captures the fraction of each type in the current population. Thus, each state is a point the m-dimensional probability simplex Δ_m,[3] and we can think of $\Omega \subseteq \Delta_m$. If $\boldsymbol{X}^{(t)}$ is the current state, then we define the *expected motion* of the chain at $\boldsymbol{X}^{(t)}$ to be the function

$$f(\boldsymbol{X}^{(t)}) \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{X}^{(t+1)} | \boldsymbol{X}^{(t)}\right]$$

where the expectation is over one step of the chain. Notice that while the domain of f is Ω, its range could be a larger subset of Δ_m. *What can the expected motion of a Markov chain tell us about the mixing time of a Markov chain?* Of course, without imposing additional structure on the Markov chain, we do not expect a very interesting answer. However, [19] suggested that, the expected motion can be helpful in establishing mixing time bounds, at least in the context of evolutionary dynamics. The first observation is that, while in the case of general Markov chains, the expected motion function is only defined at a subset of Δ_m, in the case of evolutionary Markov chains, the expected motion turns out to be a *dynamical system*; defined on *all* points of Δ_m. Further, the Markov chain can be recovered from the dynamical system: it can be shown that given a state $\boldsymbol{X}^{(t)}$ of the Markov chain, one can generate $\boldsymbol{X}^{(t+1)}$ *equivalently* by computing the probability distribution $f(\boldsymbol{X}^{(t)})$ and taking N i.i.d. samples from it. Subsequently, their main result is to prove that if this dynamical

[1] For example, even when $m = 40$ and the population is of size $10,000$, the number of states is more than 2^{300}, i.e., more than the number of atoms in the universe!

[2] It is well-known that if one is willing to pay an additional factor of $\log 1/\varepsilon$, one can bring down the error from $1/4$ to ε for any $\varepsilon > 0$; see [13].

[3] The probability simplex Δ_m is defined to be $\{p \in \mathbb{R}^m : p_i \geq 0 \; \forall i, \; \sum_i p_i = 1\}$.

system has a *unique stable fixed point* and also all the trajectories converge to this point, then the evolutionary Markov chain mixes rapidly. Roughly, this is achieved by using the geometry of the dynamical system around this unique fixed point to construct a *contractive coupling*. As an application, this enabled them to establish rapid mixing for evolutionary Markov chains in which the reproduction is *asexual*.

What if the limit sets of the expected motion are complex: multiple fixed points – some stable and some unstable, or even periodic orbits? Not only are these natural mathematical questions given the previous work, such behavior arises in several important evolutionary settings; e.g., in the case when the reproduction is sexual (see [3, 16] and Chapter 20 in [**?**]) and an equivalent model for how children acquire grammar [11, 18]. While we describe these models later, we note that, as one changes the parameters of the model, the limit sets of the expected motion can exhibit the kind of complex behavior mentioned above and a finer understanding of how they influence the mixing time is desired.

Our contribution

In this paper we introduce prove that the geometry of the dynamical system can also give tight mixing time bounds when the dynamical system has multiple fixed points and periodic orbits. This completes the picture left open by the previous work. Recall that [19] proved that when there is a unique stable fixed point, then the mixing time is about $O(\log N)$ when N is large compared to the parameters of the model. We complement their result by proving the following mixing time bounds which depend on the structure of the limit sets of the expected motion:

- One stable fixed point and multiple unstable fixed points – the mixing time is $O(\log N)$, see Theorem 6.
- Multiple stable fixed points – the mixing time is $e^{\Omega(N)}$, see Theorem 7.
- Periodic orbits – the mixing time is $e^{\Omega(N)}$, see Theorem 8.

Thus, we can prove that despite the presence of unstable fixed points the mixing time continues to remain small. On the other hand, if there are two or more stable fixed points, the mixing time can undergo a phase transition and become exponential in N.

As an application, we characterize the mixing time of the dynamics of *grammar acquisition* (or, as explained later, sexual evolution). This Markov chain attempts to model a fascinating and important problem in linguistics; to understand the mechanism by which a child acquires the capacity to comprehend a language and effectively communicate [10, 15]. Here, a parameter of interest is the *mutation rate* τ which is to be thought of as quantifying the *error of learning*; see Section 2.1. Corresponding to this, the probabilities of mutation $Q_{ij} = \tau$ for all $i \neq j$ and $Q_{ii} = 1-(m-1)\tau$. We first prove that there is a critical value where the expected motion dynamical system goes through a bifurcation from multiple stable fixed points to one stable fixed point. Our main results then imply that for $\tau < \tau_c$ the mixing time is exponential in N and for $\tau > \tau_c$ it is $O(\log N)$, see Theorem 9. Thus, we arrive at the conclusion that, in the interesting parameter regime for an important and natural dynamics, i.e., when there is a stable fixed point other than the uniform one, the mixing is very slow.

Technically, there have been several influential works in the probability literature that use dynamical systems to analyze stochastic processes, see for example [2, 14, 21, 25]. While the techniques used in these results bear some similarity to ours, to the best of our knowledge, ours is the first paper which studies the question of how the *mixing time* of a Markov chain behaves as a function of the guiding dynamical system formally.

Organization of the paper

The rest of the paper is organized as follows. In Section 2 we present the formal statement of our main theorems and the model of grammar acquisition/sexual evolution. In Section 4, we present an overview of the proofs of our main theorem. The proof of Theorems 6, 7, 8 and 9 appear in Sections 5, 6, 7 8 respectively.

2 Formal statement of our results

In this section we present formal statements of our main results. We begin by introducing the required notation and preliminaries.

Notation

We use boldface letters, e.g., $\boldsymbol{x}$, to denote column vectors (points), and denote a vector's i^{th} coordinate by x_i. We use $\boldsymbol{X}$ and $\boldsymbol{Y}$ (often with time superscripts and coordinate subscripts as appropriate) to denote random vectors. For a function $f : \Delta_m \to \Delta_m$, by f^n we denote the composition of f with itself n times, namely $\underbrace{f \circ f \circ \cdots \circ f}_{n \text{ times}}$. We use $J_f[\boldsymbol{x}]$ to denote the Jacobian matrix of f at the point $\boldsymbol{x}$. When the function f is clear from the context, we omit the subscript and simply denote it by $J[\boldsymbol{x}]$. Similarly, we sometimes use $J^n[\boldsymbol{x}]$ to denote the Jacobian of f^n at $\boldsymbol{x}$. We denote by $\mathrm{sp}(A)$ the spectral radius of a matrix A and by $(A\boldsymbol{x})_i$ the sum $\sum_j A_{ij}x_j$.

Dynamical Systems

Let $\boldsymbol{x}^{(t+1)} = f(\boldsymbol{x}^{(t)})$ be a *discrete time* dynamical system with update rule $f : \Delta_m \to \Delta_m$. The point $\boldsymbol{z}$ is called a *fixed point* of f if $f(\boldsymbol{z}) = \boldsymbol{z}$. We call a fixed point $\boldsymbol{z}$ *stable* if, for the Jacobian $J[\boldsymbol{z}]$ of f, it holds that $\mathrm{sp}(J[\boldsymbol{z}]) < \rho < 1$. A sequence $(f^t(\boldsymbol{x}^{(0)}))_{t\in\mathbb{N}}$ is called a *trajectory* of the dynamics with $x^{(0)}$ as starting point. A common technique to show that a dynamical system converges to a fixed point is to construct a function $P : \Delta_m \to \mathbb{R}$ such that $P(f(\boldsymbol{x})) > P(\boldsymbol{x})$ unless $\boldsymbol{x}$ is a fixed point. We call P a *potential* function. One of our results deals with dynamical systems that have stable periodic orbits.

Definition 1. *$C = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is called a periodic orbit of size k if $\boldsymbol{x}_{i+1} = f(\boldsymbol{x}_i)$ for $1 \le i \le k-1$ and $f(\boldsymbol{x}_k) = \boldsymbol{x}_1$. If $\mathrm{sp}\left(J_{f^k}[\boldsymbol{x}_1]\right) < \rho < 1$, we call C a* stable *periodic orbit (we also use the terminology stable limit cycle).*

Remark 1. *Since $f : \Delta_m \to \Delta_m$ and hence $\sum_i f_i(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \Delta_m$, if we define $h_i(\boldsymbol{x}) = \frac{f_i(\boldsymbol{x})}{\sum_i f_i(\boldsymbol{x})}$ so that $h(\boldsymbol{x}) = f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Delta_m$, we get that $\sum_i \frac{\partial h_i(\boldsymbol{x})}{\partial x_j} = 0$ for all $j \in [m]$. This means without loss of generality we can assume that the Jacobian $J[\boldsymbol{x}]$ of f has $\mathbf{1}^\top$ (the all-ones vector) as a left eigenvector with eigenvalue 0.*

The definition below quantifies the instability of a fixed point as is standard in the literature. Essentially, an α unstable fixed point is repelling in any direction.

Definition 2. *Let $\boldsymbol{z}$ be a fixed point of a dynamical system f. The point $\boldsymbol{z}$ is called* α-unstable *if $|\lambda_{\min}(J[\boldsymbol{z}])| > \alpha > 1$ where $\lambda_{\min}$ corresponds to the minimum eigenvalue of the Jacobian of f at the fixed point $\boldsymbol{z}$, excluding the eigenvalue 0 that corresponds to the left eigenvector $\mathbf{1}^\top$.*

Stochastic Evolution

Definition 3. *Given an $f : \Delta_m \to \Delta_m$ which is smooth,*[4] *and a population parameter N, we define a Markov chain called the stochastic evolution guided by f as follows. The state at time t is a probability vector $\boldsymbol{X}^{(t)} \in \Delta_m$. The state $\boldsymbol{X}^{(t+1)}$ is then obtained in the following manner. Define $\boldsymbol{Y}^{(t)} = f(\boldsymbol{X}^{(t)})$. Obtain N independent samples from the probability distribution $\boldsymbol{Y}^{(t)}$, and denote by $\boldsymbol{Z}^{(t)}$ the resulting counting vector over $[m]$. Then*

$$\boldsymbol{X}^{(t+1)} \stackrel{\text{def}}{=} \frac{1}{N}\boldsymbol{Z}^{(t)} \text{ and therefore } \mathbb{E}[\boldsymbol{X}^{(t+1)}|\boldsymbol{X}^{(t)}] = f(\boldsymbol{X}^{(t)}).$$

We call f the expected motion *of the stochastic evolution.*

Definition 4 (**Smooth contractive evolution**)**.** *A function $f : \Delta_m \to \Delta_m$ is said to be a* smooth contractive evolution *if it is* smooth[4]*, has a* unique *fixed point $\boldsymbol{z}$ in the interior of Δ_m, this unique point is* stable*, and, for every $\varepsilon > 0$, there exists an ℓ such that for any $\boldsymbol{x} \in \Delta_m$, it holds that $\left\|f^{\ell}(\boldsymbol{x}) - \boldsymbol{z}\right\|_1 < \varepsilon$ (i.e., f converges to the fixed point).*

The main result in [19] was Theorem 5 below. This theorem gives a bound on the mixing time of a stochastic evolution guided by a function f that satisfies Definition 4.

Theorem 5 (**Main theorem in [19]**)**.** *Let f be a smooth contractive evolution, and let $\mathscr{M}$ be the stochastic evolution guided by f on a population of size N. Then, the mixing time of $\mathscr{M}$ is $O(\log N)$.*

Our Results

Given a dynamical system f, one of the main questions that one can ask is does it converge, and if so, how fast. In general, if the behavior of a system is non-chaotic, we expect the system to reach some steady state (e.g., a fixed point or periodic orbit). This steady state might be some (local) optimum solution to a non-linear optimization problem. Therefore, it is important to understand what traits make a dynamical system converge fast. The existence of many fixed points which are unstable can slow down the speed of convergence of a dynamical system. In the case of the stochastic evolution guided by f, one would expect the existence of multiple unstable fixed points to similarly slow down the mixing time. Nevertheless, our Theorem 6 shows rapid mixing in the presence of α-unstable fixed points. Additionally, we change the assumption *convergence to the fixed point* in 5 to the assumption that for all $\boldsymbol{x} \in \Delta_m$ the limit $\lim_{t\to\infty} f^t(\boldsymbol{x})$ exists and is equal to some fixed point $\boldsymbol{z}$, i.e., as in 5, there are no limit cycles.

Theorem 6. *Let $f : \Delta_m \to \Delta_m$ be twice differentiable in the interior of Δ_m with bounded second derivative. Assume that $f(\boldsymbol{x})$ has a finite number of fixed points $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_l$ in the interior, where $\boldsymbol{z}_0$ is a stable fixed point, i.e., $\mathrm{sp}(J[\boldsymbol{z}_0]) < \rho < 1$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_l$ are α-unstable fixed points ($\alpha > 1$). Furthermore, assume that $\lim_{t\to\infty} f^t(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \Delta_m$. Then, the stochastic evolution guided by f has mixing time $O(\log N)$.*

In our second result, we allow f to have multiple stable fixed points (in addition to any number of unstable fixed points). For this setting, we prove that the stochastic evolution guided by f has mixing time $e^{\Omega(N)}$. Our phase transition result on a linguistic/sexual evolution model discussed in Section 2.1 relies crucially on Theorem 7.

[4]For our purposes, we call a function f is *smooth* if it is twice differentiable in the relative interior of Δ_m with bounded second derivative.

Theorem 7. *Let $f : \Delta_m \to \Delta_m$ be continuously differentiable in the interior of Δ_m. Assume that $f(\boldsymbol{x})$ has at least two stable fixed points in the interior $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_l$, i.e., $\mathrm{sp}(J[\boldsymbol{z}_i]) < \rho_i < 1$ for $i = 1, 2, \ldots, l$. Then, the stochastic evolution guided by f has mixing time $e^{\Omega(N)}$.*

Finally, we allow f to have a stable limit cycle. We prove that in this setting the stochastic evolution guided by f has mixing time $e^{\Omega(N)}$. This result seems important for evolutionary dynamics as periodic orbits often appear [20, 22].

Theorem 8. *Let $f : \Delta_m \to \Delta_m$ be continuously differentiable in the interior of Δ_m. Assume that $f(\boldsymbol{x})$ has a stable limit cycle with points $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_s$ of size $s \geq 2$ in the sense that $\mathrm{sp}(\prod_{i=1}^{s} J[\boldsymbol{w}_{s-i+1}]) < \rho < 1$. Then the stochastic evolution guided by f has mixing time $e^{\Omega(N)}$.*

2.1 Dynamics of grammar acquisition and sexual evolution

We begin by describing the evolutionary processes for grammar acquisition and sexual evolution. As we will explain, the two turn out to be identical and hence we primarily focus on the model for grammar acquisition in the remainder of the paper.

The starting point of the model is Chomsky's *Universal Grammar* theory [4].[5] In his theory, language learning *is facilitated by a predisposition that our brains have for certain structures of language*. This universal grammar (UG) is believed to be innate and embedded in the neuronal circuitry. Based on this theory, an influential model for how children acquire grammar was given by appealing to evolutionary dynamics for infinite and finite populations respectively in [18] and [11]. We first describe the infinite population model, which is a dynamical system that guides the stochastic, finite population model. Each individual speaks exactly one of the m grammars from the set of inherited UGs $\{G_1, \ldots, G_m\}$; denote by x_i the fraction of the population using G_i. The model associates a *fitness* to every individual on the basis of the grammar she *and others* use. Let A_{ij} be the probability that a person who speaks grammar j understands a randomly chosen sentence spoken by an individual using grammar i. This can be viewed as the fraction of sentences according to grammar i that are also valid according to grammar j. Clearly, $A_{ii} = 1$. The pairwise *compatibility* between two individuals speaking grammars i and j is $B_{ij} \stackrel{\text{def}}{=} \frac{A_{ij}+A_{ji}}{2}$, and the fitness of an individual using G_i is $f_i \stackrel{\text{def}}{=} \sum_{j=1}^{m} x_j B_{ij}$, i.e., the probability that such an individual is able to meaningfully communicate with a randomly selected member of the population.

In the reproduction phase each individual produces a number of offsprings proportional to her fitness. Each child speaks one grammar, but the exact learning model can vary and allows for the child to incorrectly learn the grammar of her parent. We define the matrix Q where the entry Q_{ij} denotes the probability that the child of an individual using grammar i learns grammar j (i.e. Q is column stochastic matrix); once a child learns a grammar it is fixed and she does not later use a different grammar. Thus, the frequency x_i' of the individuals that use grammar G_i in the next generation will be

$$x_i' = g_i(\boldsymbol{x}) \stackrel{\text{def}}{=} \sum_{j=1}^{m} \frac{Q_{ji} x_j (B\boldsymbol{x})_j}{\boldsymbol{x}^\top B \boldsymbol{x}}$$

(with $g : \Delta_m \mapsto \Delta_m$ encoding the update rule). Nowak et al. [18] study the symmetric case, i.e., $B_{ij} = b$ and $Q_{ij} = \tau \in (0, 1/m]$ for all $i \neq j$ and observe a threshold: When τ, which can be thought of as quantifying the *error of learning or mutation*, is above a critical value, the only stable fixed point is the uniform distribution (all $1/m$) and below it, there are multiple stable fixed points.

[5] Like any important problem in the sciences, Chomsky's theory is not uncontroversial; see [10] for an in-depth discussion.

Finite population models can be derived from the linguistic dynamics in a standard way. We describe the Wright-Fisher finite population model for the linguistic dynamics. The population size remains N at all times and the generations are non-overlapping. The current state of the population is described by the frequency vector $\boldsymbol{X}^{(t)}$ at time t which is a random vector in Δ_m and notice also that the population that uses G_i is $NX_i^{(t)}$. How does one generate $\boldsymbol{X}^{(t+1)}$? To do this, in the replication (R) stage, one first replaces the individuals that speak grammar G_i in the current population by $NX_i^{(t)}(B(N\boldsymbol{X}^{(t)}))_i$ and the total population has size $N^2\boldsymbol{X}^{(t)\top}B\boldsymbol{X}^{(t)}$.[6] In the selection (S) stage, one selects N individuals from this population by sampling independently with replacement. Since the evolution is error prone, in the mutation (M) stage, the grammar of each individual in this intermediate population is mutated independently at random according to the matrix Q to obtain frequency vector $\boldsymbol{X}^{(t+1)}$. Given these rules, note that

$$\mathbb{E}[\boldsymbol{X}^{(t+1)}|\boldsymbol{X}^{(t)}] = g(\boldsymbol{X}^{(t)}).$$

In other words, in expectation, fixing $\boldsymbol{X}^{(t)}$, the next generation's frequency vector $\boldsymbol{X}^{(t+1)}$ is exactly $g(\boldsymbol{X}^{(t)})$, where g is the linguistic dynamics. Of course, this holds only for one step of the process. This process is a Markov chain with state space $\{(y_1,\ldots,y_m) : y_i \in \mathbb{N}, \sum_i y_i = N\}$ of size $\binom{N+m-1}{m-1}$. If $Q > 0$ then it is ergodic (i.e., it is irreducible and aperiodic) and thus has a unique stationary distribution. In our analysis, we consider the symmetric case as in Nowak et al. [18], i.e., $B_{ij} = b$ and $Q_{ij} = \tau \in (0, 1/m]$ for all $i \neq j$.

Note that the linguistics model described above can also be seen as a (finite population) sexual evolution model: Assume there are N individuals and m types. Let $\boldsymbol{Y}^{(t)}$ be a vector of frequencies at time t, where $\boldsymbol{Y}_i^{(t)}$ denotes the fraction of individuals of type i. Let F be a fitness matrix where F_{ij} corresponds to the number of offspring of type i, if an individual of type i chooses to mate with an individual of type j (assume $F_{ij} \in \mathbb{N}$). At every generation, each individual mates with every other individual. It is not hard to show that the number of offspring after the matings will be $N^2(\boldsymbol{Y}^{(t)\top}F\boldsymbol{Y}^{(t)})$ and there will be $N^2\boldsymbol{Y}_i^{(t)}(F\boldsymbol{Y}^{(t)})_i$ individuals of type i. After the reproduction step, we select N individuals at random with replacement, i.e., we sample an individual of type i with probability $\frac{\boldsymbol{Y}_i^{(t)}(F\boldsymbol{Y}^{(t)})_i}{\boldsymbol{Y}^{(t)\top}F\boldsymbol{Y}^{(t)}}$. Finally in the mutation step, every individual of type i mutates with probability τ (*mutation parameter*) to some type j. Let $F_{ii} = A$, $F_{ij} = B$ for all $i \neq j$ with $A > B$ (this is called homozygote advantage) and set $b = \frac{B}{A} < 1$. It is self-evident that this sexual evolution model is identical with the (finite population) linguistic model described above since both end up having the same reproduction, selection and mutation rule. It holds that $E[\boldsymbol{X}^{(t+1)}|\boldsymbol{X}^{(t)}] = g(\boldsymbol{X}^t)$[7] with

$$g_i(\boldsymbol{x}) = (1-(m-1)\tau)\frac{N^2x_i(B\boldsymbol{x})_i}{N^2(\boldsymbol{x}^\top B\boldsymbol{x})} + \sum_{j\neq i}\tau\frac{N^2x_j(B\boldsymbol{x})_j}{N^2(\boldsymbol{x}^T B\boldsymbol{x})} = (1-m\tau)\frac{x_i(B\boldsymbol{x})_i}{(\boldsymbol{x}^\top B\boldsymbol{x})} + \tau$$

where $B_{ii} = 1, B_{ij} = b$ with $i \neq j$.[8] For the Markov chains described above (symmetric case) we can prove the following phase transition result.

Theorem 9. *There is a critical value τ_c of the* error in learning/mutation *parameter τ such that the mixing time is: (i) $\exp(\Omega(N))$ for $0 < \tau < \tau_c$ and (ii) $O(\log N)$ for $\tau > \tau_c$ where N is the size of the population.*

The theorem below will be used to prove the rapid mixing result for the finite linguistic model when $\tau > \tau_c$. It is used to construct a potential function and show that the deterministic dynamics g converges to fixed points.

[6] Here we assume that B_{ij} is an positive integer and thus $N^2X_i^{(t)}(B\boldsymbol{X}^{(t)})_i$ is an integer since the individuals are whole entities; this can be achieved by scaling and is without loss of generality.

[7] We use same notation for the update rule as before, i.e. g because it turns out to be the same function.

[8] Observe that this rule is invariant under scaling of fitness matrix B.

Theorem 10 (**Baum and Eagon Inequality** [1]). *Let $P(\boldsymbol{x}) = P(\{x_{ij}\})$ be a polynomial with nonnegative coefficients homogeneous of degree d in its variables $\{x_{ij}\}$. Let $\boldsymbol{x} = \{x_{ij}\}$ be any point of the domain $D : x_{ij} \geq 0, \sum_{j=1}^{q_i} x_{ij} = 1, i = 1, \ldots, p, j = 1, \ldots, q_i$. For $\boldsymbol{x} = \{x_{ij}\} \in D$, let $\Xi(\boldsymbol{x}) = \Xi\{x_{ij}\}$ denote the point of D whose i, j-th coordinate is*

$$\Xi(\boldsymbol{x})_{ij} = \left(x_{ij} \frac{\partial P}{\partial x_{ij}}\bigg|_{(\boldsymbol{x})} \right) \cdot \left(\sum_{j=1}^{q_i} x_{ij} \frac{\partial P}{\partial x_{ij}}\bigg|_{(\boldsymbol{x})} \right)^{-1}.$$

Then $P(\Xi(\boldsymbol{x})) > P(\boldsymbol{x})$ unless $\Xi(\boldsymbol{x}) = \boldsymbol{x}$.

3 Preliminaries

Couplings and Mixing Times.

Let $\boldsymbol{p}, \boldsymbol{q} \in \Delta_m$ be two probability distributions on m objects. A *coupling* $\mathscr{C}$ of $\boldsymbol{p}$ and $\boldsymbol{q}$ is a distribution on ordered pairs in $[m] \times [m]$, such that its marginal distribution on the first coordinate is equal to $\boldsymbol{p}$ and that on the second coordinate is equal to $\boldsymbol{q}$. Couplings allow a very useful dual characterization of the total variation distance, as stated in the following well known lemma.

Lemma 11 (**Coupling lemma** [13]). *Let $\boldsymbol{p}, \boldsymbol{q} \in \Delta_m$ be two probability distributions on m objects. Then,*

$$\|\boldsymbol{p} - \boldsymbol{q}\|_{\mathrm{TV}} = \frac{1}{2} \|\boldsymbol{p} - \boldsymbol{q}\|_1 = \min_{\mathscr{C}} \mathbb{P}_{(A,B)\sim\mathscr{C}} [A \neq B],$$

where the minimum is taken over all valid couplings $\mathscr{C}$ of $\boldsymbol{p}$ and $\boldsymbol{q}$.

Definition 12 (**Mixing time** [13]). *Let $\mathscr{M}$ be an ergodic Markov chain on a finite state space Ω with stationary distribution $\boldsymbol{\pi}$. Then, the* mixing time *$t_{\mathrm{mix}}(\varepsilon)$ is defined as the smallest time such that for any starting state $\boldsymbol{X}^{(0)}$, the distribution of the state $\boldsymbol{X}^{(t)}$ at time t is within total variation distance ε of $\boldsymbol{\pi}$. The term mixing time is also used for $t_{\mathrm{mix}}(\varepsilon)$ for a* fixed *values of $\varepsilon < 1/2$.*

A well-known technique for obtaining upper bounds on mixing times is to use the Coupling Lemma above. Suppose $\boldsymbol{X}^{(t)}$ and $\boldsymbol{Y}^{(t)}$ are two evolutions of an ergodic chain $\mathscr{M}$ such that their evolutions are coupled according to some coupling $\mathscr{C}$. Let T be the smallest time such that $\boldsymbol{X}^{(T)} = \boldsymbol{Y}^{(T)}$. If it can be shown that $\mathbb{P}[T > t] \leq 1/4$ for every pair of starting states $(\boldsymbol{X}^{(0)}, \boldsymbol{Y}^{(0)})$, then it follows that $t_{\mathrm{mix}} \stackrel{\mathrm{def}}{=} t_{\mathrm{mix}}(1/4) \leq t$.

Operators, Norms

The following theorem, stated here only in the special case of the $1 \to 1$ norm, relates the spectral radius with other matrix norms.

Theorem 13 (**Gelfand's formula, specialized to the $1 \to 1$ norm** [12]). *For any square matrix A, we have*

$$\mathrm{sp}(A) = \lim_{\ell \to \infty} \left\| A^{\ell} \right\|_{1\to 1}^{1/\ell}.$$

Taylor Theorem (First order Remainder)

Theorem 14. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be differentiable and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$. Then there exists some $\boldsymbol{\xi}$ in the line segment from $\boldsymbol{x}$ to $\boldsymbol{y}$ such that $f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{\xi})(\boldsymbol{y} - \boldsymbol{x})$.*

Concentration

We also mention some standard Chernoff-Hoeffding type bounds that will be used in our later arguments.

Theorem 15 (**Chernoff-Hoeffding bounds [6]**). *Let $Z_1, Z_2, \ldots, Z_N$ be i.i.d. Bernoulli random variables with mean μ. We then have for all $\varepsilon > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mu\right| > \varepsilon\right] \leq 2\exp\left(-2N\varepsilon^2\right).$$

4 Overview of proofs

We begin by explaining the proof technique of Theorem 5 in [19]. In order to prove a bound on the mixing time, the authors constructed a coupling that contracts the distance between two chains. This contraction does not happen at every step, rather at every k steps where k is some constant and depends on the function f. Essentially, it is shown that given two chains $\boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)}$ that are close to the unique fixed point $\boldsymbol{z}$ of f, it holds that

$$\left\|\boldsymbol{X}^{(t+1)} - \boldsymbol{Y}^{(t+1)}\right\|_1 \approx \left\|J[\boldsymbol{z}](\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)})\right\|_1 \leq \|J[\boldsymbol{z}]\|_1 \left\|(\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)})\right\|_1$$

with high probability due to Chernoff bounds. Thus, the ℓ_1 norm of the Jacobian captures the contraction if it indeed exists. However it might be the case that $\|J[\boldsymbol{z}]\|_1 > 1$. On the positive side, using Gelfand's Theorem they were able to show a k-step contraction, since

$$\left\|J^k[\boldsymbol{z}]\right\|_1 \approx (\mathrm{sp}\,(J[\boldsymbol{z}]))^k < \rho^k < 1$$

for some $k \in \mathbb{N}$. Our proofs also use the idea of Gelfand's formula to show contraction/expansion (in Theorems 7 and 6 respectively) and also make use of Theorem 5). Nevertheless, there are important technical barriers that need to be crossed in order to prove our results as explained below.

4.1 Overview of Theorem 6

The main difficulty to prove this theorem is the existence of multiple unstable fixed points in the simplex from which the Markov chain should get away fast. As before, we study the time T required for two stochastic evolutions with arbitrary initial states $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$, guided by some function f, to collide. By the conditions of Theorem 6, function f has a unique stable fixed point $\boldsymbol{z}_0$ with

$$\mathrm{sp}\,(J[\boldsymbol{z}_0]) < \rho < 1.$$

Additionally, it has α-unstable fixed points. Moreover, for all starting points $\boldsymbol{x}_0 \in \Delta_m$, the sequence $(f^t(\boldsymbol{x}_0))_{t\in\mathbb{N}}$ has a limit. We can show that there exists constant c_0 such that $\mathbb{P}[T > c_0 \log N] \leq \frac{1}{4}$, from which it follows that $t_{\mathrm{mix}}(1/4) \leq c_0 \log N$. In order to show collision after $O(\log N)$ steps, it suffices first to run each chain *independently* for $O(\log N)$ steps. We first show that with probability $\Theta(1)$, each chain will reach $B(\boldsymbol{z}_0, \frac{1}{N^{1-\varepsilon}})$ after at most $O(\log N)$ steps, for some $\varepsilon > 0$.[9] As long as this is true, the coupling constructed in [19] can be used to show collision (see Section 3 for the definition of a coupling). To explain why our claim holds, we break the proof into three parts.

[9] $B(\boldsymbol{x}, r)$ denotes the open ball with center $\boldsymbol{x}$ and radius r in ℓ_1, which we call an *r-neighborhood* of $\boldsymbol{x}$.

(a) First, it is shown that as long as the state of the Markov chain is within $o\left(\frac{\log^{2/3}N}{\sqrt{N}}\right)$ in ℓ_1 distance from some α-unstable fixed point $\boldsymbol{w}$, then, with probability $\Theta(1)$, it reaches distance $\Omega\left(\frac{\log^{2/3}N}{\sqrt{N}}\right)$ after $O(\log N)$ steps. Step (a) has the technical difficulty that as long as a chain starts from a $o(\frac{1}{\sqrt{N}})$ distance from an unstable fixed point, the variance of the process dominates the expansion due to the fact the fixed point is unstable.

(b) Assuming (a), we show that with probability $1-\frac{1}{\text{poly}(N)}$ the Markov chain reaches distance $\Theta(1)$ from any unstable fixed point after $O(\log N)$ steps.

(c) Finally, if the Markov chain has $\Theta(1)$ distance from any unstable fixed point (the fixed points have pairwise ℓ_1 distance independent of N, i.e., they are "well separated"), it will reach some $\frac{1}{N^{1-\varepsilon}}$-neighborhood of the stable fixed point $\boldsymbol{z}_0$ exponentially fast (i.e., after $O(\log N)$ steps). For showing (a) and (b), we must prove an expansion argument for $\|f^t(\boldsymbol{x})-\boldsymbol{w}\|_1$ as t increases, where $\boldsymbol{w}$ is an α-unstable fixed point and also taking care of the random perturbations due to the stochastic evolution. Ideally what we want (but is not true) is the following to hold:

$$\left\|f^{t+1}(\boldsymbol{x})-\boldsymbol{w}\right\|_1 \geq \alpha \left\|f^t(\boldsymbol{x})-\boldsymbol{w}\right\|_1,$$

i.e., one step expansion. The first important fact is that f^{-1} is well-defined in a small neighborhood of $\boldsymbol{w}$ due to the Inverse Function Theorem, and it also holds that

$$\left\|f^t(\boldsymbol{x})-\boldsymbol{w}\right\|_1 \approx \left\|J^{-1}[\boldsymbol{w}](f^{t+1}(\boldsymbol{x})-\boldsymbol{w})\right\|_1 \leq \left\|J^{-1}[\boldsymbol{w}]\right\|_1 \left\|f^{t+1}(\boldsymbol{x})-\boldsymbol{w}\right\|_1,$$

where $\boldsymbol{x}$ is in some neighborhood of $\boldsymbol{w}$ and $J^{-1}[\boldsymbol{w}]$ is the *pseudoinverse* of $J[\boldsymbol{w}]$ (see the remark in Section 2). However even if $\boldsymbol{w}$ is α-unstable and $\text{sp}\left(J^{-1}[\boldsymbol{w}]\right) < \frac{1}{\alpha}$, it can hold that $\left\|J^{-1}[\boldsymbol{w}]\right\|_1 > 1$. At this point, we use Gelfand's formula (Theorem 13) as in the proof of [19]. Since $\lim_{t\to\infty}(\|A^t\|_1)^{1/t} \to \text{sp}(A)$, for all $\varepsilon > 0$, there exists a k_0 such that for all $k \geq k_0$ we have

$$\left|\left\|A^k\right\|_1 - (\text{sp}(A))^k\right| < \varepsilon.$$

We use this important theorem to show that for small $\varepsilon > 0$, there exists a k such that

$$\left\|f^t(\boldsymbol{x})-\boldsymbol{w}\right\|_1 \approx \left\|(J^{-1}[\boldsymbol{w}])^k(f^{t+k}(\boldsymbol{x})-\boldsymbol{w})\right\|_1 \leq \frac{1}{\alpha^k}\left\|f^{t+k}(\boldsymbol{x})-\boldsymbol{w}\right\|_1,$$

where we used the fact that

$$\left\|(J^{-1}[\boldsymbol{w}])^k\right\|_1 < (\text{sp}\left(J^{-1}[\boldsymbol{w}]\right))^k - \varepsilon \leq \frac{1}{\alpha^k}.$$

By taking advantage of the continuity of the $J^{-1}[\boldsymbol{x}]$ around the unstable fixed point $\boldsymbol{w}$, we can show expansion for every k steps of the dynamical system. This fact is a consequence of Lemma 16. It remains to show for (a) and (b) how one can handle the perturbations due to the randomness of the stochastic evolution. In particular, if $\left\|\boldsymbol{X}^{(0)}-\boldsymbol{w}\right\|_1$ is $o\left(\frac{1}{\sqrt{N}}\right)$, even with the expansion we have from the deterministic dynamics (as discussed above), variance dominates. We examine case (b) first, which is relatively easy (the *drift* dominates at this step). Due to Chernoff bounds, the difference $\left\|\boldsymbol{X}^{(t+k)}-\boldsymbol{w}\right\|_1 - \left\|f^k(\boldsymbol{X}^{(t)})-\boldsymbol{w}\right\|_1$ is $O\left(\sqrt{\frac{\log N}{N}}\right)$ (this captures the deviation on running the stochastic evolution for k steps vs running the deterministic dynamics for k steps, both starting from $\boldsymbol{X}^{(t)}$) with probability $1-\frac{1}{\text{poly}(N)}$. Since $\left\|\boldsymbol{X}^{(t)}-\boldsymbol{w}\right\|_1$ is $\Omega\left(\frac{\log^{2/3}N}{\sqrt{N}}\right)$, then

$$\left\|\boldsymbol{X}^{(t+k)}-\boldsymbol{w}\right\|_1 \geq (\alpha^k - o_N(1))\left\|\boldsymbol{X}^{(t)}-\boldsymbol{w}\right\|_1.$$

For (a), first we show that with probability $\Theta(1)$, after one step the Markov chain has distance $\Omega(\frac{1}{\sqrt{N}})$ of $\boldsymbol{w}$. This claim just uses properties of the multinomial distribution. After reaching distance $\Omega\left(\frac{1}{\sqrt{N}}\right)$, we can use again the idea of expansion and being careful with the variance and we can show expansion with probability at least $\frac{1}{2}$, every k steps. Then we can show that with probability at least $\frac{1}{\log^{2/3} N}$, distance $\frac{\log^{2/3} N}{\sqrt{N}}$ is reached after $O(\log\log N)$ steps and basically we finish with (b). For (c), we use a couple of modified technical lemmas from [19], i.e., 34, 35 and our Lemma 21. We explain in words below: Let Δ be some compact subset of Δ_m, where we have excluded all the α-unstable fixed points along with some open ball around each unstable fixed point of constant radius. We show that given that the initial state of the Markov chain belongs to Δ, it reaches a $B(\boldsymbol{z}_0, \frac{1}{N^{1-\varepsilon}})$ for some $\varepsilon > 0$ as long as the dynamical system converges for all starting points in Δ (and from Lemma 21, it should converge to the stable fixed point $\boldsymbol{z}_0$). Lemma 34 (which uses Lemma 21) states roughly that the dynamical system converges exponentially fast for every starting point in B to the stable fixed point $\boldsymbol{z}_0$ and Lemma 35 that with probability $1 - \frac{1}{\text{poly}(n)}$ the two chains independently will reach a $\frac{1}{N^{\varepsilon}}$ neighborhood of the stable fixed point $\boldsymbol{z}_0$. Therefore by $(a), (b), (c)$ and the coupling from [19], we conclude the proof of Theorem 6.

4.2 Overview of Theorems 7 and 8

To prove Theorem 8, we make use of Theorem 7, i.e., we reduce the case of the stable limit cycle to the case of multiple stable fixed points. If s is the length of the limit cycle, roughly the bound $e^{\Omega(N)}$ on the mixing time loses a factor $\frac{1}{s}$ compared to the case of multiple stable fixed points. We now present the ideas behind the proof of Theorem 7. First as explained above, we can show contraction after k steps (for some constant k) for the deterministic dynamics around a stable fixed point $\boldsymbol{z}$ with $\text{sp}(\text{J}[\boldsymbol{z}]) < \rho < 1$, i.e.,

$$\left\|f^{t+k}(\boldsymbol{x}) - \boldsymbol{z}\right\|_1 \approx \left\|J^k[\boldsymbol{z}]\right\|_1 \left\|f^t(\boldsymbol{x}) - \boldsymbol{z}\right\|_1 \leq \rho^k \left\|f^t(\boldsymbol{x}) - \boldsymbol{z}\right\|_1.$$

This is Lemma 16 and uses Gelfand's formula, Taylor's theorem and continuity of $J[\boldsymbol{x}]$ where $\boldsymbol{x}$ lies in a neighborhood of the fixed point $\boldsymbol{z}$. Hence, due to the above contraction of the ℓ_1 norm and the concentration of Chernoff bounds, it takes a long time for the chain $\boldsymbol{X}^{(t)}$ to get out of the region of attraction of the fixed point $\boldsymbol{z}$. Technically, the error that aggregates due to the randomness of the stochastic evolution guided by f does not become large due to the convergence of the series $\sum_{i=0}^{\infty} \rho^i$. Hence, we focus on the error probability, namely the probability the stochastic evolution guided by f deviates a lot from the dynamical system with rule f if both have same starting point after one step. Since this probability is exponentially small, i.e., it holds that

$$\left\|f(\boldsymbol{X}^{(0)}) - \boldsymbol{X}^{(1)}\right\|_1 > \varepsilon m$$

with probability at most $2me^{-2\varepsilon^2 N}$, an exponential number of steps is required for the above to be violated. Finally, as we have shown that it takes exponential time to get out of the region of attraction of a stable fixed point $\boldsymbol{z}$ we do the following easy (common) trick. Since the function has at least two fixed points, we start the Markov chain very close to the fixed point that its neighborhood has mass at most $1/2$ in the stationary distribution (this can happen since we have at least 2 fixed points that are well separated). Then, after exponential number of steps, it will follow that the total variation distance between the distribution of the chain and the stationary will be at least $1/4$.

4.3 Overview of Theorem 9

Below we give the necessary ingredients of the proof of Theorem 9. Our previous results, along with some analysis on the fixed points of g (function of Linguistic Dynamics) suffice to show the phase transition result.

To prove Theorem 9, initially we show that the model (finite population) is essentially a stochastic evolution (see Definition 3) guided by g as defined in Section 2.1 and proceed as follows: We prove that in the interval $0 < \tau < \tau_c$, the function g has multiple fixed points whose Jacobian have spectral radius less than 1. Therefore due to Theorem 7 discussed above, the mixing time will be exponential in N. For $\tau = \tau_c$ a bifurcation takes place which results in function g of linguistic dynamics having only one fixed point inside simplex (specifically, the uniform point $(1/m, \ldots, 1/m)$). In dynamical systems, a local bifurcation occurs when a parameter (in particular the mutation parameter τ) change causes two (or more) fixed points to collide or the stability of an equilibrium (or fixed point) to change. To prove fast mixing in the case $\tau_c < \tau \leq 1/m$, we make use of the result in [19] (see Theorem 5). One of the assumptions is that the dynamical system with g as update rule needs to converge to the unique fixed point for all initial points in simplex. To prove convergence to the unique fixed point, we define a Lyapunov function P such that

$$P(g(\boldsymbol{x})) > P(\boldsymbol{x}) \text{ unless } \boldsymbol{x} \text{ is a fixed point.} \tag{1}$$

As a consequence, the (infinite population) linguistic dynamics converge to the unique fixed point $(1/m, \ldots, 1/m)$. To show Equation (1), we use an inequality that dates back in 1967 (see Theorem 10, [1]), which intuitively states the discrete analogue of proving that for a gradient system $\frac{d\boldsymbol{x}}{dt} = \nabla V(\boldsymbol{x})$ it is true that $\frac{dV}{dt} \geq 0$.

5 One stable fixed point

We start by proving some technical lemmas that will be very useful for our proofs. A modified version of the following lemma appeared in [19]. It roughly states that there exists a k (derived from Theorem 13) such that after k steps in the vicinity a stable fixed point $\boldsymbol{z}$, there is as expected a contraction of the ℓ_1 distance between the frequency vector of the deterministic dynamics and the fixed point.

Important Lemmas

Lemma 16 ([19] Modified)**.** *Let $f : \Delta_m \to \Delta_m$ and $\boldsymbol{z}$ be a stable fixed point of f with $\mathrm{sp}(J[\boldsymbol{z}]) < \rho$. Assume that f is continuously differentiable for all $\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{z}\|_1 < \delta$ for some positive δ. From Gelfand's formula (Theorem 13) consider a positive integer k such that $\left\|J^k[\boldsymbol{z}]\right\|_1 < \rho^k$. There exist $\varepsilon \in (0,1], \varepsilon$ depending upon f and k for which the following is true. Let $\left(\boldsymbol{x}^{(i)}\right)_{i=0}^{k}$ be sequences of vectors with $\boldsymbol{x}^{(i)} \in \Delta_m$ which satisfy the following conditions:*

1. *For $1 \leq i \leq k$, it holds that*
$$\boldsymbol{x}^{(i)} = f(\boldsymbol{x}^{(i-1)}).$$
2. *For $0 \leq i \leq k$, $\left\|\boldsymbol{x}^{(i)} - \boldsymbol{z}\right\|_1 \leq \varepsilon$.*

Then, we have

$$\left\|\boldsymbol{x}^{(k)} - \boldsymbol{z}\right\|_1 \leq \rho^k \left\|\boldsymbol{x}^{(0)} - \boldsymbol{z}\right\|_1.$$

Proof. We denote the set $\{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{z}\|_1 < \delta\}$ by $B(\boldsymbol{z}, \delta)$. Since f is continuously differentiable on $B(\boldsymbol{z}, \delta)$, $\nabla f_i(\boldsymbol{x})$ is continuous on $B(\boldsymbol{z}, \delta)$ for $i = 1, ..., m$. Let $A(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ be a matrix so that $A_{ij}(\boldsymbol{y}_1, ..., \boldsymbol{y}_m) = (\nabla f_i(\boldsymbol{y}_i))_j$.[10] This implies that the function on $\times_{i=1}^{mk} B(\boldsymbol{z}, \delta)$ defined by $\boldsymbol{w}_{11}, \boldsymbol{w}_{12}, \ldots, \boldsymbol{w}_{1m}, \boldsymbol{w}_{21}, \ldots \boldsymbol{w}_{mk} \mapsto$

[10] Easy to see that $A(\boldsymbol{z}, \ldots, \boldsymbol{z}) = J[\boldsymbol{z}]$.

$\prod_{i=1}^{k} A(\boldsymbol{w}_{i1},\ldots,\boldsymbol{w}_{im})$ is also continuous. Hence, there exist $\varepsilon_1,\varepsilon_2>0$ smaller than 1 such that if $\left\|\boldsymbol{w}_{ij}-\boldsymbol{z}\right\| \le \varepsilon_1$ for $1\le i\le k, 1\le j\le m$ then

$$\left\|\prod_{i=1}^{k} A(\boldsymbol{w}_{i1},\ldots,\boldsymbol{w}_{im})\right\|_1 \quad \le \quad \left\|J^k[\boldsymbol{z}]\right\|_1 - \varepsilon_2 < \rho^k. \tag{2}$$

From Taylor's theorem (Theorem 14) we have that $\boldsymbol{x}^{(t+1)} = A(\xi_1^{(k-t)},\ldots,\xi_m^{(k-t)})(\boldsymbol{x}^{(t)}-\boldsymbol{z})$ where $\boldsymbol{\xi}_i^{(k-t)}$ lies in the line segment from $\boldsymbol{z}$ to $\boldsymbol{x}^{(t)}$ for $i=1,\ldots,m$. By induction we get that

$$\boldsymbol{x}^{(k)}-\boldsymbol{z} = \prod_{j=1}^{k} A(\xi_1^{(j)},\ldots,\xi_m^{(j)})(\boldsymbol{x}^{(0)}-\boldsymbol{z}).$$

We choose $\varepsilon=\min(\varepsilon_1,\delta)$. Therefore since $\boldsymbol{\xi}_i^{(j)} \in B(\boldsymbol{z},\varepsilon)$ for $i=1,\ldots,m$ and $j=1,\ldots,k$, from inequality 2 we get that $\left\|\boldsymbol{x}^{(k)}-\boldsymbol{z}\right\|_1 < \rho^k \left\|\boldsymbol{x}^{(0)}-\boldsymbol{z}\right\|_1$. □

Lemma 17 below roughly says that the stochastic evolution guided by f does not deviate by much from the deterministic dynamics with update rule f after t steps, for t some small positive integer.

Lemma 17. *Let $f:\Delta_m\to\Delta_m$ be continuously differentiable in the interior of Δ_m. Let $\boldsymbol{X}^{(0)}$ be the state of a stochastic evolution guided by f at time 0. Then with probability $1-2t\cdot m\cdot e^{-2\varepsilon^2 N}$ we have that $\left\|\boldsymbol{X}^{(t)}-f^t(\boldsymbol{X}^{(0)})\right\|_1 \le t\beta^t\varepsilon m$, where $\beta \stackrel{\text{def}}{=} \sup_{\boldsymbol{x}\in\Delta_m}\|J[\boldsymbol{x}]\|_1$.*

Proof. We proceed by induction. For $t=1$ the result follows from concentration (Chernoff bounds, Theorem 15). Using the triangle inequality we get that

$$\left\|\boldsymbol{X}^{(t+1)}-f^{t+1}(\boldsymbol{X}^{(0)})\right\|_1 \le \left\|\boldsymbol{X}^{(t+1)}-f(\boldsymbol{X}^{(t)})\right\|_1 + \left\|f(\boldsymbol{X}^{(t)})-f^{t+1}(\boldsymbol{X}^{(0)})\right\|_1.$$

With probability at least $1-2m\cdot e^{-2\varepsilon^2 N}$ (Chernoff bounds, Theorem 15) we have that

$$\left\|\boldsymbol{X}^{(t+1)}-f(\boldsymbol{X}^{(t)})\right\|_1 \le \varepsilon m, \tag{3}$$

and also by the fact that $\|f(\boldsymbol{x})-f(\boldsymbol{x}')\|_1 \le \beta\|\boldsymbol{x}-\boldsymbol{x}'\|_1$ and induction we get that with probability at least $1-2t\cdot m\cdot e^{-2\varepsilon^2 N}$

$$\left\|f(\boldsymbol{X}^{(t)})-f^{t+1}(\boldsymbol{X}^{(0)})\right\|_1 \le \beta\left\|\boldsymbol{X}^{(t)}-f^t(\boldsymbol{X}^{(0)})\right\|_1 \le \beta\cdot t\beta^t\varepsilon m. \tag{4}$$

It is easy to see that $\varepsilon m + t\beta^{t+1}\varepsilon m \le (t+1)\beta^{t+1}\varepsilon m$, hence from inequalities 3 and 4 the result follows with probability at least $1-2(t+1)\cdot m\cdot e^{-2\varepsilon^2 N}$. □

Existence of Inverse function

For the rest of this section when we talk about the inverse of the Jacobian of a function f at an α-unstable fixed point, we mean the pseudoinverse which also has left eigenvector all ones $\mathbf{1}^\top$ with eigenvalue 0 (see also Remark in Section 2). Since we use a lot the inverse of a function f around a neighborhood of α-unstable fixed points in our lemmas, we need to prove that the inverse is well defined.

Lemma 18. *Let $f:\Delta_m\to\Delta_m$ be continuously differentiable in the interior of Δ_m. Let $\boldsymbol{z}$ be an α-unstable fixed point ($\alpha>1$). Then $f^{-1}(\boldsymbol{x})$ is well-defined in a neighborhood of $\boldsymbol{z}$ and is also continuously differentiable in that neighborhood. Also $J_{f^{-1}}[\boldsymbol{z}] = J^{-1}[\boldsymbol{z}]$ where $J_{f^{-1}}[\boldsymbol{z}]$ is the Jacobian of f^{-1} at $\boldsymbol{z}$.*

Proof. This comes from the Inverse function theorem. It suffices to show that $J[\boldsymbol{z}]\boldsymbol{x} = 0$ iff $\sum_i x_i = 0$, namely the differential is invertible on the simplex Δ_m. This is true by assumption since the minimum eigenvalue $\lambda_{\min}$ of $(J[\boldsymbol{z}])$, excluding the one with left eigenvector $\mathbf{1}^\top$, will satisfy $\lambda_{\min} > \alpha > 1 > 0$. Finally the Jacobian of f^{-1} at $\boldsymbol{z}$ is just the pseudoinverse $J^{-1}[\boldsymbol{z}]$ (which will have as well $\mathbf{1}^\top$ as a left eigenvector with eigenvalue 0). □

Distance $\Omega\left(\frac{\log^{2/3} N}{\sqrt{N}}\right)$

Lemma 19. *Let $f : \Delta_m \to \Delta_m$ be continuously differentiable in the interior of Δ_m. Let $\boldsymbol{X}^{(0)}$ be the state of a stochastic evolution guided by f at time 0 and also $\boldsymbol{z}$ be an α-unstable fixed point of f such that $\left\|\boldsymbol{X}^{(0)} - \boldsymbol{z}\right\|_1$ is $O\left(\frac{\log^{2/3} N}{\sqrt{N}}\right)$. Then with probability at least $\Theta(1)$ we get that*

$$\left\|\boldsymbol{X}^{(t)} - \boldsymbol{z}\right\|_1 \geq \frac{\log^{2/3} N}{\sqrt{N}}$$

after at most $O(\log N)$ steps.

Proof. We assume that $\boldsymbol{X}^{(t)}$ is in a neighborhood of $\boldsymbol{z}$ which is $o_N(1)$ for the rest of the proof, otherwise the lemma holds trivially. Let q be a positive integer such that $\left\|(J^{-1}[\boldsymbol{z}])^q\right\|_1 < \frac{1}{\alpha^q} < \frac{2}{5}$ (using Gelfand's formula 13 and the fact that $\alpha > 1$). First of all, it is easy to see that if $\left\|X^{(0)} - \boldsymbol{z}\right\|_1$ is $o\left(\frac{1}{\sqrt{N}}\right)$ then with probability at least $\Theta(1) = c_1$ we have after one step that $\left\|X^{(1)} - \boldsymbol{z}\right\|_1 > \frac{c}{\sqrt{N}}$ (this is true because the variance of binomial is $\Theta(N)$ and by CLT). We choose $c = \sqrt{2}\log(4mq)q\beta^q m$ where $\beta \stackrel{\text{def}}{=} \sup_{\boldsymbol{x}\in\Delta_m} \|J[\boldsymbol{x}]\|_1$. From Lemma 17 we get that with probability at least $\frac{1}{2}$ the deviation between the deterministic dynamics and the stochastic evolution after q steps is at most $\frac{\log(4mq)q\beta^q m}{\sqrt{2N}}$ (by substitute $\varepsilon = \frac{\log(4mq)}{\sqrt{2N}}$ in Lemma 17). Hence, using Lemma 16 for the function $h = f^{-1}$ around $\boldsymbol{z}$ and $k = q$, $\mathrm{sp}\,(J-1[\boldsymbol{z}]) < \frac{1}{\alpha}$, after q steps we get that $\left\|f^q(\boldsymbol{X}^{(1)}) - \boldsymbol{z}\right\|_1 \geq \alpha^q \left\|\boldsymbol{X}^{(1)} - \boldsymbol{z}\right\|_1$ with probability at least $\frac{1}{2}c_1$. From Lemma 17 and using the facts that $\alpha^q > 5/2$ and $\left\|\boldsymbol{X}^{(1)} - \boldsymbol{z}\right\|_1 \geq 2\frac{\log(4mq)q\beta^q m}{\sqrt{2N}}$ we conclude that

$$\begin{aligned}\left\|\boldsymbol{X}^{(q+1)} - \boldsymbol{z}\right\|_1 &\geq \left\|f^q(\boldsymbol{X}^{(1)}) - \boldsymbol{z}\right\|_1 - \frac{\log(4mq)q\beta^q m}{\sqrt{2N}} \\ &\geq \alpha^q \left\|\boldsymbol{X}^{(1)} - \boldsymbol{z}\right\|_1 - \frac{\log(4mq)q\beta^q m}{\sqrt{2N}} \geq 2\left\|\boldsymbol{X}^{(1)} - \boldsymbol{z}\right\|_1.\end{aligned}$$

By induction, we conclude that $\left\|\boldsymbol{X}^{(qt+1)} - \boldsymbol{z}\right\|_1 \geq \frac{\log^{2/3} N}{\sqrt{N}}$ with t to be at most $2/3(\log\log N)$ with probability at least $\frac{c_1}{(\log N)^{2/3}}$. Since we have made no assumptions on the position of the chain (except the distance), it follows that after at most $c_2(\log N)^{2/3} \cdot (\log\log N) = O(\log N)$ steps, the Markov chain has reached distance greater than $\frac{\log^{2/3} N}{\sqrt{N}}$ from the fixed point with probability $\Theta(1)$. □

Distance $\Theta(1)$

Combining Lemma 19 with the lemma below we can show that after $O(\log N)$ number of steps, the Markov chain will have distance from an α-unstable fixed point lower bounded by a constant $\Theta(1)$ with sufficient probability.

Lemma 20. *Let $f:\Delta_m\to\Delta_m$ be continuously differentiable in the interior of Δ_m. Let $\boldsymbol{X}^{(0)}$ be the state of a stochastic evolution guided by f at time 0 and also $\boldsymbol{z}$ be an α-unstable fixed point of f such that $\left\|\boldsymbol{X}^{(0)}-\boldsymbol{z}\right\|_1\geq\frac{\log^{2/3}N}{\sqrt{N}}$. Then with probability $1-\frac{1}{poly(N)}$ we have that $\left\|\boldsymbol{X}^{(t)}-\boldsymbol{z}\right\|_1$ is $r\overset{\text{def}}{=}\Theta(1)$ after at most $O(\log N)$ steps.*

Proof. Let r be such that we can apply Lemma 16 for f^{-1} with fixed point $\boldsymbol{z}$ and parameters $\rho=\frac{1}{a}$ and q such that $a^q<\frac{1}{2}$ since $\text{sp}\left(J^{-1}[\boldsymbol{z}]\right)<\frac{1}{a}$ and q is given from Gelfand's formula. Using Lemma 17 for $\varepsilon=\sqrt{\frac{\gamma\log N}{N}}$ we get that $\boldsymbol{X}^{(1)},\ldots,\boldsymbol{X}^{(q)}$ have ℓ_1 distance $\Omega\left(\frac{\log^{2/3}N}{\sqrt{N}}\right)$ from $\boldsymbol{z}$, with probability at least $1-2\frac{mq}{N^{2\gamma}}$. Then by induction for some t follows that

$$\begin{aligned}\left\|\boldsymbol{X}^{(t)}-\boldsymbol{z}\right\|_1&\geq\left\|f^q(\boldsymbol{X}^{(t-q)})-\boldsymbol{z}\right\|_1-q\beta^q m\sqrt{\frac{\gamma\log N}{N}}\\&=(1-o_N(1))\left\|f^q(\boldsymbol{X}^{(t-q)})-\boldsymbol{z}\right\|_1\\&\geq(1-o_N(1))\alpha^q\left\|\boldsymbol{X}^{(t-q)}-\boldsymbol{z}\right\|_1>2\left\|\boldsymbol{X}^{(t-q)}\right\|_1.\text{(Lemma 17)}\end{aligned}$$

Therefore, after at most $T=q\log N$ steps we get that $\left\|\boldsymbol{X}^{(T)}-\boldsymbol{z}\right\|_1\geq r$ with probability at least $1-\frac{2mq^2\log N}{N^{2\gamma}}$ from union bound (and choose $\gamma=2$). □

Below we show the last technical lemma of the section. Intuitively says that given a dynamical system where the update rule is defined in the simplex, if for every initial condition, the dynamics converges to some fixed point $\boldsymbol{z}$, then $\boldsymbol{z}$ cannot be an α-unstable unless the initial condition is $\boldsymbol{z}$.

Lemma 21. *Let $f:\Delta_m\to\Delta_m$ be continuously differentiable and assume that f has $\boldsymbol{z}_0,\ldots,\boldsymbol{z}_{l+1}$ (l is finite) fixed points, where $\boldsymbol{z}_0$ is stable such that $\text{sp}\left(J[\boldsymbol{z}_0]\right)<\rho<1$ and $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{l+1}$ are α-unstable with $\alpha>1$. Assume also that $\lim_{q\to\infty}f^q(\boldsymbol{x})$ exists for all $\boldsymbol{x}\in\Delta_m$ (and it is some fixed point). Let $B=\cup_{i=1}^{l}B(\boldsymbol{z}_i,r_i)$, where $B(\boldsymbol{z}_i,r_i)$ denotes the open ball of radius r_i around $\boldsymbol{z}_i$ and set $\Delta=\Delta_m-B$. Then for every ε, there exists a t such that*

$$\left\|f^t(\boldsymbol{x})-\boldsymbol{z}_0\right\|_1<\varepsilon$$

for all $\boldsymbol{x}\in\Delta$.

Proof. If Δ is empty, then it holds trivially. By assumption we have that for all $\boldsymbol{x}\in\Delta$, $\lim_{q\to\infty}f^q(\boldsymbol{x})=\boldsymbol{z}_i$ for some $i=0,\ldots,l+1$. Let $\boldsymbol{z}$ be an α-unstable fixed point. We claim that the if $\lim_{t\to\infty}f^t(\boldsymbol{x})=\boldsymbol{z}$ then $\boldsymbol{x}=\boldsymbol{z}$. Let us prove this claim. Assume $\boldsymbol{x}_0\in\Delta$ and that $\boldsymbol{x}_0$ is not a fixed point. By assumption $\lim_{q\to\infty}f^q(\boldsymbol{x}_0)=\boldsymbol{z}_i$ for some $i>0$, hence for every $\delta>0$, there exists a q_0 such that for $q\geq q_0$ we get that $\|f^q(\boldsymbol{x}_0)-\boldsymbol{z}_i\|_1\leq\delta$. We choose some k such that $\text{sp}\left((J^{-1}[\boldsymbol{z}_i])^k\right)<\frac{1}{\alpha^k}$ and we consider an ε such that Theorem 16 holds for function f^{-1} and k. We pick $\delta=\frac{\min(\varepsilon,r_i)}{2}$ and assume a q_0 such that by convergence assumption $\|f^q(\boldsymbol{x}_0)-\boldsymbol{z}_i\|_1\leq\delta$ for $q\geq q_0$. Hence Theorem 16 holds for the trajectory $(f^{t+q_0}(\boldsymbol{x}_0))_{t\in\mathbb{N}}$. Set $s=\|f^{q_0}(\boldsymbol{x}_0)-\boldsymbol{z}_i\|_1$ and observe that for $t=q_0+k\lceil\frac{\log_\alpha\frac{2\delta}{s}}{k}\rceil$ it holds that $\|f^t(\boldsymbol{x}_0)-\boldsymbol{z}_i\|_1\geq a^{t-q_0}\|f^{q_0}(\boldsymbol{x}_0)-\boldsymbol{z}\|_1\geq2\delta$ (due to Lemma 16), i.e., we reached a contradiction. Hence $\lim_{t\to\infty}f^t(\boldsymbol{x})=\boldsymbol{z}_0$ for all $\boldsymbol{x}\in\Delta$. The rest follows from Lemma 22 which is stated below. □

Lemma 22. *Let $S\subset\Delta_m$ be compact and assume that $\lim_{t\to\infty}f^t(\boldsymbol{x})=\boldsymbol{z}$ for all $\boldsymbol{x}\in S$. Then for every ε, there exists a q such that*

$$\|f^q(\boldsymbol{x})-\boldsymbol{z}\|_1<\varepsilon$$

for all $\boldsymbol{x}\in S$.

Proof. Because of the convergence assumption, for every $\varepsilon > 0$ and every $x \in S$, there exists an $d = d_x$ (depends on $\boldsymbol{x}$) such that

$$\left\| f^d(\boldsymbol{x}) - \boldsymbol{z} \right\|_1 < \varepsilon.$$

Define the sets $A_i = \left\{ \boldsymbol{y} \in S \mid \left\| f^i(\boldsymbol{y}) - \boldsymbol{z} \right\|_1 < \varepsilon \right\}$ for each positive integer i. Then, since f^i is continuous, the sets A_i are open in S, and therefore, by the above condition, form an open cover of S (since every $\boldsymbol{y}$ must lie in some A_i). By compactness, some finite collection of them must therefore cover S, and hence by taking q to be the maximum of the indices of the sets in this finite collection the lemma follows.

□

We are now able to prove the main theorem of the section, i.e., Theorem 6.

Proof of Theorem 6. Consider $r_1, \ldots, r_l$ as can occur from Lemma 19 and assume without loss of generality that the open balls $B(\boldsymbol{z}_i, r_i)$ for $i = 1, \ldots, l$, with center $\boldsymbol{z}_i$ and radius r_i (in ℓ_1 distance) are disjoint sets and that $\Delta \stackrel{\text{def}}{=} \Delta_m \setminus \cup_{i=1}^{l} B(\boldsymbol{z}_i, r_i)$ is not empty (otherwise we could decrease r_i's since they remain constants and Lemma 19 would still hold). We consider two chains $\boldsymbol{X}^{(0)}, \boldsymbol{Y}^{(0)}$. We claim that with probability $\Theta(1)$ (which can be boosted to any constant) each chain reaches within $\frac{1}{N^w}$ distance of the stable fixed point $\boldsymbol{z}_0$ for some $w > 0$, after at most $T = O(\log N)$ steps. Then the coupling constructed in [19] works because it uses the smoothness of f and the stability of the fixed point, as long as the two chains are within $\frac{1}{N^w}$ for some $w > 0$ distance of $\boldsymbol{z}_0$. Due to the coupling, as the two chains reach within $\frac{1}{N^w}$ distance of $\boldsymbol{z}_0$, they collide after $O(\log N)$ steps (with probability $\Theta(1)$ which also can be boosted to any constant) and hence the mixing time will be $O(\log N)$. To prove the claim, we first use Lemmas 20 and 19. It occurs that with probability say $\Theta(1)$ after at most $O(\log^{2/3} N \log\log N) + O(\log N)$ steps, each chain will have reached the compact set Δ. Moreover, from Lemma 34 (A.1 in [19]) we have that for all $\boldsymbol{x} \in \Delta$, $f^t(\boldsymbol{x})$ converges to fixed point $\boldsymbol{z}_0$ exponentially fast. Hence, using Lemma 35 (Claim 5.16 in [19]) follows that after $O(\log N)$ steps, each chain that started in Δ comes within $\frac{1}{N^w}$ distance of $\boldsymbol{z}_0$ with sufficiently enough probability. □

6 Multiple Stable fixed points

Staying close to fixed point

We prove the main lemma of this section, then our second result will be a corollary. The main lemma states that as long as the Markov chain starts from a neighborhood of one stable fixed point, it takes at least exponential time to get away from that neighborhood with probability say $\frac{9}{10}$.

Lemma 23. *Let $f : \Delta_m \to \Delta_m$ be continuously differentiable in the interior of Δ_m with stable fixed points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_l$ and k (independent of N) be such that $\left\| J[\boldsymbol{z}_i]^k \right\|_1 < \rho_i^k < 1$ for all $i = 1, \ldots, l$. Let $\boldsymbol{X}^{(0)}$ be the state of a stochastic evolution guided by f at time 0. There exists a small constant ε_i (independent of N) such that given that $\boldsymbol{X}^{(0)}$ satisfies $\left\| \boldsymbol{X}^{(0)} - \boldsymbol{z}_i \right\|_1 \leq m\varepsilon_i$ for some stable fixed point $\boldsymbol{z}_i$, after $t = \frac{e^{2\varepsilon_i{}^2 N}}{20mk}$ steps it holds that $\left\| \boldsymbol{X}^{(t)} - \boldsymbol{z}_i \right\|_1 \leq \frac{(k+1)\beta^k \varepsilon_i m}{1-\rho_i}$ with probability at least $\frac{9}{10}$.*

Proof. ε_i will be chosen later. By Lemma 17 it follows that

$$\left\| \boldsymbol{X}^{(t)} - \boldsymbol{z} \right\|_1 \leq \left\| f^t(\boldsymbol{X}^{(0)}) - \boldsymbol{z} \right\|_1 + t\beta^t \varepsilon_i m$$

with probability at least $1-2m\cdot ke^{-2\varepsilon_i^2 N}$ for $t=1,\ldots,k$. Since $\left\|f^t(\boldsymbol{X}^{(0)})-\boldsymbol{z}\right\|_1 \le \beta^t\left\|\boldsymbol{X}^{(0)}-\boldsymbol{z}\right\|_1$, it follows that $\left\|\boldsymbol{X}^{(t)}-\boldsymbol{z}\right\|_1 \le (t+1)\beta^t\varepsilon_i m$ with probability at least $1-2m\cdot ke^{-2\varepsilon_i^2 N}$ for $t=1,\ldots,k-1$. Assume that $\left\|\boldsymbol{X}^{(t)}-\boldsymbol{z}\right\|_1 \le (t+1)\beta^t\varepsilon_i m$ is true for $t=1,\ldots,k-1$. We choose ε_i small enough constant such that Lemma 16 holds with $\varepsilon=\frac{(k+1)\beta^k\varepsilon_i m}{1-\rho_i}$. To prove the lemma, we use induction on t and show that $\left\|\boldsymbol{X}^{(t)}-\boldsymbol{z}_i\right\|_1 \le \left((k+1)\beta^k\varepsilon_i m\right)\cdot\left(\sum_{j=0}^t\rho_i^j\right) < \frac{\left((k+1)\beta^k\varepsilon_i m\right)}{1-\rho_i} < \varepsilon$ and hence Lemma 16 will hold. For $t=k$ we have that

$$\begin{aligned}
\left\|\boldsymbol{X}^{(k)}-\boldsymbol{z}_i\right\|_1 &\le \left\|f^k(\boldsymbol{X}^{(0)})-\boldsymbol{z}_i\right\|_1 + \left\|f^k(\boldsymbol{X}^{(0)})-\boldsymbol{X}^{(k)}\right\|_1 \text{ (triangle inequality)}\\
&\le \rho_i^k\left\|\boldsymbol{X}^{(0)}-\boldsymbol{z}_i\right\|_1 + k\beta^k\varepsilon_i m \text{ (Lemma 16 and Lemma 17)}\\
&< (1+\rho_i^k)(k+1)\beta^k\varepsilon_i m < (\sum_{j=0}^k\rho_i^j)(k+1)\beta^k\varepsilon_i m.
\end{aligned}$$

Let $t'=t-k$, be a time index. We do the same trick as for the base case and we get that

$$\begin{aligned}
\left\|\boldsymbol{X}^{(t)}-\boldsymbol{z}_i\right\|_1 &\le \left\|f^k(\boldsymbol{X}^{(t')})-\boldsymbol{z}_i\right\|_1 + \left\|f^k(\boldsymbol{X}^{(t')})-\boldsymbol{X}^{(t)}\right\|_1\\
&\le \rho_i^k\left\|\boldsymbol{X}^{(t')}-\boldsymbol{z}\right\|_1 + k\beta^k\varepsilon_i m\\
&\le \rho_i^k\left((k+1)\beta^k\varepsilon_i m\right)\cdot\left(\sum_{j=0}^{t'}\rho_i^j\right) + (k+1)\beta^k\varepsilon_i m \text{ (induction)}\\
&= \left((k+1)\beta^k\varepsilon_i m\right)\cdot\left(1+\sum_{j=k}^{t}\rho_i^j\right) < \left((k+1)\beta^k\varepsilon_i m\right)\cdot\left(\sum_{j=0}^{t}\rho_i^j\right).
\end{aligned}$$

The error probability, i.e., at least one of the steps above fails and the chain gets larger noise than $k\beta^k\varepsilon_1 m$, by union bound will be at most $\frac{e^{2\varepsilon_i^2 N}}{20mk}\cdot 2mk\cdot e^{-2\varepsilon_i^2 N}=\frac{1}{10}$ (by Lemma 17). □

We can now prove the main theorem 7 which follows as a corollary from Lemma 23.

Proof of Thereom 7. Two stable fixed points suffice; let $\boldsymbol{z}_1,\boldsymbol{z}_2$. Consider the ε_i's from the previous lemma (Lemma 23) and set $S_i=\{\boldsymbol{x}:\|\boldsymbol{x}-\boldsymbol{z}_i\|_1\le\frac{(k+1)\beta^k\varepsilon_i m}{1-\rho_i}\}$ for $i=1,2$ where $\beta\stackrel{\text{def}}{=}\sup_{\boldsymbol{x}\in\Delta_m}\|J[\boldsymbol{x}]\|_1$. We can choose $\varepsilon_1,\varepsilon_2$ so small such that $S_1\cap S_2=\varnothing$ (by continuity). Let μ be the stationary distribution. Set $S=S_1, T=\frac{e^{2\varepsilon_1^2 N}}{20mk}$ and $\boldsymbol{y}=\boldsymbol{z}_1$ if $\mu(S_1)\le\frac{1}{2}$, otherwise set $S=S_2, T=\frac{e^{2\varepsilon_2^2 N}}{20mk}$ and $\boldsymbol{y}=\boldsymbol{z}_2$. Assume $\left\|X^{(0)}-\boldsymbol{y}\right\|_1\le\varepsilon m$. Therefore from Lemma 23 we get that $\mathbb{P}\left[X^{(T)}\in\bar{S}\right]\le\frac{1}{10}$. and also by assumption $\mu(\bar{S})\ge\frac{1}{2}$. Let $\nu^{(T)}$ be the distribution of $X^{(T)}$. However

$$\left\|\mu,\nu^{(T)}\right\|_{\text{TV}}\ge\left|\mu(\bar{S})-\mathbb{P}\left[X^{(T)}\in\bar{S}\right]\right|>\frac{1}{4}$$

and the result follows, i.e., $t_{mix}(1/4)$ is $e^{\Omega(N)}$. □

7 Stable limit cycle

This part technically is small, because it depends on the previous section. We denote by $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_s$ ($s\ge 2$) the points in the stable limit cycle. Again we assume that $\boldsymbol{w}_i$'s are *well separated*.

Proof of Theorem 8. Let $h(\boldsymbol{x}) = f^s(\boldsymbol{x})$. It is clear to see that the Markov chain guided by h satisfies the assumptions of 7. The fixed points of h are just the points in the limit cycle, i.e., $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_s$. Additionally, it easy to see (via chain rule) that $J_{f^s}[\boldsymbol{w}_i] = J_{f^{s-1}}[f(\boldsymbol{w}_i)]J[\boldsymbol{w}_i] = J_{f^{s-1}}[\boldsymbol{w}_{i+1}]J[\boldsymbol{w}_i]$, where we denote by J_{f^i} the Jacobian of function $f^i(\boldsymbol{x})$ and $\boldsymbol{w}_{l+1} = \boldsymbol{w}_1$. Therefore

$$J_h[\boldsymbol{w}_i] = \prod_{j=1}^{i-1} J[\boldsymbol{w}_{i-j}] \prod_{j=i}^{s} J[\boldsymbol{w}_{s+i-j}].$$

Matrices don't commute in general but it is true that AB, BA have the same eigenvalues hence $\mathrm{sp}\,(J_h[\boldsymbol{w}_i]) < \rho$ is the same for all $i = 1, \ldots, s$. Finally, let k be such that $\left\| J_{f^s}[\boldsymbol{w}_i]^k \right\|_1 < \rho^k$ (using Gelfand's formula 13). For each $\boldsymbol{w}_i$ consider ε_i as in the proof of 23, for function h and upper bound ρ on the spectral radius of $J_{f^s}[w_i]$. Then analogously follows that for $t = \frac{e^{2\varepsilon_i^2 N}}{20mk \cdot s}$ with probability at least $\frac{9}{10}$ we have that $\left\| \boldsymbol{X}^{(t)} - \boldsymbol{w}_i \right\|_1 \leq \frac{(k+1)\beta^k \varepsilon_i m}{1-\rho}$ and the proof for $e^{\Omega(N)}$ mixing follows from Theorem 7. □

8 Phase transitions in Linguistic/Sexual Evolutionary Models

8.1 Sampling from distribution $g(\mathbf{x})$

In this section, we prove that the finite population linguistic model discussed in preliminaries can be seen as a stochastic evolution guided by the function g defined by $g(\boldsymbol{x}) = (1 - m\tau)\frac{x_i(B\boldsymbol{x})_i}{\boldsymbol{x}^\top B\boldsymbol{x}} + \tau$ (we assume that we have m grammars and $g : \Delta_m \to \Delta_m$, see Definition 3 to check what a stochastic evolution guided by a function is). Given a starting population of size N on m types represented by a $1/N$-integral probability vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ we consider the following process P1:

1. Reproduction, i.e., the number of individuals that use grammar G_i becomes $N^2 x_i (B\boldsymbol{x})_i$ and the total number is $N^2 \boldsymbol{x}^\top B\boldsymbol{x}$.

2. Each individual that uses grammar S can end up using grammar T with probability Q_{ST}.

We now show that sampling from P1 is exactly the same as sampling from the multinomial distribution $g(\boldsymbol{x})$. Taking one sample (individual) we compute the probability to use grammar t.

Claim 24. $\mathbb{P}\,[type\ t] = \frac{N^2 \sum_j Q_{jt} x_j (B\boldsymbol{x})_j}{N^2 \boldsymbol{x}^\top B\boldsymbol{x}} = (1 - m\tau)\frac{x_t (B\boldsymbol{x})_t}{\boldsymbol{x}^\top B\boldsymbol{x}} + \tau.$

Proof. We have

$$\begin{aligned}
\mathbb{P}\,[\text{type } t] &:= \sum_{i=1}^{m} Q_{it} \cdot \frac{x_i (B\boldsymbol{x})_i}{\boldsymbol{x}^\top B\boldsymbol{x}} \\
&= (1 - m\tau)\frac{x_t (B\boldsymbol{x})_t}{\boldsymbol{x}^\top B\boldsymbol{x}} + \tau \frac{x_t (B\boldsymbol{x})_t}{\boldsymbol{x}^\top B\boldsymbol{x}} + \tau \sum_{i \neq t} \frac{x_i (B\boldsymbol{x})_i}{\boldsymbol{x}^\top B\boldsymbol{x}} \\
&= (1 - m\tau)\frac{x_t (B\boldsymbol{x})_t}{\boldsymbol{x}^\top B\boldsymbol{x}} + \tau \frac{\boldsymbol{x}^\top B\boldsymbol{x}}{\boldsymbol{x}^\top B\boldsymbol{x}}.
\end{aligned}$$

□

From 24, we see that producing N independent samples from the process P1 described above (which is the finite linguistic model discussed in the introduction) produces the same distribution as producing N independent samples from the distribution $g(\boldsymbol{x})$. So, we assume that the finite linguistic model is a stochastic evolution guided by g (see Definition 3).

8.2 Analyzing the Infinite Population Dynamics

In this section we prove several structural properties of the linguistic dynamics. We start this section by proving that the linguistic dynamics converges to fixed points. [11]

Theorem 25 (Convergence of Linguistic Dynamics). *The linguistic dynamics converges to fixed points. In particular, the Lyapunov function* $P(\boldsymbol{x}) = (x^\top Bx)^{\frac{1}{\tau}-m}\prod_i x_i^2$ *is strictly increasing along the trajectories for* $0 \le \tau \le 1/m$.

Proof. We first prove the results for rational τ; let $\tau = \kappa/\lambda$. We use the theorem of Baum and Eagon [1]. Let

$$L(\boldsymbol{x}) = (x^\top Bx)^{\lambda - m\kappa}\prod_i x_i^{2\kappa}.$$

Then

$$x_i\frac{\partial L}{\partial x_i} = 2\kappa L + \frac{2x_i(B\boldsymbol{x})_i(\lambda - m\kappa)L}{\boldsymbol{x}^\top B\boldsymbol{x}}.$$

It follows that

$$\begin{aligned}\frac{x_i\frac{\partial L}{\partial x_i}}{\sum_i x_i\frac{\partial L}{\partial x_i}} &= \frac{2\kappa L + \frac{2x_i(B\boldsymbol{x})_i(\lambda-m\kappa)L}{\boldsymbol{x}^\top B\boldsymbol{x}}}{2m\kappa L + 2(\lambda - m\kappa)L}\\ &= \frac{2\kappa L}{2\lambda L} + \frac{2L(\lambda - m\kappa)x_i(B\boldsymbol{x})_i}{2\lambda L\boldsymbol{x}^\top B\boldsymbol{x}}\\ &= (1-m\tau)x_i\frac{(B\boldsymbol{x})_i}{\boldsymbol{x}^\top B\boldsymbol{x}} + \tau\end{aligned}$$

where the first equality comes from the fact that $\sum_{i=1}^m x_i(B\boldsymbol{x})_i = \boldsymbol{x}^\top B\boldsymbol{x}$. Since L is a homogeneous polynomial of degree 2λ, from Theorem 10 we get that L is strictly increasing along the trajectories, namely

$$L(g(\boldsymbol{x})) > L(\boldsymbol{x})$$

unless $\boldsymbol{x}$ is a fixed point. So $P(\boldsymbol{x}) = L^{1/\kappa}(\boldsymbol{x})$ is a potential function for the dynamics.

To prove the result for irrational τ, we just have to see that the proof of [1] holds for all homogeneous polynomials with degree d, even irrational.

To finish the proof let $\Omega \subset \Delta_m$ be the set of limit points of an orbit $\boldsymbol{x}(t)$ (frequencies at time t for $t \in \mathbb{N}$). $P(\boldsymbol{x}(t))$ is increasing with respect to time t by above and so, because P is bounded on Δ_m, $P(\boldsymbol{x}(t))$ converges as $t \to \infty$ to $P^* = \sup_t\{P(\boldsymbol{x}(t))\}$. By continuity of P we get that $P(\boldsymbol{y}) = \lim_{t\to\infty} P(\boldsymbol{x}(t)) = P^*$ for all $\boldsymbol{y} \in \Omega$. So P is constant on Ω. Also $\boldsymbol{y}(t) = \lim_{n\to\infty}\boldsymbol{x}(t_n + t)$ as $n \to \infty$ for some sequence of times $\{t_i\}$ and so $\boldsymbol{y}(t)$ lies in Ω, i.e. Ω is invariant. Thus, if $\boldsymbol{y} \equiv \boldsymbol{y}(0) \in \Omega$ the orbit $\boldsymbol{y}(t)$ lies in Ω and so $P(\boldsymbol{y}(t)) = P^*$ on the orbit. But P is strictly increasing except on equilibrium orbits and so Ω consists entirely of fixed points. □

8.3 Fixed points and bifurcation

Let $\boldsymbol{z}$ be a fixed point. $\boldsymbol{z}$ satisfies the following equations:

$$\frac{z_i - \tau}{z_i(B\boldsymbol{z})_i} = \frac{z_j - \tau}{z_j(B\boldsymbol{z})_j} = \frac{1-m\tau}{\boldsymbol{z}^T B\boldsymbol{z}} \text{ for all } i, j. \tag{5}$$

[11] This requires proof since convergence to limit cycles or the existence of strange attractors are a priori not ruled out.

The previous equations can be derived by solving $z_i = (1-m\tau)z_i \frac{z_i(B\boldsymbol{z})_i}{\boldsymbol{z}^\top B\boldsymbol{z}} + \tau$. By solving with respect to τ we get that

$$\tau = \frac{z_i z_j((B\boldsymbol{z})_i - (B\boldsymbol{z})_j)}{z_i(B\boldsymbol{z})_i - z_j(B\boldsymbol{z})_j} \text{ for } z_i(B\boldsymbol{z})_i \neq z_j(B\boldsymbol{z})_j.$$

Fact 26. *The uniform point $(1/m,\ldots,1/m)$ is a fixed point of the dynamics for all values of τ.*

To see why 26 is true, observe that $g_i(1/m,\ldots,1/m) = (1-m\tau)\frac{1}{m} + \tau = \frac{1}{m}$ for all i and hence $g(1/m,\ldots,1/m) = (1/m,\ldots,1/m)$. The fixed points satisfy the following property:

Lemma 27 (Two Distinct Values)**.** *Let $(x_1,\ldots,x_m)$ be a fixed point. Then $x_1,\ldots,x_m$ take at most two distinct values.*

Proof. Let $x_i \neq x_j$ for some i,j. Then it follows that

$$\tau = \frac{x_i x_j((B\boldsymbol{x})_i - (B\boldsymbol{x})_j)}{x_i(B\boldsymbol{x})_i - x_j(B\boldsymbol{x})_j} = \frac{x_i x_j(1-b)}{(1-b)(x_i+x_j)+b}.$$

Hence if $x_{j'} \neq x_i$ then

$$\frac{x_{j'}}{(1-b)(x_i+x_{j'})+b} = \frac{x_j}{(1-b)(x_i+x_j)+b}$$

from which follows that $x_j = x_{j'}$. Finally, the uniform fixed point satisfies trivially the property. □

We shall compute the threshold τ_c such that for $0 < \tau < \tau_c$ the dynamics has multiple fixed points and for $1/m \geq \tau > \tau_c$ we have only one fixed point (which by Fact 26 must be the uniform one). Let

$$h(x) = -x^2(m-2)(1-b) - 2x(1+b(m-2)) + 1 + b(m-2).$$

By Bolzano's theorem and the fact that $h(0) = 1+b(m-2) > 0$ and $h(-1) < 0$, $h(1) = 1-m < 0$, it follows that there exists one positive solution for $h(x) = 0$ which is between 0 and 1; we denote it by s_1.

We can now define

$$\tau_c \stackrel{\text{def}}{=} \frac{(1-b)s_1(1-s_1)}{(m-1)b + (1-b)(1+(m-2)s_1)}.$$

Lemma 28 (Bifurcation). *If $\tau_c < \tau \leq 1/m$ then the only fixed point is the uniform one. If $0 \leq \tau < \tau_c$ then there exist multiple fixed points.*

Proof. Assume that there are multiple fixed points (apart from the uniform, see 26) and let $(x_1,\ldots,x_m)$ be a fixed point, where x and y being the two values that the coordinates x_i take (by Lemma 27). Let $k \geq 1$ be the number of coordinates with value x and $m-k$ the coordinates with values y where $m > k$ and $kx + (m-k)y = 1$ (in case $k = 0$ or $m = k$ we get the uniform fixed point). Solving by τ we get that $\tau = \frac{xy(1-b)}{b+(1-b)(x+y)}$. We set $y = \frac{1-kx}{m-k}$ and we analyze the function

$$f(x,k) = \frac{(1-b)x(1-kx)}{(m-k)b + (1-b)(1+(m-2k)x)}$$

It follows that f is decreasing with respect to k (assuming $x < 1/k+1$ such that $y > 0$, see appendix B for Mathematica code for proving $f(x,k)$ is decreasing with respect to k). Hence the maximum is attained for $k = 1$. Hence, we can consider

$$f(x) \stackrel{\text{def}}{=} f(x,1) = \frac{(1-b)x(1-x)}{(m-1)b + (1-b)(1+(m-2)x)}.$$

By solving $\frac{df}{dx} = 0$ it follows that $h(x) = 0$ (where $h(x)$ is the numerator of the derivative of f). This occurs at s_1. For $\tau > \tau_c$ there exist no fixed points whose coordinates can take on more than one value by construction of f, namely the only fixed point is the uniform one. □

8.4 Stability analysis

The equations of the Jacobian are given below:

$$\frac{\partial g_i}{\partial x_i} = (1-m\tau)\left(\frac{(B\boldsymbol{x})_i + x_i B_{ii}}{\boldsymbol{x}^\top B\boldsymbol{x}} - \frac{x_i(B\boldsymbol{x})_i \cdot 2(B\boldsymbol{x})_i}{(\boldsymbol{x}^\top B\boldsymbol{x})^2}\right), \tag{6}$$

$$\frac{\partial g_j}{\partial x_i} = (1-m\tau)\left(\frac{x_j B_{ji}}{\boldsymbol{x}^\top B\boldsymbol{x}} - \frac{x_j(B\boldsymbol{x})_j \cdot 2(B\boldsymbol{x})_i}{(\boldsymbol{x}^\top B\boldsymbol{x})^2}\right) \text{ for } j \neq i. \tag{7}$$

Fact 29. *The all ones vector $(1,\ldots,1)$ is a left eigenvector of the Jacobian with corresponding eigenvalue 0.*

Proof. This can be derived by computing

$$\sum_{j=1}^{m} \frac{\partial g_j}{\partial x_i} = (1-m\tau)\left(\frac{2(B\boldsymbol{x})_i}{\boldsymbol{x}^\top B\boldsymbol{x}} - \frac{2\boldsymbol{x}^\top B\boldsymbol{x}(B\boldsymbol{x})_i}{(\boldsymbol{x}^\top B\boldsymbol{x})^2}\right) = 0.$$

□

We will focus on two specific classes of fixed points. The first one is the uniform, i.e., $(1/m,\ldots,1/m)$ which we denote by $\boldsymbol{z}_u$ and the other one is $(y,\ldots,y,\underbrace{x}_{i^{th}},y,\ldots,y)$ with $x+(m-1)y=1$ and $x > s_1$, which we denote by $\boldsymbol{z}_i$ (for $1 \leq i \leq m$).

Stability of $\boldsymbol{z}_u$. Let

$$\tau_u \stackrel{\text{def}}{=} \frac{1-b}{m(2-2b+mb)}.$$

Lemma 30. *If $\tau_u < \tau \leq 1/m$, then $\mathrm{sp}(J[z_u]) < 1$ and if $0 \leq \tau < \tau_u$, then $\mathrm{sp}(J[z_u]) > 1$.*

Proof. The Jacobian of the uniform fixed point has diagonal entries $(1-m\tau)\left(1-\frac{2}{m}+\frac{1}{1+(m-1)b}\right)$ and non-diagonal entries $(1-m\tau)\left(\frac{b}{1+(m-1)b}-\frac{2}{m}\right)$. Consider the matrix

$$W_u \stackrel{\text{def}}{=} J[\boldsymbol{z}_u] - (1-m\tau)\left(1+\frac{1-b}{1+(m-1)b}\right) I_m$$

where I_m is the identity matrix of size $m \times m$. The matrix W_u has eigenvalue 0 with multiplicity $m-1$ and eigenvalue $m(1-m\tau)\left(\frac{b}{1+(m-1)b}-\frac{2}{m}\right)$ with multiplicity 1. Hence the eigenvalues of $J[\boldsymbol{z}_u]$ are 0 with multiplicity 1 and $(1-m\tau)(1+\frac{1-b}{1+(m-1)b})$ with multiplicity $m-1$. Thus, the Jacobian of $\boldsymbol{z}_u$ has spectral radius less than one if and only if $-1 < (1-m\tau)(1+\frac{1-b}{1+(m-1)b}) < 1$. By solving with respect to τ it follows that

$$\frac{1-b}{m(2-2b+mb)} < \tau < \frac{3-3b+2bm}{m(2-2b+mb)}.$$

Because $1/m < \frac{3-3b+2bm}{m(2-2b+mb)}$ (as $b \leq 1$), the first part of the lemma follows. In case $0 \leq \tau < \frac{1-b}{m(2-2b+mb)}$ then $(1-m\tau)(1+\frac{1-b}{1+(m-1)b})$ and the second part follows. □

Hence, we conclude that τ_u is the threshold below which the uniform fixed point satisfies $\mathrm{sp}\,(J[\boldsymbol{z}_u]) > 1$ and above which $\mathrm{sp}\,(J[\boldsymbol{z}_u]) < 1$.

Stability of $\boldsymbol{z}_i$.

Lemma 31. *If* $0 \leq \tau < \tau_c$ *then* $\mathrm{sp}\,(J[z_i]) < 1$.

Proof. Consider the matrix

$$W_i \stackrel{\text{def}}{=} J[\boldsymbol{z}_i] - (1 - m\tau)\frac{y + b + (1-2b)y}{\boldsymbol{z}_i^\top B \boldsymbol{z}_i} I_m$$

where I_m is the identity matrix of size $m \times m$. The matrix W_i has eigenvectors of the form

$$(w_1, \ldots, w_{i-1}, 0, w_{i+1}, \ldots, w_m)$$

with $\sum_{j=1, j\neq i}^{m} w_j = 0$ (the dimension of the subspace is $m-2$) and corresponding eigenvalues 0. Hence the Jacobian has $m-2$ eigenvalues of value $(1-m\tau)\frac{y+b+(1-2b)y}{\boldsymbol{z}_i^\top B \boldsymbol{z}_i}$. It is true that $0 < (1-m\tau)\frac{y+b+(1-2b)y}{\boldsymbol{z}_i^\top B \boldsymbol{z}_i} < 1$ (see appendix B for Mathematica code). Finally, since $J[z_i]$ has an eigenvalue zero (see Fact 29), the last eigenvalue is

$$\begin{aligned}
&\mathrm{Tr}(J[z_i]) - (1-m\tau)(m-2)\frac{y+b+(1-2b)y}{\boldsymbol{z}_i^\top B \boldsymbol{z}_i} \\
=&(1-m\tau)\cdot\left(\frac{x+2b+(1-b)x+2y+(m-3)by}{\boldsymbol{z}_i^\top B \boldsymbol{z}_i} - \frac{2x(b+(1-b)x)^2 + 2(m-1)y(b+(1-b)y)^2}{(\boldsymbol{z}_i^\top B \boldsymbol{z}_i)^2}\right)
\end{aligned}$$

which is also less than 1 and greater than 0 (see appendix B for Mathematica code). □

Remark. In the case where $m = 2$ it follows that $\tau_u = \tau_c = \frac{1-b}{4}$. For $m > 2$ we have $\tau_u < \tau_c$ (see Mathematica code in Lemma B.3).

8.5 Mixing Time

In this section we prove our result concerning the linguistic model (finite population). The structural lemmas proved in the previous section are used here. Now, we proceed by analysing the mixing time of the Markov chain for the two intervals $(0, \tau_c)$ and $(\tau_c, 1/m]$.

Regime $0 < \tau < \tau_c$

Lemma 32. *For the interval* $0 < \tau < \tau_c$*. the mixing time of the Markov chain is* $\exp(\Omega(N))$.

Proof. By Lemma 31 it is true that there exist m fixed points $\boldsymbol{z}_i$ with $\mathrm{sp}\,(J[\boldsymbol{z}_i]) < 1$ and their pairwise distance is some positive constant independent of N (well-separated). Hence using Theorem 7 and because the Markov chain is a stochastic evolution guided by g (see 24), we conclude that the mixing time is $e^{\Omega(N)}$. □

Regime $\tau_c < \tau \leq 1/m$

We prove the second part of the of Theorem 9.

Lemma 33. *For the interval $\tau_c < \tau \leq 1/m$, the assumptions of the main theorem of [19] are satisfied, namely the mixing time of the Markov chain is $O(\log N)$.*

Proof. By Lemma 28, we know that in the interval $\tau_c < \tau \leq 1/m$ there is a unique fixed point (the uniform $\boldsymbol{z}_u$) and also by Lemma 30 that $\mathrm{sp}\,(J[\boldsymbol{z}_u]) < 1$. It is trivial to check that g is twice differentiable with bounded second derivative. It suffices to show the 4th condition in the Definition 4. Due to Theorem 25 we have $\lim_{k\to\infty} g^k(\boldsymbol{x}) \to \boldsymbol{z}_u$ for all $\boldsymbol{x} \in \Delta_m$. The rest follows from Lemma 22 (by setting $S = \Delta_m$). □

Our result on linguistic model is a consequence of 32, 33.

Remark. For $\tau = 1/m$ the Markov chain mixes in one step. This is trivial since g maps every point to the uniform fixed point $\boldsymbol{z}_u$.

References

[1] L. Baum and J. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73:360–363, 1967.

[2] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1999.

[3] Erick Chastain, Adi Livnat, Christos Papadimitriou, and Umesh Vazirani. Algorithms, games, and evolution. *Proceedings of the National Academy of Sciences*, 2014.

[4] Noam A. Chomsky. Rules and Representations. *Behavioral and Brain Sciences*, 3(127):1–61, 1980.

[5] Narendra Dixit, Piyush Srivastava, and Nisheeth K. Vishnoi. A finite population model of molecular evolution: Theory and computation. *Journal of Computational Biology*, 19(10):1176–1202, 2012.

[6] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[7] Richard Durrett. *Probability models for DNA sequence evolution*. Springer, 2008.

[8] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[9] Warren J. Ewens. *Mathematical Population Genetics I. Theoretical Introduction*. Springer, 2004.

[10] W.T. Fitch. *The Evolution of Language*. Approaches to the Evolution of Language. Cambridge University Press, 2010.

[11] Natalia L. Komarova and Martin A. Nowak. Language dynamics in finite populations. *Journal of Theoretical Biology*, 221(3):445 – 457, 2003.

[12] Erwin Kreyszig. *Introductory Functional Analysis with Applications*. Wiley, 1978.

[13] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

[14] Yun Long, Asaf Nachmias, and Yuval Peres. Mixing time power laws at criticality. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 205–214, Washington, DC, USA, 2007. IEEE Computer Society.

[15] J. Maynard-Smith and E. Szathmary. *The Major Transitions in Evolution*. New York: Oxford University Press, 1997.

[16] Ruta Mehta, Ioannis Panageas, and Georgios Piliouras. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics. In *Innovations in Theoretical Computer Science*, 2015.

[17] M.A. Nowak. *Evolutionary Dynamics*. Harvard University Press, 2006.

[18] Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Evolution of universal grammar. *Science*, 2001.

[19] Ioannis Panageas, Piyush Srivastava, and Nisheeth K. Vishnoi. Evolutionary dynamics in finite populations mix rapidly. *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, 2016.

[20] Christos H. Papadimitriou and Nisheeth K. Vishnoi. On the computational complexity of limit cycles in dynamical systems. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, page 403, 2016.

[21] Robin Pemantle. When are touchpoints limits for generalized pólya urns? *Proceedings of the American Mathematical Society*, pages 235–243, 1991.

[22] Georgios Piliouras, Carlos Nieto-Granda, Henrik I. Christensen, and Jeff S. Shamma. Persistent patterns: Multi-agent learning beyond equilibrium and utility. In *AAMAS*, pages 181–188, 2014.

[23] Kushal Tripathi, Rajesh Balagam, Nisheeth K. Vishnoi, and Narendra M. Dixit. Stochastic simulations suggest that HIV-1 survives close to its error threshold. *PLoS Comput Biol*, 8(9):e1002684, 09 2012.

[24] Nisheeth K. Vishnoi. The speed of evolution. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1590–1601, 2015.

[25] Nicholas C Wormald. Differential equations for random processes and random graphs. *The annals of applied probability*, pages 1217–1235, 1995.

A Lemmas from [19]

Lemma 34 (**Exponential convergence** [19] Modified)**.** *Choose z_0, ρ, Δ as in the proof of Theorem 6, and set $\beta \stackrel{\text{def}}{=} \sup_{\boldsymbol{x}\in\Delta_m} \|J[\boldsymbol{x}]\|_1$. Then there exist a positive r such that for every $\boldsymbol{x} \in \Delta$, and every positive integer t,*

$$\left\|f^t(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 \leq r\rho^t.$$

Proof. Let ε and k be as defined in 16. From Lemma 21, we know that there exists an ℓ such that for all $\boldsymbol{x} \in \Delta$,

$$\left\|f^{\ell}(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 \leq \frac{\varepsilon}{\beta^k}. \tag{8}$$

Note that this implies that $f^{\ell+i}(\boldsymbol{x})$ is within distance ε of $\boldsymbol{z}_0$ for $i = 0, 1, \ldots, k$, so that 16 can be applied to the sequence of vectors $f^{\ell}(\boldsymbol{x}), f^{\ell+1}(\boldsymbol{x}), \ldots, f^{\ell+k}(\boldsymbol{x})$ and $\boldsymbol{z}_0$. Thus, we get

$$\left\|f^{\ell+k}(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 \leq \rho^k \left\|f^{\ell}(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 \leq \frac{\rho^k \varepsilon}{\beta^k}.$$

Since $\rho < 1$, we can iterate this process. Using also the fact that the $1 \to 1$ norm of the Jacobian of f is at most β (which we can assume without loss of generality to be at least 1), we therefore get for every $\boldsymbol{x} \in \Delta$, and every $i \geq 0$ and $0 \leq j < k$

$$\begin{aligned}\left\|f^{\ell+ik+j}(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 &\leq \rho^{ki+j} \frac{\beta^j}{\rho^j} \left\|f^{\ell}(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 \\ &\leq \rho^{ki+j+\ell} \frac{\beta^{j+\ell}}{\rho^{j+\ell}} \left\|\boldsymbol{x} - \boldsymbol{z}_0\right\|_1 \leq \rho^{ki+j+\ell} \frac{\beta^{k+\ell}}{\rho^{k+\ell}} \left\|\boldsymbol{x} - \boldsymbol{z}_0\right\|_1\end{aligned}$$

where in the last line we use the facts that $\beta > 1$, $\rho < 1$ and $j < k$. Noting that any $t \geq \ell$ is of the form $\ell + ki + j$ for some i and j as above, we have shown that for every $t \geq \ell$ and every $\boldsymbol{x} \in \Delta$

$$\left\|f^{t}(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 \leq \left(\frac{\beta}{\rho}\right)^{k+\ell} \rho^t \left\|\boldsymbol{x} - \boldsymbol{z}_0\right\|_1. \tag{9}$$

Similarly, for $t < \ell$, we have, for any $\boldsymbol{z} \in \Delta$

$$\begin{aligned}\left\|f^{t}(\boldsymbol{x}) - \boldsymbol{z}_0\right\|_1 &\leq \beta^t \left\|\boldsymbol{x} - \boldsymbol{z}_0\right\|_1 \\ &\leq \left(\frac{\beta}{\rho}\right)^{t} \rho^t \left\|\boldsymbol{x} - \boldsymbol{z}_0\right\|_1 \leq \left(\frac{\beta}{\rho}\right)^{\ell} \rho^t \left\|\boldsymbol{x} - \boldsymbol{z}_0\right\|_1,\end{aligned} \tag{10}$$

where in the last line we have again used $\beta > 1, \rho < 1$ and $t < \ell$. From 10, 9, we get the claimed result with $r \stackrel{\text{def}}{=} \left(\frac{\beta}{\rho}\right)^{k+\ell}$. □

Lemma 35 ([19]). *Choose $\boldsymbol{z}_0, \rho, \Delta$ as in the proof of Theorem 6, set $\beta \stackrel{\text{def}}{=} \sup_{\boldsymbol{x} \in \Delta_m} \|J[\boldsymbol{x}]\|_1$ and consider r from Lemma 34. Define T_{start} to be the first time such that*

$$\left\|\boldsymbol{X}^{(T_{start}+i)} - \boldsymbol{z}_0\right\|_1, \left\|\boldsymbol{Y}^{(T_{start}+i)} - \boldsymbol{z}_0\right\|_1 \leq \frac{\alpha}{N^w} \text{ for } 0 \leq i \leq k-1,$$

where $\alpha \stackrel{\text{def}}{=} m + r$ and $w = \min\left(\frac{1}{6}, \frac{\log(1/\rho)}{6\log(\beta+1)}\right)$. It holds that

$$\mathbb{P}\left[T_{start} > t_{start} \log N\right] \leq 4mkt_o \log N \exp\left(-N^{1/3}\right), \tag{11}$$

where $t_{start} \stackrel{\text{def}}{=} \frac{1}{6\log(\beta+1)}$. The probability itself is upper bounded by $\exp\left(-N^{1/4}\right)$ for N large enough.

B Mathematica Code

B.1 Mathematica code for proving Lemma 28

```
Reduce[((1 - k*x)/(m - k))/(b + (1 - b)*(x + (1 - k*x)/(m - k))) < ((1 - (k + 1)* x)
       /(m - k - 1))/(b + (1 - b)*(x + (1 - (k + 1)*x)/(m - k - 1))) && 1 > b > 0 &&
       1 > x > 0 && 1/(k + 1) > x > 1/m && m >= 3 && m >= k + 2 && k >= 1]
```

```
False
```

B.2 Mathematica code for proving Lemma 31

First inequality in Lemma 31:

```
Reduce[1 > b > 0 && m >= 3 && -(m - 2) (1 - b) s^2 - 2 s (1 + b (m - 2)) + 1 +
    b (m - 2) == 0 && 0 < s < x < 1 && y == (1 - x)/(m - 1) &&
    t == (x*y*(1 - b))/(b + (1 - b)*(x + y)) && t <= 1/m &&
    (1 - m*t)*(y + b + (1 - 2*b)*y)/(b + (1 - b)*x^2 + (1 - b)*(m - 1)*y^2) >= 1]
```

```
False
```

Second inequality in Lemma 31:

```
Reduce[1 > b > 0 && m >= 3 && -(m - 2) (1 - b) s^2 - 2 s (1 + b (m - 2)) + 1 +
    b (m - 2) == 0 && 0 < s < x < 1 && y == (1 - x)/(m - 1) && 1/m >= t &&
  t == (x*y*(1 - b))/(b + (1 - b)*(x + y)) && ((1 - m*t)*((2*(x + y) +
          b*(2 - x + (m - 3)*y))/(b + (1 - b)*x^2 + (1 - b)*(m - 1)*y^2) -
          (2*x*(b + (1 - b)*x)^2 + 2*(m - 1)*y*(b + (1 - b)*y)^2)
          /((b + (1 - b)*x^2 + (1 - b)*(m - 1)*y^2)^2)) >= 1)]
```

```
False
```

B.3 Mathematica code for proving $\tau_c > \tau_u$ when $m > 2$

```
Reduce[1 > b > 0 && m >= 3 && -(m - 2) (1 - b) s^2 - 2 s (1 + b (m - 2)) + 1 +
    b (m - 2) == 0 && 0 < s < 1 && (s*(1 - s)*(1 - b))/((m - 1)*
       b + (1 - b)*(1 + (m - 2)*s)) <= (1 - b)/(m*(2 - 2*b + m*b))]
```

```
False
```