EVOLUTIONARY DYNAMICS IN FINITE POPULATIONS MIX RAPIDLY

IOANNIS PANAGEAS, PIYUSH SRIVASTAVA, AND NISHEETH K. VISHNOI

ABSTRACT. In this paper we prove that the rate of convergence, or mixing time, of a broad class of evolutionary dynamics in finite populations is roughly logarithmic in the size of the state space. An important special case of such a stochastic process is the Wright-Fisher model from evolutionary biology (with selection and mutation) on a population of size N over m genotypes. Our main result implies that the mixing time of this process is $O(\log N)$ for all mutation rates and fitness landscapes, and solves the main open problem from [DSV12]. In particular, it significantly extends the main result in [Vis15] who proved this for $m = 2$. Biologically, such models have been used to study the evolution of viral populations with applications to drug design strategies countering them. Here the time it takes for the population to reach a steady state is important both for the estimation of the steady-state structure of the population as well in the modeling of the treatment strength and duration. Our result, that such populations exhibit rapid mixing, makes both of these approaches sound.

Technically, we make a novel connection between Markov chains arising in evolutionary dynamics and dynamical systems on the probability simplex. This allows us to use the local and global stability properties of the fixed points of such dynamical systems to construct a contractive coupling in a fairly general setting. We expect that our mixing time result would be useful beyond the evolutionary biology setting, and the techniques used here would find applications in bounding the mixing times of Markov chains which have a natural underlying dynamical system.

CONTENTS

1. Introduction

Evolutionary dynamical systems are central to the sciences due to their versatility in modeling a wide variety of biological, social and cultural phenomena, see [Now06]. Such dynamics are often used to capture the deterministic, *infinite population* setting, and are typically the first step in our understanding of seemingly complex processes. However, real populations are finite and often lend themselves to substantial stochastic effects (such as random drift) and it is often important to understand these effects as the population size varies. Hence, stochastic or *finite population* versions of evolutionary dynamical systems are appealed to in order to study such phenomena. While there are many ways to translate a deterministic dynamical system into a stochastic one, one thing remains common: the mathematical analysis becomes much harder as differential equations are easier to analyze and understand than stochastic processes.

Consider the example of the error-prone evolution of an asexual haploid population in evolutionary biology; this is also the main motivation for this work. Each individual in the population could be one of m types. An individual of type i has a *fitness* (that translates to the ability to reproduce) which is specified by a positive integer a_i, and captured as a whole by a diagonal $m \times m$ matrix A whose (i,i)th entry is a_i. The reproduction is error-prone and this is captured by an $m \times m$ stochastic matrix Q whose (i,j)th entry captures the probability that the jth type will mutate to the ith type during reproduction.[1] If the population is assumed to be infinite and its evolution deterministic, then one can track the fraction of each type at step t of the evolution by a vector $\boldsymbol{x}^{(t)} \in \Delta_m$ (the probability simplex of dimension m) whose evolution is then governed by the difference equation $\boldsymbol{x}^{(t+1)} = \frac{QA\boldsymbol{x}^{(t)}}{\|QA\boldsymbol{x}^{(t)}\|_1}$. This was studied in the pioneering work of Eigen and co-authors [Eig71, ES77]. Of interest is the steady state[2] or the limiting distribution of this process and how it changes as one changes the evolutionary parameters Q and A. Importantly, this particular dynamical system has found use in modeling rapidly evolving viral populations (such as HIV), which in turn has guided drug and vaccine design strategies. As a result, these dynamics are well-studied; see [DSV12, Vis] for an in depth discussion.

However, viral populations are generally not infinite and show stochastic effects: e.g., the effective population size of HIV-1 in infected individuals is approximately $10^3 - 10^6$ [KAB06, BSSD11], which is believed to be responsible for the strongly stochastic nature of its evolution. Several researchers have studied stochastic versions of Eigen's deterministic evolution equations [NS89, CF98, AF98, vNCM99, Wil05, RPM08, SRA08, Mus11, TBVD12]. One such stochastic version, motivated by the Wright-Fisher model in population genetics, was studied by Dixit et al. [DSV12]. Here, the population is fixed to a size N and, after normalization, is a random point in Δ_m; say $\boldsymbol{X}^{(t)}$ at time t. How does one generate $\boldsymbol{X}^{(t+1)}$ in this model when the parameters are still described by the matrices Q and A as in the infinite population case? To do this, in the replication (R) stage, one first replaces an individual of type i in the current population by a_i individuals of type i: the total number of individuals of type i in the intermediate population is therefore $a_i N X_i^{(t)}$. In the selection (S) stage, the population is culled back to size N by sampling without replacement N individuals from this intermediate population. Finally, since the evolution is error prone, in the mutation (M) stage, one then mutates each individual in this intermediate population independently and stochastically according to the matrix Q. The vector $\boldsymbol{X}^{(t+1)}$ then is the normalized frequency vector of the resulting population.

While this stochastic (RSM) model captures the effect of population size, it is useful only if the mixing time, or the time it takes for the population to reach (close to) its steady state, is much smaller than the size of the state space. For example, in simulations, samples from close to steady state are needed and this is only computationally feasible when the mixing time is small. Moreover, the efficacy of drug design strategies depends on the time it takes the population to evolve to steady state – the mixing time therefore models the minimum required duration of treatment. Further, predictions about the real world, stochastic, version of the population are often made by studying the fixed point of the *deterministic* evolution. In order to argue that these predictions are valid for observed populations, one must show that the finite population reaches close to

---

[1]We follow the convention that a matrix if stochastic if its *columns* sum up to 1.

[2]Note that there is a unique steady state, called the *quasispecies*, when $QA > 0$.

the steady state predicted by the deterministic infinite population dynamics in a reasonable amount of time. However, the number of states in the RSM process is roughly N^m, and this can be prohibitively large. For example, even for a small constant $m = 30$ and a population of size $10,000$, the number of states can grow to more than 2^{300}! While the only way to know that the population is from a distribution close to the steady state is to prove an upper bound on the mixing time of the RSM process, in practice, specific statistics of the population are tracked until they seem to stabilize. This, however, can be misleading and there are well known examples of stochastic processes that appear to get stuck before going to the true steady state.[3]

The importance of obtaining rigorous bounds for mixing time of RSM model was first pointed out in [DSV12]. Rigorous mixing time results are far and few; they have either ignored mutation, assumed that the model is *neutral* (i.e., types have the same fitness), or moved to the diffusion limit which requires both mutation and selection pressure to be *weak*. Recently, the mixing time problem when $m = 2$ for all Q and A was solved [Vis15]. The author proved that the mixing time for the two type case is roughly $\log N$ when all other parameters of the model are constants when compared to N. As discussed in the technical overview, there are significant hurdles to extend this result to $m > 2$ and this problem has remained open since it was raised in [DSV12].

1.1. **Our contribution.**
In this paper we prove that the mixing time of any RSM model parametrized by $m \times m$ matrices Q and A which has a unique stationary distribution is $O(\log N)$ (recall that N is the population size). Thus, we resolve the main problem concerning the RSM model.

Perhaps more interestingly this result turns out to be a corollary of a rapid mixing result for a fairly general class of evolutionary dynamics in finite populations which should have applicability beyond evolutionary biology. Before we describe the result we present the setup: consider an infinite population whose evolution is described by a function $f : \Delta_m \mapsto \Delta_m$ where the population at time t, captured by the vector $\boldsymbol{x}^{(t)} \in \Delta_m$, evolves according to the rule $\boldsymbol{x}^{(t+1)} = f(\boldsymbol{x}^{(t)})$. Inspired by the Wright-Fisher model, we convert an infinite population dynamics to a finite one, say of size N, as follows: imagine that now the fraction of each type is captured by a random vector $\boldsymbol{X}^{(t)} \in \Delta_m$ and the population at time $t + 1$ is obtained by sampling N times independently from the distribution $f(\boldsymbol{X}^{(t)})$. Our main result is that, under mild assumptions on f, the mixing time is bounded by $O(\log N)$.

Theorem 1.1 (Informal statement - see Theorem 3.6). *Let $f : \Delta_m \mapsto \Delta_m$ be an evolution function which satisfies certain mild conditions explained below, and consider the stochastic evolution guided by f on a population of size N. Then, the stochastic evolution converges to its stationary distribution in time $O(\log N)$.*

We first explain the conditions on f required by the theorem; see Definition 3.3 for a formal description. The key conditions are 1) that f has a "unique fixed point" $\boldsymbol{\tau}$ which lies in the interior of Δ_m, 2) that the unique fixed point of f shows "contraction near the fixed point" and 3) that f shows "convergence to fixed point". While it is clear what (1) and (3) mean, the second condition roughly means that there is a $k > 0$ such that if $J(\boldsymbol{x})$ denotes the Jacobian of f at $\boldsymbol{x}$, then for all $\boldsymbol{x}$ which are near $\boldsymbol{\tau}$, $\|J^k(\boldsymbol{x})\|_1 < 1$. For technical reasons, we also need that f is twice differentiable in the interior of Δ_m. While we would have liked the contractive condition to hold for $k = 1$, unfortunately, there are simple examples where this is false. As we shall see later, the fact that we cannot take $k = 1$ is the main source of all the technical difficulties that arise in our work.

While we do not prove that these assumptions on f are necessary for the theorem to hold, there is strong justification for them as we discuss below. Although the case of f having multiple fixed points is interesting in its own right as a model for populations with several distinct "phases", we would expect the stochastic population in this setting to exhibit slow mixing, since it can get stuck close to one of the fixed points and not get the chance to explore the part of the state space near the other fixed points. We therefore concentrate on the case where there is a unique fixed point. The assumption that this fixed point be located in the interior is motivated by the fact that this is the case in most applications, e.g. Eigen's model has this property when $Q > 0$ and $A > 0$, entry-wise. The "contraction near the fixed point" condition ensures that the fixed point

---

[3]In fact, there are striking examples of this, see for instance [MV05, SV07].

is indeed a steady state of the deterministic evolution in the sense of being asymptotically stable: once the evolution reaches close enough to the fixed point, it converges to the latter. Together with the "convergence to fixed point" condition (which, again, is satisfied by models such as Eigen's model), this condition also ensures that the behavior of the deterministic system is free of exotic features, e.g. the presence of cycles, which may present a barrier to the fast mixing of the stochastic evolution The smoothness condition is of a more technical nature, but models our expectation that any evolutionary dynamics should not be too susceptible to small changes in the population profile.

Several other remarks are in order. 1) We treat m, a bound on the derivatives of f, and the rate of contraction as constants and, hence, we do not optimize the dependence on these parameters in this paper. We leave it as an open problem to determine to what extent our results can be generalized when m is a growing function of N. In this paper, our emphasis instead is on circumventing the obstacles that arose in previous attempts to prove a mixing time of close to $O(\log N)$ for general models: earlier works in this direction either put stringent conditions on the parameters of the model (e.g., Dixit et al. had to place very strong conditions on the matrices Q and A for their mixing time result to hold), or were valid only when the number of genotypes was very small (e.g. Vishnoi [Vis15] required the condition $m = 2$). 2) It is not obvious that the desired mixing time result for the RSM model is a corollary of the main theorem and we give a proof in Section 3.2. This result should be of independent interest in population genetics. 3) If one wishes, one can obtain a version of our result for the RSM model in the sampling without replacement setting. We omit the details from this version of the paper. Note however that, for the main theorem, it makes little sense to talk about sampling without replacement.

Finally, we give a quick comparison of our techniques with that of [Vis15] who proved a similar result for $m = 2$. (See the technical overview below for a more detailed discussion.) While [Vis15] also used the underlying deterministic process, it was only to bring two copies of the Markov chain close enough (about a distance $\sqrt{1/N}$). Subsequently, his argument involved the construction of an ad-hoc coupling which contracts when $m = 2$. However, there appear to be serious hurdles when one tries to generalize this coupling for $m > 2$. We bypass these obstacles by again resorting to the properties of f listed above, however, in novel ways. We now move on to illustrating our key techniques.

2. Technical overview

We analyze the mixing time of our stochastic process by studying the time required for evolutions started at two arbitrary starting states $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$ to collide. More precisely, let $\mathscr{C}$ be any Markovian coupling of two stochastic evolutions $\boldsymbol{X}$ and $\boldsymbol{Y}$, both guided by a smooth contractive evolution f, which are started at $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$. Let T be the first (random) time such that $\boldsymbol{X}^{(T)} = \boldsymbol{Y}^{(T)}$. It is well known that if it can be shown that $\mathbb{P}[T > t] \leq 1/4$ for every pair of starting states $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$ then $t_{\mathrm{mix}}(1/4) \leq t$. We show that such a bound on $\mathbb{P}[T > t]$ holds if we couple the chains using the optimal coupling of two multinomial distributions (see Section 3 for a definition of this coupling).

Our starting point is the observation that the optimal coupling and the definition of the evolutions implies that for any time t,

$$\mathbb{E}\left[\left\|\boldsymbol{X}^{(t+1)} - \boldsymbol{Y}^{(t+1)}\right\|_1 \mid \boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)}\right] = \left\|f(\boldsymbol{X}^{(t)}) - f(\boldsymbol{Y}^{(t)})\right\|_1. \tag{1}$$

Now, if f were *globally* contractive, so that the right hand side of eq. (1) was always bounded above by $\rho' \left\|\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)}\right\|_1$ for some constant $\rho' < 1$, then we would get that the expected distance between the two copies of the chains contracts at a constant rate. Since the minimum possible positive ℓ_1 distance between two copies of the chain is $1/N$, this would have implied an $O(\log N)$ mixing time using standard arguments. However, such a global assumption on f, which is equivalent to requiring that the Jacobian J of f satisfies $\|J(\boldsymbol{x})\|_1 < 1$ for all $\boldsymbol{x} \in \Delta_m$, is far too strong. In particular, it is not satisfied by standard systems such as Eigen's dynamics discussed above.

Nevertheless, these dynamics do satisfy a more local version of the above condition. That is, they have a unique fixed point $\boldsymbol{\tau}$ to which they converge quickly, and *in the vicinity* of this fixed point, *some* form of

3

contraction holds. These conditions motivate the "unique fixed point", "contraction near the fixed point", and the "convergence to fixed point" conditions in our definition of a smooth contractive evolution (Definition 3.3). However, crucially, the "contraction near the fixed point" condition, inspired from the definition of "asymptotically stable" fixed points in dynamical systems, is weaker than the stepwise contraction condition described in the last paragraph, even in the vicinity of the fixed point. As we shall see shortly, this weakening is essential for generalizing the earlier results of [Vis15] to the $m > 2$ case, but comes at the cost of making the analysis more challenging.

However, we first describe how the "convergence to fixed point" condition is used to argue that the chains come close to the fixed point in $O(\log N)$ time. This step of our argument is the only one technically quite similar to the development in [Vis15]; our later arguments need to diverge widely from that paper. Although this step is essentially an iterated application of appropriate concentration results along with the fact that the "convergence to fixed point" condition implies that the deterministic evolution f comes close to the fixed point $\boldsymbol{\tau}$ at an exponential rate, complications arise because f can amplify the effect of the random perturbations that arise at each step. In particular, if $L > 1$ is the maximum of $\|J(\boldsymbol{x})\|_1$ over Δ_m, then after ℓ steps, a random perturbation can become amplified by a factor of L^ℓ. As such, if ℓ is taken to be too large, these accumulated errors can swamp the progress made due to the fast convergence of the deterministic evolution to the fixed point. These considerations imply that the argument can only be used for $\ell = \ell_0 \log N$ steps for some small constant ℓ_0, and hence we are only able to get the chains within $\Theta(N^{-\gamma})$ distance of the fixed point, where $\gamma < 1/3$ is a small constant. In particular, the argument cannot be carried out all the way down to distance $O(1/N)$, which, if possible, would have been sufficient to show that the coupling time is small with high probability. Nevertheless, it does allow us to argue that both copies of the chain enter an $O(N^{-\gamma})$ neighborhood of the fixed point in $O(\log N)$ steps.

At this point, [Vis15] showed that in the $m = 2$ case, one could take advantage of the contractive behavior near the fixed point to construct a coupling obeying eq. (1) in which the right hand side was indeed contractive: in essence, this amounted to a proof that $\|J\|_1 < 1$ was indeed satisfied in the small $O(N^{-\gamma})$ neighborhood reached at the end of the last step. This allowed [Vis15] to complete the proof using standard arguments, after some technicalities about ensuring that the chains remained for a sufficiently long time in the neighborhood of the fixed point had been taken care of.

The situation however changes completely in the $m > 2$ case. It is no longer possible to argue in general that $\|J(\boldsymbol{x})\|_1 < 1$ when $\boldsymbol{x}$ is in the vicinity of the fixed point, even when there is fast convergence to the fixed point. Instead, we have to work with a weaker condition (the "contraction to the fixed point" condition alluded to earlier) which only implies that there is a positive integer k, possibly larger than 1, such that in some vicinity of the fixed point, $\|J^k\|_1 < 1$. In the setting used by [Vis15], k could be taken to be 1, and hence it could be argued via eq. (1) that the distance between the two coupled copies of the chains contracts in each step. This argument however does not go through when only a kth power of J is guaranteed to be contractive while J itself could have $1 \to 1$ norm larger than 1. This inability to argue stepwise contraction is the major technical obstacle in our work when compared to the work of [Vis15], and the source of all the new difficulties that arise in this more general setting.

As a first step toward getting around the difficulty of not having stepwise contraction, we prove Theorem 4.1, which shows that the eventual contraction after k steps can be used to ensure that the distance between two evolutions $\boldsymbol{x}^{(t)}$ and $\boldsymbol{y}^{(t)}$ close to the fixed point contracts by a factor $\rho^k < 1$ over an *epoch* of k steps (where k is as described in the last paragraph), even when the evolutions undergo arbitrary perturbations $\boldsymbol{u}^{(t)}$ and $\boldsymbol{v}^{(t)}$ at each step, *provided* that the difference $\boldsymbol{u}^{(t)} - \boldsymbol{v}^{(t)}$ between the two perturbations is small compared to the difference $\boldsymbol{x}^{(t-1)} - \boldsymbol{y}^{(t-1)}$ between the evolutions at the previous step. The last condition actually asks for a relative notion of smallness, i.e., it requires that

$$\left\| \boldsymbol{\xi}^{(t)} \right\|_1 := \left\| \boldsymbol{u}^{(t)} - \boldsymbol{v}^{(t)} \right\|_1 \leq \delta \left\| \boldsymbol{x}^{(t-1)} - \boldsymbol{y}^{(t-1)} \right\|_1, \tag{2}$$

4

where δ is a constant specified in the theorem. Note that the theorem is a statement about deterministic evolutions against possibly adversarial perturbations, and does *not* require $\boldsymbol{u}^{(t)}$ and $\boldsymbol{v}^{(t)}$ to be stochastic, but only that they follow the required conditions on the difference of the norm (in addition to the implied condition that the evolution $\boldsymbol{x}^{(t)}$ and $\boldsymbol{y}^{(t)}$ remain close to the fixed point during the epoch).

Thus, in order to use Theorem 4.1 for showing that the distance $\left\|\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)}\right\|_1$ between the two coupled chains contracts after every k iterations of eq. (1), we need to argue that the required condition on the perturbations in eq. (2) holds with high probability over a given epoch during the coupled stochastic evolution of $\boldsymbol{X}^{(t)}$ and $\boldsymbol{Y}^{(t)}$. (In fact, we also need to argue that the two chains individually remain close to the fixed point, but this is easier to handle and we ignore this technicality in this proof overview; the details appear in the formal proofs in Section 5.)

However, at this point, a complication arises from the fact that Theorem 4.1 requires the difference $\boldsymbol{\xi}^{(t)}$ between the perturbations at time t to be bounded *relative* to the difference $\left\|\boldsymbol{X}^{(t-1)} - \boldsymbol{Y}^{(t-1)}\right\|_1$ at time $t-1$. In other words, the upper bounds required on the $\boldsymbol{\xi}^{(t)}$ become more stringent as the two chains come closer to each other. This fact creates a trade-off between the probability with which the condition in eq. (2) can be enforced in an epoch, and the required *lower* bound on the distance between the chains required during the epoch so as to ensure that probability (this trade-off is technically based on Lemma 3.5). To take a couple of concrete examples, when $\left\|\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)}\right\|_1$ is $\Omega(\log N/N)$ in an epoch, we can ensure that eq. (2) remains valid with probability at least $1 - N^{-\Theta(1)}$ (see the discussion following Theorem 4.4), so that with high probability $\Omega(\log N)$ consecutive epochs admit a contraction allowing the distance between the chains to come down from $\Theta(N^{-\gamma})$ at the end of the first step to $\Theta(\log N/N)$ at the end of this set of epochs.

Ideally, we would have liked to continue this argument till the distance between the chains is $\Theta(1/N)$ and (due to the properties of the optimal coupling) they have a constant probability of colliding in a single step. However, due to the trade-off referred to earlier, when we know only that $\left\|\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)}\right\|_1$ is $\Omega(1/N)$ during the epoch, we can only guarantee the condition of eq. (2) with probability $\Theta(1)$ (see the discussion following the proof of Theorem 4.4). Thus, we cannot claim directly that once the distance between the chains is $O(\log N/N)$, the next $\Omega(\log\log N)$ epochs will exhibit contraction in distance leading the chain to come as close as $O(1/N)$ with a high enough positive probability. To get around this difficulty, we consider $O(\log\log N)$ epochs with successively weaker guaranteed upper bounds on $\left\|\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)}\right\|_1$. Although the weaker lower bounds on the distances lead in turn to weaker concentration results when Theorem 4.4 is applied, we show that this trade-off is such that we can choose these progressively decreasing guarantees so that after this set of epochs, the distance between the chains is $O(1/N)$ with probability that it is small but at least a *constant*. Since the previous steps, i.e., those involving making both chains come within distance $O(N^{-\gamma})$ of the fixed point (for some small constant $\gamma < 1$), and then making sure that the distance between them drops to $O(\log N/N)$, take time $O(\log N)$ with probability $1 - o(1)$, we can conclude that under the optimal coupling, the collision or coupling time T satisfies

$$\mathbb{P}\left[T > O(\log N)\right] \leq 1 - q, \tag{3}$$

for some small enough constant q, irrespective of the starting states $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$ (note that here we are also using the fact that once the chains are within distance $O(1/N)$, the optimal coupling has a constant probability of causing a collision in a single step). The lack of dependence on the starting states allows us to iterate eq. (3) for $\Theta(1)$ consecutive "blocks" of time $O(\log N)$ each to get

$$\mathbb{P}\left[T > O\left(\log N\right)\right] \leq \frac{1}{4},$$

which gives us the claimed mixing time.

3. Preliminaries and formal statement of results

In preparation for formally stating our main result, we now discuss some preliminary notation, definitions and technical tools that will be used later on. We then formally state our main theorem in Sections 3.1 and 3.2.

Notation. We denote the probability simplex on a set of size m as Δ_m. Vectors in $\mathbb{R}^m$, and probability distributions in Δ_m are both denoted in boldface, and x_j denotes the jth co-ordinate of a given vector $\boldsymbol{x}$. Time indices are denoted by superscripts. Thus, a time indexed scalar s at time t is denoted as $s^{(t)}$, while a time indexed vector $\boldsymbol{x}$ at time t is denotes as $\boldsymbol{x}^{(t)}$. The letters $\boldsymbol{X}$ and $\boldsymbol{Y}$ (with time superscripts and co-ordinate subscripts, as appropriate) will be used to denote random vectors. Scalar random vectors and matrices are denoted by other capital letters. Boldface $\mathbf{1}$ denotes a vector all whose entries are 1.

Operators and norms. For any square matrix M, we denote by $\|M\|_1$ its $1 \to 1$ norm defined as $\max_{\|v\|_1=1} \|Mv\|_1$, by $\|M\|_2$ its *operator norm* defined as $\max_{\|v\|_2=1} \|Mv\|_2$, and by $\mathrm{sp}(M)$ its *spectral radius* defined as the maximum of the absolute values of its eigenvalues. 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 3.1 (**Gelfand's formula, specialized to the** $1 \to 1$ **norm**). *For any square matrix M, we have*

$$\mathrm{sp}(M) = \lim_{l \to \infty} \left\|M^l\right\|_1^{1/l}.$$

Derivatives. Let $g : \mathbb{R}^m \to \mathbb{R}^m$ be any differentiable function, whose co-ordinates are denoted as $g_i : \mathbb{R}^m \to R$. The *Jacobian* $J(\boldsymbol{x})$ of g at a point $\boldsymbol{x} \in \mathbb{R}^m$ is the $m \times m$ matrix whose (i,j) entry is $\frac{\partial g_i(\boldsymbol{x})}{\partial x_j}$. The *Hessian* of a twice differentiable function $h : \mathbb{R}^m \to \mathbb{R}$ at a given point $\boldsymbol{x}$ is the $m \times m$ symmetric matrix whose (i,j) entry is $\frac{\partial^2 h(\boldsymbol{x})}{\partial x_i \partial x_j}$. We use the following special case of Taylor's theorem.

Theorem 3.2 (**Taylor's theorem, truncated**). *Let $g : \mathbb{R}^m \to \mathbb{R}^m$ be a twice differentiable function, and let $J(\boldsymbol{z})$ denote the Jacobian of g at $\boldsymbol{z}$. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ be two points, and suppose there exists a positive constant B such that at every point on the line segment joining $\boldsymbol{x}$ to $\boldsymbol{y}$, the Hessians of each of the m co-ordinates g_i of g have operator norm at most $2B$. Then, there exists a $\boldsymbol{v} \in \mathbb{R}^m$ such that*

$$g(\boldsymbol{x}) = g(\boldsymbol{y}) + J(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) + \boldsymbol{v},$$

and $|v_i| \le B \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ for each $i \in [m]$.

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 co-ordinate is equal to $\boldsymbol{p}$ and that on the second co-ordinate is equal to $\boldsymbol{q}$. A simple, if trivial, example of a coupling is the joint distribution obtained by sampling the two co-ordinates independently, one from $\boldsymbol{p}$ and the other from $\boldsymbol{q}$.

Couplings allow a very useful dual characterization of the total variation distance, as stated in the following well known lemma.

Lemma 3.3 (**Coupling lemma**). *Let $\boldsymbol{p}, \boldsymbol{q} \in \Delta_m$ be two probability distributions on m objects. Then,*

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

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

Moreover, the coupling in the lemma can be explicitly described. We use this coupling extensively in our arguments, and hence we record some of its useful properties here.

Definition 3.1 (**Optimal coupling**). Let $\boldsymbol{p}, \boldsymbol{q} \in \Delta_m$ be two probability distributions on m objects. For each $i \in [m]$, let $s_i := \min(p_i, q_i)$, and $s := \sum_{i=1}^m s_i$. Sample U, V, W independently at random as follows:

$$\mathbb{P}[U = i] = \frac{s_i}{s}, \ \mathbb{P}[V = i] = \frac{p_i - s_i}{1 - s}, \ \text{and } \mathbb{P}[W = i] = \frac{q_i - s_i}{1 - s}, \text{ for all } i \in [m].$$

We then sample (independent of U, V, W) a Bernoulli random variable H with mean s. The sample (A, B) given by the coupling is (U, U) if $H = 1$ and (V, W) otherwise.

It is easy to verify that $A \sim \boldsymbol{p}$, $B \sim \boldsymbol{q}$ and $\mathbb{P}[A = B] = s = 1 - \|\boldsymbol{p} - \boldsymbol{q}\|_{\text{TV}}$. Another easily verified but important property is that for any $i \in [m]$

$$\mathbb{P}[A = i, B \neq i] = \begin{cases} 0 & \text{if } p_i < q_i, \\ p_i - q_i & \text{if } p_i \geq q_i. \end{cases}$$

Definition 3.2 (**Mixing time**). Let $\mathscr{M}$ be an ergodic Markov chain on a finite state space Ω with stationary distribution $\boldsymbol{\pi}$. Then, the *mixing time* $t_{\text{mix}}(\varepsilon)$ is defined as the smallest time such that for any starting state $S^{(0)}$, the distribution of the state $S^{(t)}$ at time t is within total variation distance ε of $\boldsymbol{\pi}$. The term mixing time is also used for $t_{\text{mix}}(\varepsilon)$ for a *fixed* values of $\varepsilon < 1/2$. For concreteness, in the rest of this paper, we use t_{mix} to refer to $t_{\text{mix}}(1/e)$ (though any other constant smaller than $1/2$ could be chosen as well in place of $1/e$ without changing any of the claims).

A standard technique for obtaining upper bounds on mixing times is to use the Coupling Lemma above. Suppose $S_1^{(t)}$ and $S_2^{(t)}$ are two evolutions of an ergodic chain M such that their evolutions are coupled according to some coupling $\mathscr{C}$. Let T be the stopping time such that $S_1^{(T)} = S_2^{(T)}$. Then, if it can be shown that $\mathbb{P}[T > t] \leq 1/e$ for every pair of starting states $(S_1^{(0)}, S_2^{(0)})$, then it follows that $t_{\text{mix}} := t_{\text{mix}}(1/e) \leq t$.

Concentration. We now discuss some concentration results that are used extensively in our later arguments. We begin with some standard Chernoff-Hoeffding type bounds.

Theorem 3.4 (**Chernoff-Hoeffding bounds [DP09]**). *Let $Z_1, Z_2, \ldots, Z_N$ be i.i.d Bernoulli random variables with mean μ. We then have*

(1) *When $0 < \delta \leq 1$,*

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mu\right| > \mu\delta\right] \leq 2\exp\left(-N\mu\delta^2/3\right).$$

(2) *When $\delta \geq 1$,*

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mu\right| > \mu\delta\right] \leq \exp\left(-N\mu\delta/3\right).$$

An important tool in our later development is the following lemma, which bounds additional "discrepancies" that can arise when one samples from two distribution $\boldsymbol{p}$ and $\boldsymbol{q}$ using an optimal coupling. The important feature for us is the fact that the additional discrepancy (denoted as $\boldsymbol{e}$ in the lemma) is bounded as a *fraction* of the "initial discrepancy" $\|\boldsymbol{p} - \boldsymbol{q}\|_1$. However, such relative bounds on the discrepancy are less likely to hold when the initial discrepancy itself is very small, and hence, there is a trade-off between the lower bound that needs to be imposed on the initial discrepancy $\|\boldsymbol{p} - \boldsymbol{q}\|_1$, and the desired probability with which the claimed relative bound on the additional discrepancy $\boldsymbol{e}$ is to hold. The lemma makes this delicate trade-off precise.

Lemma 3.5. *Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be probability distributions on a universe of size m, so that $\boldsymbol{p}, \boldsymbol{q} \in \Delta_m$. Consider an optimal coupling of the two distributions, and let $\boldsymbol{x}$ and $\boldsymbol{y}$ be random frequency vectors with m co-ordinates (normalized to sum to 1) obtained by taking N independent samples from the coupled distributions, so that $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{p}$ and $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{q}$. Define the random error vector $\boldsymbol{e}$ as*

$$\boldsymbol{e} := (\boldsymbol{x} - \boldsymbol{y}) - (\boldsymbol{p} - \boldsymbol{q}).$$

Suppose $c > 1$ and t (possibly dependent upon N) are such that $\|\boldsymbol{p} - \boldsymbol{q}\|_1 \geq \frac{ctm}{N}$. We then have

$$\|\boldsymbol{e}\|_1 \leq \left(\frac{2}{\sqrt{c}}\right) \|\boldsymbol{p} - \boldsymbol{q}\|_1$$

with probability at least $1 - 2m\exp(-t/3)$.

Proof. The properties of the optimal coupling of the distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ imply that since the N coupled samples are taken independently,

(1) $|x_i - y_i| = \frac{1}{N}\sum_{j=1}^{N} R_j$, where R_j are i.i.d. Bernoulli random variables with mean $|p_i - q_i|$, and
(2) $x_i - y_i$ has the same sign as $p_i - q_i$.

The second fact implies that $|e_i| = ||x_i - y_i| - |(p_i - q_i)||$. By applying the concentration bounds from Theorem 3.4 to the first fact, we then get (for any arbitrary $i \in [m]$)

$$\mathbb{P}\left[|e_i| > \sqrt{\frac{t}{N|p_i - q_i|}} \cdot |p_i - q_i|\right] \leq 2\exp(-t/3), \text{ if } \frac{t}{N|p_i - q_i|} \leq 1, \text{ and}$$

$$\mathbb{P}\left[|e_i| > \frac{t}{N|p_i - q_i|} \cdot |p_i - q_i|\right] \leq \exp(-t/3), \text{ if } \frac{t}{N|p_i - q_i|} > 1.$$

One of the two bounds applies to every $i \in [m]$ (except those i for which $|p_i - q_i| = 0$, but in those cases, we have $|e_i| = 0$, so the bounds below will apply nonetheless). Thus, taking a union bound over all the indices, we see that with probability at least $1 - 2m\exp(-t/3)$, we have

$$\|\boldsymbol{e}\|_1 = \sum_{i=1}^{m} |e_i| \leq \sqrt{\frac{t}{N}} \sum_{i=1}^{m} \sqrt{|p_i - q_i|} + \frac{tm}{N}$$

$$\leq \sqrt{\frac{tm}{N}} \sqrt{\|\boldsymbol{p} - \boldsymbol{q}\|_1} + \frac{tm}{M} \tag{4}$$

$$\leq \left(\frac{1}{\sqrt{c}} + \frac{1}{c}\right) \|\boldsymbol{p} - \boldsymbol{q}\|_1. \tag{5}$$

Here, eq. (4) uses the Cauchy-Schwarz inequality to bound the first term while eq. (5) uses the hypothesis in the lemma that $\frac{ctm}{N} \leq \|\boldsymbol{p} - \boldsymbol{q}\|_1$. The claim of the lemma follows since $c > 1$. $\qquad \square$

3.1. **Main theorem.** We are now ready to state our main theorem. We begin by formally defining the conditions on the evolution function required by the theorem.

Definition 3.3 (Smooth contractive evolution). A function $f : \Delta_m \to \Delta_m$ is said to be a (L, B, ρ) *smooth contractive evolution* if it has the following properties:

Smoothness: f is twice differentiable in the interior of Δ_m. Further, the Jacobian J of f satisfies $\|J(x)\|_1 \leq L$ for every x in the interior of Δ_m, and the operator norms of the Hessians of its coordinates are uniformly bounded above by $2B$ at all points in the interior of Δ_m.

Unique fixed point: f has a unique fixed point $\boldsymbol{\tau}$ in Δ_m which lies in the interior of Δ_m.

Contraction near the fixed point: At the fixed point $\boldsymbol{\tau}$, the Jacobian $J(\boldsymbol{\tau})$ of f satisfies

$$\text{sp}(J(\boldsymbol{\tau})) < \rho < 1.$$

Convergence to fixed point: For every $\varepsilon > 0$, there exists an ℓ such that for any $x \in \Delta_m$,

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

Remark 3.1. Note that the last condition implies that $\|f^t(x) - \boldsymbol{\tau}\|_1 = O(\rho^t)$ in the light of the previous condition and the smoothness condition (see Lemma A.1). Also, it is easy to see that the last two conditions imply the uniqueness of the fixed point, i.e., the second condition. However, the last condition on global convergence does not by itself imply the third condition on contraction near the fixed point. Consider, e.g., $g : [-1, 1] \to [-1, 1]$ defined as $g(x) = x - x^3$. The unique fixed point of g in its domain is 0, and we have $g'(0) = 1$, so that the third condition is not satisfied. On the other hand, the last condition is satisfied, since for $x \in [-1, 1]$ satisfying $|x| \geq \varepsilon$, we have $|g(x)| \leq |x|(1 - \varepsilon^2)$. In order to construct a function $f : [0, 1] \to [0, 1]$ with the same properties, we note that the range of g is $[-x_0, x_0]$ where $x_0 = 2/(3\sqrt{3})$, and consider $f : [0, 1] \to [0, 1]$ defined as $f(x) = x_0 + g(x - x_0)$. Then, the unique fixed point of f in $[0, 1]$ is x_0, $f'(x_0) = g'(0) = 1$, the range of f is contained in $[0, 2x_0] \subseteq [0, 1]$, and f satisfies the fourth condition in the definition but does not satisfy the third condition.

Given an f which is a smooth contractive evolution, and a population parameter N, we can define a 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 frequency vector over $[m]$. Then

$$\boldsymbol{x}^{(t+1)} := \frac{1}{N}\boldsymbol{z}^{(t)}.$$

Theorem 3.6 (**Main Theorem**). *Let f be a (L,B,ρ) smooth contractive evolution, and consider the stochastic evolution guided by f on a population of size N. Then, there exists c_0 and N_0 depending upon L, B and ρ and f such that for any $N > N_0$, the mixing time $t_{\mathrm{mix}}(1/e)$ of the stochastic evolution is at most $c_0 \left((\log N)\right)$.*

3.2. **The RSM model as a special case.** We now show that the Eigen or RSM model discussed in the introduction is a special case of the abstract model defined in the last section, and hence satisfies the mixing time bound in Theorem 3.6. Our first step is to show that the RSM model can be seen as a stochastic evolution guided by the function f defined by $f(\boldsymbol{p}) = \frac{(QA\boldsymbol{p})_t}{\|QA\boldsymbol{p}\|_1}$, where Q and A are matrices with positive entries, with Q stochastic (i.e., columns summing up to 1), as described in the introduction. We will then show that this f is a smooth contractive evolution, which implies that Theorem 3.6 applies to the RSM process.

We begin by recalling the definition of the RSM process. Given a starting population of size N on m types represented by a $1/N$-integral probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)$, the RSM process produces the population at the next step by independently sampling N times from the following process:

(1) Sample a type T from the probability distribution $\frac{A\boldsymbol{p}}{\|A\boldsymbol{p}\|_1}$.
(2) Mutate T to the result type S with probability Q_{ST}.

We now show that sampling from this process is exactly the same as sampling from the multinomial distribution $f(\boldsymbol{p}) = \frac{QA\boldsymbol{p}}{\|QA\boldsymbol{p}\|_1}$. To do this, we only need to establish the following claim:

Claim 3.7. *For any type $t \in [m]$, $\mathbb{P}\left[S = t\right] = \frac{\sum_j Q_{tj} A_{jj} p_j}{\sum_j A_{jj} p_j} = \frac{(QA\boldsymbol{p})_t}{\|QA\boldsymbol{p}\|_1}$.*

Proof. We first note that $\|QA\boldsymbol{p}\|_1 = \sum_{ij} Q_{ij} A_{jj} p_j = (\sum_i Q_{ij}) \cdot (\sum_j A_{jj} p_j) = \sum_j A_{jj} p_j = \|A\boldsymbol{p}\|_1$, where in the last equality we used the fact that the columns of Q sum up to 1. Now, we have

$$\mathbb{P}\left[S = t\right] := \sum_i^m Q_{ti} \cdot \frac{(A\boldsymbol{p})_i}{\|A\boldsymbol{p}\|_1} = \frac{\sum_{i=1} Q_{ti} A_{ii} p_i}{\sum_j A_{jj} p_j} = \frac{(QA\boldsymbol{p})_t}{\|QA\boldsymbol{p}\|_1}. \qquad \square$$

From Claim 3.7, we see that producing N independent samples from the process described above (which corresponds exactly to the RSM model) produces the same distribution as producing N independent samples from the distribution $\frac{(QA\boldsymbol{p})}{\|QA\boldsymbol{p}\|_1}$. Thus, the RSM process is a stochastic evolution guided by $f(\boldsymbol{x}) := \frac{QA\boldsymbol{x}}{\|QA\boldsymbol{x}\|_1}$. We now proceed to verify that this f is a smooth contractive evolution. We first note that the "smoothness" condition is directly implied by the definition of f. For the "uniqueness of fixed point" condition, we observe that every fixed point of $\frac{QA\boldsymbol{x}}{\|QA\boldsymbol{x}\|_1}$ in the simplex Δ_m must be an eigenvector of QA. Since QA is a matrix with positive entries, the Perron-Frobenius theorem implies that it has a unique positive eigenvector $\boldsymbol{v}$ (for which we can assume without loss of generality that $\|\boldsymbol{v}\|_1 = 1$) with a positive eigenvalue λ_1. Therefore $f(\boldsymbol{x})$ has a unique fixed point $\boldsymbol{\tau} = \boldsymbol{v}$ in the simplex Δ_m which is in its interior. The Perron-Frobenius theorem also implies that for every $\boldsymbol{x} \in \Delta_m$, $\lim_{t \to \infty}(QA)^t \boldsymbol{x}/\lambda_1^t \to \boldsymbol{v}$. In fact, this convergence can be made uniform over Δ_m (meaning that given an $\varepsilon > 0$ we can choose t_0 such that for all $t > t_0$, $\|(QA)^t \boldsymbol{x}/\lambda_1^t - \boldsymbol{v}\|_1 < \varepsilon$ for all $\boldsymbol{x} \in \Delta_m$) since each point $\boldsymbol{x} \in \Delta_m$ is a convex combination of the extreme points of Δ_m and the left hand side is a linear function of $\boldsymbol{x}$. From this uniform convergence, it then follows easily that $\lim_{t \to \infty} f^t(\boldsymbol{x}) = \boldsymbol{v}$, and that the convergence in this limit is also uniform. The "convergence to fixed point" condition follows directly from this observation.

Finally, we need to establish that the spectral radius of the Jacobian $J := J(\boldsymbol{v})$ of f at its fixed point is less than 1. A simple computation shows that the Jacobian at $\boldsymbol{v}$ is $J = \frac{1}{\lambda_1}(I - V)QA$ where V is the matrix each of

9

whose columns is the vector $\boldsymbol{v}$. Since QA has positive entries, we know from the Perron-Frobenius theorem that λ_1 as defined above is real, positive, and strictly larger in magnitude than any other eigenvalue of QA. Let $\lambda_2, \lambda_3, \ldots, \lambda_m$ be the other, possibly complex, eigenvalues arranged in decreasing order of magnitude (so that $\lambda_1 > |\lambda_2|$). We now establish the following claim from which it immediately follows that $\text{sp}(J) = \frac{|\lambda_2|}{\lambda_1} < 1$ as required.

Claim 3.8. *The eigenvalues of $M := (I - V)QA$ are $\lambda_2, \lambda_3, \ldots, \lambda_m, 0$.*

Proof. Let D be the Jordan canonical form of QA, so that $D = U^{-1}QAU$ for some invertible matrix U. Note that D is an upper triangular matrix with $\lambda_1, \lambda_2, \ldots, \lambda_m$ on the diagonal. Further, the Perron-Frobenius theorem applied to QA implies that λ_1 is an eigenvalue of both algebraic and geometric multiplicity 1, so that we can assume that the topmost Jordan block in D is of size 1 and is equal to λ_1. Further, we can assume the corresponding first column of U is equal to the corresponding positive eigenvector $\boldsymbol{v}$ satisfying $\|\boldsymbol{v}\|_1 = 1$. It therefore follows that $U^{-1}V = U^{-1}\boldsymbol{v}\mathbf{1}^T$ is the matrix $\boldsymbol{e}_1\mathbf{1}^T$, where $\boldsymbol{e}_1$ is the first standard basis vector.

Now, since U is invertible, M has the same eigenvalues as $U^{-1}MU = (U^{-1} - U^{-1}V)QAU = (I - \boldsymbol{e}_1\mathbf{1}^T U)D$, where in the last line we use $UD = QAU$. Now, note that all rows except the first of the matrix $\boldsymbol{e}_1\mathbf{1}^T U$ are zero, and its $(1,1)$ entry is 1 since the first column of U is $\boldsymbol{v}$, which in turn is chosen so that $\mathbf{1}^T\boldsymbol{v} = 1$. Thus, we get that $(I - \boldsymbol{e}_1\mathbf{1}^T U)D$ is an upper triangular matrix with the same diagonal entries as D except that its $(1,1)$ entry is 0. Since the $(1,1)$ entry of D was λ_1 while its other diagonal entries were $\lambda_2, \lambda_3, \ldots, \lambda_m$, it follows that the eigenvalues of $(I - \boldsymbol{e}_1\mathbf{1}^T U)D$ (and hence those of M) are $\lambda_2, \lambda_2, \ldots, \lambda_m, 0$, as claimed. □

We thus see that the RSM process satisfies the condition of being guided by a smooth contractive evolution and hence has the mixing time implied by Theorem 3.6.

4. Perturbed evolution near the fixed point

As discussed in Section 2, the crux of the proof of our main theorem is analyzing how the distance between two copies of a stochastic evolution guided by a smooth contractive evolution evolves in the presence of small perturbations at every step. In this section, we present our main tool, Theorem 4.1, to study this phenomenon. We then describe how the theorem, which itself is presented in a completely deterministic setting, applies to stochastic evolutions.

Fix any (L, B, ρ)-smooth contractive evolution f on Δ_m, with fixed point $\boldsymbol{\tau}$. As we noted in Section 2, since the Jacobian of f does not necessarily have operator (or $1 \to 1$) norm less than 1, we cannot argue that the effect of perturbations shrinks in every step. Instead, we need to argue that the condition on the spectral radius of the Jacobian of f at its fixed point implies that there is eventual contraction of distance between the two evolutions, even though this distance might increase in any given step. Indeed, the fact that the spectral radius $\text{sp}(J)$ of the Jacobian at the fixed point $\boldsymbol{\tau}$ of f is less than $\rho < 1$ implies that a suitable iterate of f has a Jacobian with operator (and $1 \to 1$) norm less than 1 at $\boldsymbol{\tau}$. This is because Gelfand's formula (Theorem 3.1) implies that for all large enough positive integers k',

$$\left\| J^k(\boldsymbol{\tau}) \right\|_1 < \rho^k.$$

We now use the above condition to argue that after k steps in the vicinity of the fixed point, there is indeed a contraction of the distance between two evolutions guided by f, even in the presence of adversarial perturbations, as long as those perturbations are small. The precise statement is given below; the vectors $\boldsymbol{\xi}^{(i)}$ in the theorem model the small perturbations.

Theorem 4.1 (**Perturbed evolution**). *Let f be a (L, B, ρ)-smooth contractive evolution, and let $\boldsymbol{\tau}$ be its fixed point. For all positive integers $k > k_0$ (where k_0 is a constant that depends upon f) there exist $\varepsilon, \delta \in (0, 1]$ depending upon f and k for which the following is true. Let $\left(\boldsymbol{x}^{(i)} \right)_{i=0}^k$, $\left(\boldsymbol{y}^{(i)} \right)_{i=0}^k$, and $\left(\boldsymbol{\xi}^{(i)} \right)_{i=1}^k$ be sequences of vectors with $\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)} \in \Delta_m$ and $\boldsymbol{\xi}^{(i)}$ orthogonal to $\mathbf{1}$, which satisfy the following conditions:*

(1) **(Definition).** *For* $1 \leq i \leq k$, *there exist vectors* $\boldsymbol{u}^{(i)}$ *and* $\boldsymbol{v}^{(i)}$ *such that*

$$\boldsymbol{x}^{(i)} = f(\boldsymbol{x}^{(i-1)}) + \boldsymbol{u}^{(i)}, \ \boldsymbol{y}^{(i)} = f(\boldsymbol{y}^{(i-1)}) + \boldsymbol{v}^{(i)}, \ \text{and } \boldsymbol{\xi}^{(i)} = \boldsymbol{u}^{(i)} - \boldsymbol{v}^{(i)}.$$

(2) **(Closeness to fixed point).** *For* $0 \leq i \leq k$, $\left\| \boldsymbol{x}^{(i)} - \boldsymbol{\tau} \right\|_1 \leq \varepsilon$ *and* $\left\| \boldsymbol{y}^{(i)} - \boldsymbol{\tau} \right\|_1 \leq \varepsilon$.

(3) **(Small perturbations).** *For* $1 \leq i \leq k$, $\left\| \boldsymbol{\xi}^{(i)} \right\|_1 \leq \delta \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1$.

Then, we have

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

In the theorem, the vectors $\boldsymbol{x}^{(i)}$ and $\boldsymbol{y}^{(i)}$ model the two chains, while the vectors $\boldsymbol{u}^{(i)}$ and $\boldsymbol{v}^{(i)}$ model the individual perturbations from the evolution dictated by f. The theorem says that if the perturbations $\boldsymbol{\xi}^{(i)}$ to the *distance* are not too large, then the distance between the two chains indeed contracts after every k steps.

Proof. As observed above, we can use Gelfand's formula to conclude that there exists a positive integer k_0 (depending upon f) such that we have $\left\| J(\tau)^k \right\|_1 < \rho^k$ for all $k > k_0$. This k_0 will be the sought k_0 in the theorem, and we fix some appropriate $k > k_0$ for the rest of the paper.

Since f is twice differentiable, J is continuous on Δ_m. This implies that the function on Δ_m^k defined by $z_1, z_2, \ldots, z_k \mapsto \prod_{i=1}^k J(z_i)$ is also continuous. Hence, there exist $\varepsilon_1, \varepsilon_2 > 0$ smaller than 1 such that if $\| z_i - \boldsymbol{\tau} \| \leq \varepsilon_1$ for $1 \leq i \leq k$ then

$$\left\| \prod_{i=1}^k J(z_i) \right\|_1 \leq \rho^k - \varepsilon_2. \tag{6}$$

Further, since Δ_m is compact and f is continuously differentiable, $\|J\|_1$ is bounded above on Δ_m by some positive constant L, which we assume without loss of generality to be greater than 1. Similarly, since f has bounded second derivatives, it follows from the multivariate Taylor's theorem that there exists a positive constant B (which we can again assume to be greater than 1) such that for any $\boldsymbol{x}, \boldsymbol{y} \in \Delta_m$, we can find a vector $\boldsymbol{v}$ such that $\|\boldsymbol{v}\|_1 \leq Bm \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leq Bm \|\boldsymbol{x} - \boldsymbol{y}\|_1^2$ such that

$$f(\boldsymbol{x}) = f(\boldsymbol{y}) + J(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) + \boldsymbol{v}. \tag{7}$$

We can now choose

$$\varepsilon = \min\left\{ \varepsilon_1, \frac{\varepsilon_2}{4Bmk(L+1)^{k-1}} \right\} \leq 1, \ \text{and } \delta = 2Bm\varepsilon \leq 1.$$

With this setup, we are now ready to proceed with the proof. Our starting point is the use of a first order Taylor expansion to control the error $\boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)}$ in terms of $\boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)}$. Indeed, eq. (7) when applied to this situation (along with the hypotheses of the theorem) yields for any $1 \leq i \leq k$ that

$$\boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)} = f(\boldsymbol{x}^{(i-1)}) - f(\boldsymbol{y}^{(i-1)}) + \boldsymbol{\xi}^{(i)} \tag{8}$$

$$= J(\boldsymbol{y}^{(i-1)})(\boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)}) + (\boldsymbol{v}^{(i)} + \boldsymbol{\xi}^{(i)}), \tag{9}$$

where $\boldsymbol{v}^{(i)}$ satisfies $\left\| \boldsymbol{v}^{(i)} \right\|_1 \leq Bm \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1^2$. Before proceeding, we first take note of a simple consequence of eq. (9). Taking the ℓ_1 norm of both sides, and using the conditions on the norms of $\boldsymbol{v}^{(i)}$ and $\boldsymbol{\xi}^{(i)}$, we have

$$\left\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)} \right\|_1 \leq \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 \left(L + \delta + Bm \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 \right).$$

Since both $\boldsymbol{x}^{(i-1)}$ and $\boldsymbol{y}^{(i-1)}$ are within distance ε of τ by the hypothesis of the theorem, the above calculation and the definition of ε and δ imply that

$$\left\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)} \right\|_1 \leq (L + 4Bm\varepsilon) \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 \leq (L+1) \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1,$$

where in the last inequality we use $4Bm\varepsilon \leq \varepsilon_2 \leq 1$. This, in turn, implies via induction that for every $1 \leq i \leq k$,

$$\left\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)} \right\|_1 \leq (L+1)^i \left\| \boldsymbol{x}^{(0)} - \boldsymbol{y}^{(0)} \right\|_1 . \tag{10}$$

We now return to the proof. By iterating eq. (9), we can control $\boldsymbol{x}^{(k)} - \boldsymbol{y}^{(k)}$ in terms of a product of k Jacobians, as follows:

$$\boldsymbol{x}^{(k)} - \boldsymbol{y}^{(k)} = \left(\prod_{i=1}^{i=k} J(\boldsymbol{y}^{(k-i)}) \right) \left(\boldsymbol{x}^{(0)} - \boldsymbol{y}^{(0)} \right) + \sum_{i=1}^{k} \left(\prod_{j=0}^{j=k-i-1} J(\boldsymbol{y}^{(k-j)}) \right) \left(\boldsymbol{\xi}^{(i)} + \boldsymbol{v}^{(i)} \right).$$

Since $\left\| \boldsymbol{y}^{(i)} - \boldsymbol{\tau} \right\|_1 \leq \varepsilon$ by the hypothesis of the theorem, we get from eq. (6) that the leftmost term in the above sum has ℓ_1 norm less than $(\rho^k - \varepsilon_2) \left\| \boldsymbol{x}^{(0)} - \boldsymbol{y}^{(0)} \right\|_1$.

We now proceed to estimate the terms in the summation. Our first step to use the conditions on the norms of $\boldsymbol{v}^{(i)}$ and $\boldsymbol{\xi}^{(i)}$ and the fact that $\|J\|_1 \leq L$ uniformly to obtain the upper bound

$$\sum_{i=1}^{k} L^{k-i} \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 \left(Bm \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 + \delta \right).$$

Now, recalling that $\boldsymbol{x}^{(i)}$ and $\boldsymbol{y}^{(i)}$ are both within an ε ℓ_1-neighborhood of $\boldsymbol{\tau}$ so that $\left\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)} \right\|_1 \leq 2\varepsilon$, we can estimate the above upper bound as follows:

$$\sum_{i=1}^{k} L^{k-i} \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 \left(Bm \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 + \delta \right) \leq (L+1)^{k-1} \left\| \boldsymbol{x}^{(0)} - \boldsymbol{y}^{(0)} \right\|_1 \sum_{i=1}^{k} \left(Bm \left\| \boldsymbol{x}^{(i-1)} - \boldsymbol{y}^{(i-1)} \right\|_1 + \delta \right)$$

$$\leq k(L+1)^{k-1} (2Bm\varepsilon + \delta) \left\| \boldsymbol{x}^{(0)} - \boldsymbol{y}^{(0)} \right\|_1$$

$$\leq \varepsilon_2 \left\| \boldsymbol{x}^{(0)} - \boldsymbol{y}^{(0)} \right\|_1 ,$$

where the first inequality is an application of eq. (10), and the last uses the definitions of ε and δ. Combining with the upper bound of $(\rho^k - \varepsilon_2) \left\| \boldsymbol{x}^{(0)} - \boldsymbol{y}^{(0)} \right\|_1$ obtained above for the first term, this yields the result. $\quad\square$

Remark 4.1. Note that k in the theorem can be chosen as large as we want. However, for simplicity, we fix some $k > k_0$ in the rest of the discussion, and revisit the freedom of choice of k only toward the end of the proof of the main theorem (Theorem 3.6) on page 17.

4.1. **Evolution under random perturbations.** We now explore some consequences of the above theorem for stochastic evolutions. Our main goal in this subsection is to highlight the subtleties that arise in ensuring that the third "small perturbations" condition during a random evolution, and strategies that can be used to avoid them. However, we first begin by showing the second condition, that of "closeness to the fixed point" is actually quite simple to maintain. It will be convenient to define for this purpose the notion of an *epoch*, which is simply the set of $(k+1)$ initial and final states of k consecutive steps of a stochastic evolution.

Definition 4.1 (**Epoch**). Let f be a smooth contractive evolution and let k be as in the statement of Theorem 4.1 when applied to f. An *epoch* is a set of $k+1$ consecutive states in a stochastic evolution guided by f. By a slight abuse of terminology we also use the same term to refer to a set of $k+1$ consecutive states in a *pair* of stochastic evolutions guided by f that have been coupled using the optimal coupling.

Suppose we want to apply Theorem 4.1 to a pair of stochastic evolutions guided by f. Recall the parameter ε in the statement of Theorem 4.1. Ideally, we would likely to show that if both the states in the pair at the beginning of an epoch are within some distance $\varepsilon' < \varepsilon$ of the fixed point $\boldsymbol{\tau}$, then (1) all the consequent steps in the epoch are within distance ε of the fixed point (so that the closeness condition in the theorem is satisfied), and more importantly (2) that the states at the last step of the epoch are *again* within the same distance ε' of the fixed point, so that we have the ability to apply the theorem to the *next* epoch. Of course, we also need to ensure that the condition on the perturbations being true also holds during the epoch, but as stated above,

this is somewhat more tricky to maintain than the closeness condition, so we defer its discussion to later in the section. Here, we state the following lemma which shows that the closeness condition can indeed be maintained at the end of the epoch.

Lemma 4.2 (**Remaining close to the fixed point**). *Let $w < w' < 1/3$ be fixed constants. Consider a stochastic evolution $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots$ on a population of size N guided by a (L, B, ρ)-smooth contractive evolution $f : \Delta_m \to \Delta_m$ with fixed point $\boldsymbol{\tau}$. Suppose $\alpha > 1$ is such that $\left\| \mathbf{X}^{(0)} - \boldsymbol{\tau} \right\|_1 \le \frac{\alpha}{N^w}$. If N is chosen large enough (as a function of L, α, m, k, w, w' and ρ), then with probability at least $1 - 2mk \exp\left(-N^{w'}/2 \right)$ we have*

- $\left\| \mathbf{X}^{(i)} - \boldsymbol{\tau} \right\|_1 \le \frac{(\alpha+m)(L+1)^i}{N^w}$, *for $1 \le i \le k-1$.*
- $\left\| \mathbf{X}^{(k)} - \boldsymbol{\tau} \right\|_1 \le \frac{\alpha}{N^w}$.

To prove the lemma, we need the following simple concentration result the proof of which is deferred to Appendix B.

Lemma 4.3. *Let $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots$ be a stochastic evolution on a population of size N which is guided by a (L, B, ρ)-smooth contractive evolution $f : \Delta_m \to \Delta_m$ with fixed point $\boldsymbol{\tau}$. For any $t > 0$ and $\gamma \le 1/3$, it holds with probability at least $1 - 2mt \exp(-N^\gamma/2)$ that*

$$\left\| \mathbf{X}^{(i)} - f^i(\mathbf{X}^{(0)}) \right\|_1 \le \frac{(L+1)^i m}{N^\gamma} \text{ for all } 1 \le i \le t.$$

Proof of Lemma 4.2. Lemma 4.3 implies that with probability at least $1 - 2mk \exp\left(-N^{w'}/2 \right)$ we have

$$\left\| \mathbf{X}^{(i)} - f^i(\mathbf{X}^{(0)}) \right\|_1 \le \frac{(L+1)^i m}{N^{w'}}, \text{ for } 1 \le i \le k. \tag{11}$$

On the other hand, the fact that $\max \| J(\mathbf{x}) \|_1 \le L$ implies that

$$\left\| f^{i+1}(\mathbf{X}^{(0)}) - \boldsymbol{\tau} \right\|_1 = \left\| f^{i+1}(\mathbf{X}^{(0)}) - f(\boldsymbol{\tau}) \right\|_1 \le L \left\| f^i(\mathbf{X}^{(0)}) - \boldsymbol{\tau} \right\|_1,$$

so that

$$\left\| f^i(\mathbf{X}^{(0)}) - \boldsymbol{\tau} \right\|_1 \le L^i \left\| \mathbf{X}^{(0)} - \boldsymbol{\tau} \right\|_1 \le \frac{\alpha L^i}{N^w}, \text{ for } 1 \le i \le k. \tag{12}$$

Combining eqs. (11) and (12), we already get the first item in the lemma. However, for $i = k$, we can do much better than the above estimate (and indeed, this is the most important part of the lemma). Recall the parameter ε in Theorem 4.1. If we choose N large enough so that

$$\frac{(\alpha + m)(L+1)^k}{N^w} \le \varepsilon, \tag{13}$$

then the above argument shows that with probability at least $1 - 2mk \exp\left(-N^{w'} \right)$, the sequences $\mathbf{y}^{(i)} = f^i(\mathbf{X}^{(0)})$ and $\mathbf{z}^{(i)} = \boldsymbol{\tau}$ (for $0 \le i \le k$) satisfy the hypotheses of Theorem 4.1: the perturbations in this case are simply 0. Hence, we get that

$$\left\| f^k(\mathbf{X}^{(0)}) - \boldsymbol{\tau} \right\|_1 \le \rho^k \left\| \mathbf{X}^{(0)} - \boldsymbol{\tau} \right\|_1 \le \frac{\alpha \rho^k}{N^w}. \tag{14}$$

Using eq. (14) with the $i = k$ case of eq. (11), we then have

$$\left\| \mathbf{X}^{(k)} - \boldsymbol{\tau} \right\|_1 \le \frac{(L+1)^k m}{N^{w'}} + \frac{\rho^k \alpha}{N^w}.$$

Thus, (since $\rho < 1$) we only need to choose N so that

$$N^{w'-w} \ge \frac{(L+1)^k m}{\alpha(1-\rho^k)} \tag{15}$$

13

in order to get the second item in the lemma. Since $w > 0$ and $w' > w$, it follows that all large enough N will satisfy the conditions in both eqs. (13) and (15), and this completes the proof. $\qquad\square$

4.2. Controlling the size of random perturbations.

We now address the "small perturbations" condition of Theorem 4.1. For a given smooth contractive evolution f, let α, w, w' be any constants satisfying the hypotheses of Lemma 4.2 (the precise values of these constants will specified in the next section). For some N as large as required by the lemma, consider a pair $\boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)}$ of stochastic evolutions guided by f on a population of size N, which are coupled according to the optimal coupling. Now, let us call an epoch *decent* if the first states $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$ in the epoch satisfy $\left\| \boldsymbol{X}^{(0)} - \boldsymbol{\tau} \right\|_1, \left\| \boldsymbol{Y}^{(0)} - \boldsymbol{\tau} \right\|_1 \le \alpha N^{-w}$. The lemma (because of the choice of N made in eq. (13)) shows that if an epoch is decent, then except with probability that is sub-exponentially small in N,

   (1) the current epoch satisfies the "closeness to fixed point" condition in Theorem 4.1, and
   (2) the *next* epoch is decent as well.

Thus, the lemma implies that if a certain epoch is decent, then with all but sub-exponential (in N) probability, a polynomial (in N) number of subsequent epochs are also decent, and hence satisfy the "closeness to fixed point" condition of Theorem 4.1. Hypothetically, if these epochs also satisfied the "small perturbation" condition, then we would be done, since in such a situation, the distance between the two chains will drop to less than $1/N$ within $O(\log N)$ time, implying that they would collide. This would in turn imply a $O(\log N)$ mixing time.

However, as alluded to above, ensuring the "small perturbations" condition turns out to be more subtle. In particular, the fact that the perturbations $\boldsymbol{\xi}^{(i)}$ need to be *multiplicatively* smaller than the actual differences $\left\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)} \right\|_1$ pose a problem in achieving adequate concentration, and we cannot hope to prove that the "small perturbations" condition holds with very high probability over an epoch when the staring difference $\left\| \boldsymbol{X}^{(0)} - \boldsymbol{Y}^{(0)} \right\|_1$ is very small. As such, we need to break the arguments into two stages based on the starting differences at the start of the epochs lying in the two stages.

To make this more precise (and to state a result which provides examples of the above phenomenon and will also be a building block in the coupling proof), we define the notion of an epoch being *good* with a *goal g*. As before let $\boldsymbol{X}^{(t)}$ and $\boldsymbol{Y}^{(t)}$ be two stochastic evolutions guided by f which are coupled according to the optimal coupling, and let $\boldsymbol{\xi}^{(i)}$ be the perturbations as defined in Theorem 4.1. Then, we say that a decent epoch (which we can assume, without loss of generality, to start at $t = 0$) is *good* with goal g if one of following two conditions holds. Either (1) there is a j, $0 \le j \le k - 1$ such that $\left\| f(\boldsymbol{X}^{(j)}) - f(\boldsymbol{Y}^{(j)}) \right\|_1 \le g$, or otherwise, (2) it holds that the next epoch is also decent, and, further

$$\left\| \boldsymbol{\xi}^{(i)} \right\|_1 \le \delta \left\| \boldsymbol{X}^{(i)} - \boldsymbol{Y}^{(i)} \right\|_1 \text{ for } 0 \le i \le k,$$

where δ again is as defined in Theorem 4.1. Note that if an epoch is good with goal g, then either the expected difference between the two chains drops below g sometime during the epoch, or else, all conditions of Theorem 4.1 are satisfied during the epoch, and the distance between the chains drops by a factor of ρ^k while the next epoch is also decent. Further, in terms of this notion, the preceding discussion can be summarized as "the probability of an epoch being good depends upon the goal g, and can be small if g is too small". To make this concrete, we prove the following theorem which quantifies this trade-off between the size of the goal g and the probability with which an epoch is good with that goal.

Theorem 4.4 (Goodness with a given goal). *Let the chains $\boldsymbol{X}^{(t)}$, $\boldsymbol{Y}^{(t)}$, and the quantities $N, m, w, w', k, L, \varepsilon$ and δ be as defined above, and let $\beta < (\log N)^2$. If N is large enough, then a decent epoch is good with goal $g := \frac{4L^2 m \beta}{\delta^2 N}$ with probability at least $1 - 2mk \left(\exp(-\beta/3) + \exp(-N^{w'}/2) \right)$.*

Proof. Let $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$ denote the first states in the epoch. Since the current epoch is assumed to be decent, Lemma 4.2 implies that with probability at least $1 - 2mk \exp(-N^{-w'}/2)$, the "closeness to fixed point"

14

condition of Theorem 4.1 holds throughout the epoch, and the next epoch is also decent. If there is a $j \leq k-1$ such that $\left\| f(\boldsymbol{X}^{(j)}) - f(\boldsymbol{Y}^{(j)}) \right\|_1 \leq g$, then the current epoch is already good with goal g. So let us assume that

$$\left\| f(\boldsymbol{X}^{(i-1)}) - f(\boldsymbol{Y}^{(i-1)}) \right\|_1 \geq g = \frac{4L}{\delta^2} \cdot \frac{\beta m}{N} \text{ for } 1 \leq i \leq k.$$

However, in this case, we can apply the concentration result in Lemma 3.5 with $c = 4L^2/\delta^2$ and $t = \beta$ to get that with probability at least $1 - 2mk\exp(-\beta/3)$,

$$\left\| \boldsymbol{\xi}^{(i)} \right\|_1 \leq \frac{\delta}{L} \left\| f(\boldsymbol{X}^{(i-1)}) - f(\boldsymbol{Y}^{(i-1)}) \right\|_1 \leq \delta \left\| \boldsymbol{X}^{(i-1)} - \boldsymbol{Y}^{(i-1)} \right\|_1 \text{ for } 1 \leq i \leq k.$$

Hence, both conditions ("closeness to fixed point" and "small perturbations") for being good with goal g hold with the claimed probability. $\square$

Note that we need to take β to a large constant, at least $\Omega(\log(mk))$ even to make the result non-trivial. In particular, if we take $\beta = 3\log(4mk)$, then if N is large enough, the probability of success is at least $1/e$. However, with a slightly larger goal g, it is possible to reduce the probability of an epoch not being good to $o_N(1)$: if we choose $\beta = \log N$, then a decent epoch is good with the corresponding goal with probability at least $1 - N^{-1/4}$, for N large enough. In the next section, we use both these settings of parameters in the above theorem to complete the proof of the mixing time result. As described in Section 2, the two settings above will be used in different stages of the evolution of two coupled chains in order to argue that the time to collision of the chains is indeed small.

5. Proof of the main theorem: Analyzing the coupling time

Our goal is now to show that if we couple two stochastic evolutions guided by the same smooth contractive evolution f using the optimal coupling, then irrespective of their starting positions, they reach the same state in a small number of steps, with reasonably high probability. More precisely, our proof would be structured as follows. Fix any starting states $\boldsymbol{X}^{(0)}$ and $\boldsymbol{Y}^{(0)}$ of the two chains, and couple their evolutions according to the optimal coupling. Let T be the first time such that $\boldsymbol{X}^{(T)} = \boldsymbol{Y}^{(T)}$. Suppose that we establish that $\mathbb{P}[T < t] \geq q$, where t and p do not depend upon the starting states $(\boldsymbol{X}^{(0)}, \boldsymbol{Y}^{(0)})$. Then, we can dovetail this argument for ℓ "windows" of time t each to see that $\mathbb{P}[T > \ell \cdot t] \leq (1-q)^\ell$: this is possible because the probability bounds for T did not depend upon the starting positions $(\boldsymbol{X}^{(0)}, \boldsymbol{Y}^{(0)})$ and hence can be applied again to the starting positions $(\boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)})$ if $\boldsymbol{X}^{(t)} \neq \boldsymbol{Y}^{(t)}$. By choosing ℓ large enough so that $(1-q)^\ell$ is at most $1/e$ (or any other constant less than $1/2$), we obtain a mixing time of ℓt. We therefore proceed to obtain an upper bound on $\mathbb{P}[T < t]$ for some $t = \Theta(\log N)$.

As discussed earlier, we need to split the evolution of the chains into several stages in order to complete the argument outlined above. We now describe these four different stages. Recall that f is assumed to be a (L, B, ρ)-smooth contractive evolution. Without loss of generality we assume that $L > 1$. The parameter r appearing below is a function of these parameters and k and is defined in Lemma A.1. Further, as we noted after the proof of Theorem 4.1, k can be chosen to be as large as desired. We now exercise this choice by choosing k to be large enough so that

$$\rho^k \leq e^{-1}. \tag{16}$$

The other parameters below are chosen to ease the application of the framework developed in the previous section.

(1) **Approaching the fixed point**. We define T_{start} to be the first time such that

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

15

where $\alpha := m + r$ and $w = \min\left(\frac{1}{6}, \frac{\log(1/\rho)}{6\log(L+1)}\right)$. We show below that

$$\mathbb{P}\left[T_{\text{start}} > t_{\text{start}} \log N\right] \leq 4mkt_o \log N \exp\left(-N^{1/3}\right), \tag{17}$$

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

(2) **Coming within distance $\Theta\left(\frac{\log N}{N}\right)$.** Let $\beta_0 := (8/\rho^k)\log(17mk)$ and $h = \frac{4L^2 m}{\delta^2}$. Then, we define T_0 to be the smallest number of steps needed after T_{start} such that

either $\left\|X^{(T_{\text{start}}+T_0)} - Y^{(T_{\text{start}}+T_0)}\right\|_1 \leq \frac{h\beta_0 \log N}{N}$ or $\left\|f(X^{(T_{\text{start}}+T_0)}) - f(Y^{(T_{\text{start}}+T_0)})\right\|_1 \leq \frac{h\beta_0 \log N}{N(1+\delta)}$.

We prove below that when N is large enough

$$\mathbb{P}\left[T_0 > kt_0 \log N\right] \leq \frac{1}{N^{\beta_0/7}}, \tag{18}$$

where $t_0 := \frac{1}{k\log(1/\rho)}$.

(3) **Coming within distance $\Theta(1/N)$.** Let β_0 and h be as defined in the last item. We now define a sequence of $\ell_1 := \left\lceil \frac{\log\log n}{k\log(1/\rho)} \right\rceil$ random variables $T_1, T_2, \ldots T_\ell$. We begin by defining the stopping time $S_0 := T_{\text{start}} + T_0$. For $i \geq 1$, T_i is defined to be the smallest number of steps after S_{i-1} such that the corresponding stopping time $S_i := S_{i-1} + T_i$ satisfies

either $\left\|X^{(S_i)} - Y^{(S_i)}\right\|_1 \leq \frac{h}{N}$ or $\left\|f(X^{(S_i)}) - f(Y^{(S_i)})\right\|_1 \leq \frac{h}{N}$.

Note that T_i is defined to be 0 if setting $S_i = S_{i-1}$ already satisfies the above conditions. Define $\beta_i = \rho^{ik}\beta_0$. We prove below that when N is large enough

$$\mathbb{P}\left[T_i > k+1\right] \leq 4mk \exp\left(-(\beta_i \log N)/8\right), \text{ for } 1 \leq i \leq \ell_1. \tag{19}$$

(4) **Collision.** Let β_0 and h be as defined in the last two items. Note that after time S_{ℓ_1}, we have

$$\left\|f(X^{(S_{\ell_1})}) - f(Y^{(S_{\ell_1})})\right\|_1 \leq \frac{Lh\beta_{\ell_1} \log N}{N} = \frac{Lh\beta_0 \rho^{k\ell_1} \log N}{N} \leq \frac{Lh\beta_0}{N}.$$

Then, from the properties of the optimal coupling we have that $X^{(S_{\ell_1}+1)} = Y^{(S_{\ell_1}+1)}$ with probability at least $\left(1 - \frac{Lh\beta_0}{2N}\right)^N$ which is at least $\exp\left(-L\beta_0 h\right)$ when N is so large that $N > hL$.

Assuming eqs. (17) to (19), we can complete the proof of Theorem 3.6 as follows.

Proof of Theorem 3.6. Let $X^{(0)}, Y^{(0)}$ be the arbitrary starting states of two stochastic evolutions guided by f, whose evolution is coupled using the optimal coupling. Let T be the minimum time t satisfying $X^{(t)} = Y^{(t)}$. By the Markovian property and the probability bounds in items 1 to 4 above, we have (for large enough N)

$$\mathbb{P}\left[T \leq t_{\text{start}} \log N + kt_0 \log N + (k+1)\ell_1\right] \geq \mathbb{P}\left[T_{\text{start}} \leq t_{\text{start}} \log N\right] \cdot \mathbb{P}\left[T_0 \leq kt_0 \log N\right]$$

$$\cdot \left(\prod_{i=1}^{\ell_1} \mathbb{P}\left[T_i \leq k+1\right]\right) \cdot e^{-L\beta_0 h}$$

$$\geq e^{-L\beta_0 h}\left(1 - \exp(-N^{1/4})\right)\left(1 - N^{-\beta/7}\right)$$

$$\cdot \prod_{i=1}^{\ell_1}\left(1 - 4mk\exp(-(\rho^{ik}\beta_0 \log N)/8)\right)$$

$$\geq \exp\left(-L\beta_0 h - 1\right)\left(1 - 4mk\sum_{i=1}^{\ell_1}\exp(-(\rho^{ik}\beta_0 \log N)/8)\right)$$

16

where the last inequality is true for large enough N. Applying Lemma C.1 to the above sum (with the parameters x and α in the lemma defined as $x = \exp(-(\beta_0 \log N)/8)$ and $\alpha = \rho^k \le 1/e$ by the assumption in eq. (16)), we can put a upper bound on it as follows:

$$\sum_{i=1}^{\ell_1} \exp(-(\rho^{ik}\beta_0 \log N)/8) \le \frac{1}{\exp(\rho^k \beta_0/8) - 1}$$
$$= \frac{1}{17mk - 1} \le \frac{1}{16mk},$$

where the first inequality follows from the lemma and the fact that $\frac{\log\log N}{k\log(1/\rho)} \le \ell_1 \le 1 + \frac{\log\log N}{k\log(1/\rho)}$, and the last inequality uses the definition of β_0 and $m, k \ge 1$. Thus, for large enough N, we have

$$\mathbb{P}\left[T \le c\log N\right] \ge q,$$

where $c := 2(t_{\text{start}} + kt_0)$ and $q := (3/4)\exp(-L\beta_0 h - 1)$. Since this estimate does not depend upon the starting states, we can bootstrap the estimate after every $c\log N$ steps to get

$$\mathbb{P}\left[T > c\ell \log N\right] < (1-q)^\ell \le e^{-q\ell},$$

which shows that

$$\frac{c}{q} \cdot \log N$$

is the mixing time of the chain for total variation distance $1/e$, when N is large enough. $\qquad\square$

We now proceed to prove the claimed equations, starting with eq. (17). Let $t := t_{\text{start}} \log N$ for convenience of notation. From Lemma A.1 we have

$$\left\| f^t(\boldsymbol{X}^{(0)}) - \boldsymbol{\tau} \right\|_1 \le r\rho^t.$$

On the other hand, applying Lemma 4.3 to the chain $\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \ldots$ with $\gamma = 1/3$, we have

$$\left\| f^t(\boldsymbol{X}^{(0)}) - \boldsymbol{X}^{(t)} \right\|_1 \le \frac{(L+1)^t m}{N^{1/3}} \text{ with probability at least } 1 - 2mt\exp(-N^{1/3}).$$

From the triangle inequality and the definition of t, we then see that with probability at least $1 - 2mt\exp\left(-N^{1/3}\right)$, we have

$$\left\| \boldsymbol{X}^{(t)} - \boldsymbol{\tau} \right\|_1 \le \frac{m}{N^{1/6}} + \frac{r}{N^{t_{\text{start}}\log(1/\rho)}} \le \frac{m+r}{N^w} = \frac{\alpha}{N^w},$$

where α, w are as defined in item 1 above. Now, if we instead looked at the chain starting at some $i < k$, the same result would hold for $\boldsymbol{X}^{(t+i)}$. Further, the same analysis applies also to $\boldsymbol{Y}^{(t+i)}$. Taking an union bound over these $2k$ events, we get the required result.

Before proceeding with the proof of the other two equations, we record an important consequences of eq. (17). Let w, α be as defined above, and let $w' > w$ be such that $w' < 1/3$. Recall that an epoch starting at time 0 is decent if both $\boldsymbol{X}^{(t)}$ and $\boldsymbol{Y}^{(t)}$ are within distance α/N^w of τ.

Observation 5.1. *For large enough N, it holds with probability at least $1 - \exp(-N^{w'/4})$ that for $1 \le i \le kN$, $\boldsymbol{X}^{(T_{\text{start}}+i)}, \boldsymbol{Y}^{(T_{\text{start}}+i)}$ are within ℓ_1 distance α/N^w of $\boldsymbol{\tau}$.*

Proof. We know from item 1 that the epochs starting at times $T_{\text{start}} + i$ for $0 \le i < k$ are all decent. For large enough N, Lemma 4.2 followed by a union bound implies that the N consecutive epochs starting at $T + j + k\ell$ where $\ell \le N$ and $0 \le j \le N$ are also all decent with probability at least $1 - 2mk^2 N\exp(-N^{w'/2})$, which upper bounds the claimed probability for large enough N. $\qquad\square$

We denote by E the event that the epochs starting at $T_{\text{start}} + i$ for $1 \le i \le kN$ are all decent. The above observation says that $P(E) \ge 1 - \exp(-N^{-w'/4})$ for N large enough.

We now consider T_0. Let $g_0 := \frac{h\beta_0 \log N}{N(1+\delta)}$, where h is as defined in items 2 and 3 above. From Theorem 4.4 followed by an union bound, we see that the first $t_1 \log N$ consecutive epochs starting at $T_{\text{start}}, T_{\text{start}} + k, T_{\text{start}} + 2k, \ldots$ are good with goal g_0 (they are already known to be decent with probability at least $P(E)$ from the above observation) with probability at least

$$1 - 2mkt_1 \left(N^{-\beta_0/(3(1+\delta))} + \exp(-N^{w'}/4) \right) \log N - P(\neg E),$$

which is larger than $1 - N^{-\beta_0/7}$ for N large enough (since $\delta < 1$). Now, if we have $\left\| f(\boldsymbol{X}^{(i)}) - f(\boldsymbol{Y}^{(i)}) \right\|_1 \le g$ for some time i during these $t_1 \log N$ good epochs then $T_0 \le k t_1 \log N$ follows immediately. Otherwise, the goodness condition implies that the hypotheses of Theorem 4.1 are satisfied across all these epochs, and we get

$$\left\| \boldsymbol{X}^{(T_{\text{start}} + kt_1 \log N)} - \boldsymbol{Y}^{(T_{\text{start}} + kt_1 \log N)} \right\|_1 \le \rho^{kt_1 \log N} \left\| \boldsymbol{X}^{(T_{\text{start}})} - \boldsymbol{Y}^{(T_{\text{start}})} \right\|_1 \le \rho^{kt_1 \log N} \frac{\alpha}{N^w} \le g_0 \le \frac{h\beta_0 \log N}{N},$$

where the second last inequality is true for large enough N.

Finally, we analyze T_i for $i \ge 1$. For this, we need to consider cases according to the state of the chain at time S_{i-1}. However, we first observe that plugging our choice of h into Theorem 4.4 shows that any decent epoch is good with goal $g_i := \frac{h\beta_i \log N}{N(1+\delta)}$ with probability at least

$$1 - 2mk \left(\exp(-(\beta_i \log N)/(3(1+\delta))) + \exp(-N^{w'}/4) \right),$$

which is at least $1 - 2mk \exp(-(\beta_i \log N)/7)$ for N large enough (since $\delta < 1$). Further, since we can assume via the above observation that all the epochs we consider are decent with probability at least $P(E)$, it follows that the epoch starting at S_{i-1} (and also the one starting at $S_{i-1} + 1$) is good with goal g_i with probability at least

$$p := 1 - 2mk \exp(-(\beta_i \log N)/7) - P(\neg E) \ge 1 - 2mk \exp(-(\beta_i \log N)/8),$$

where the last inequality holds whenever $\beta_i \le \log N$ and N is large enough (we will use at most one of these two epochs in each of the exhaustive cases we consider below). Note that if at any time $S_{i-1} + j$ (where $j \le k+1$) during one of these two good epochs it happens that $\left\| f(\boldsymbol{X}^{(S_{i-1}+j)}) - f(\boldsymbol{Y}^{(S_{i-1}+j)}) \right\|_1 \le g_i$, then we immediately get $T_i \le k+1$ as required. We can therefore assume that this does not happen, so that the hypotheses of Theorem 4.1 are satisfied across these epochs.

Now, the first case to consider is $\left\| \boldsymbol{X}^{(S_{i-1})} - \boldsymbol{Y}^{(S_{i-1})} \right\|_1 \le \frac{h\beta_{i-1} \log N}{N}$. Since we are assuming that Theorem 4.1 is satisfied across the epoch starting at S_{i-1}, we get

$$\left\| \boldsymbol{X}^{(S+i-1+k)} - \boldsymbol{Y}^{(S_{i-1}+k)} \right\|_1 \le \rho^k \left\| \boldsymbol{X}^{(S_{i-1})} - \boldsymbol{Y}^{(S_{i-1})} \right\|_1 \le \rho^k \frac{h\beta_{i-1} \log N}{N} = \frac{h\beta_i \log N}{N}. \tag{20}$$

Thus, in this case, we have $T_i \le k$ with probability at least p as defined in the last paragraph.

Even simpler is the case $\left\| f(\boldsymbol{X}^{(S_{i-1})}) - f(\boldsymbol{Y}^{(S_{i-1})}) \right\|_1 \le \frac{h\beta_i \log N}{N}$ in which case T_i is zero by definition. Thus the only remaining case left to consider is

$$\frac{h\beta_i \log N}{N} < \left\| f(\boldsymbol{X}^{(S_{i-1})}) - f(\boldsymbol{Y}^{(S_{i-1})}) \right\|_1 \le \frac{h\beta_{i-1} \log N}{N(1+\delta)}.$$

Since $h = \frac{4L^2 m}{\delta^2}$, the first inequality allows us to use Lemma 3.5 with the parameters c and t in that lemma set to $c = 4/\delta^2$ and $t = \beta_i L^2 \log N$, and we obtain

$$\left\| \boldsymbol{X}^{(S_{i-1}+1)} - \boldsymbol{Y}^{(S_{i-1}+1)} \right\|_1 \le (1+\delta) \left\| f(\boldsymbol{X}^{(S_{i-1})}) - f(\boldsymbol{Y}^{(S_{i-1})}) \right\|_1 \le \frac{h\beta_{i-1} \log N}{N},$$

18

with probability at least $1 - 2m \exp\left(-(\beta_i L^2 \log N)/3\right)$. Using the same analysis as the first case from this point onward (the only difference being that we need to use the epoch starting at $S_{i-1} + 1$ instead of the epoch starting at S_{i-1} used in that case), we get that

$$\mathbb{P}\left[T_i \leq 1 + kt\right] \geq p - 2m \exp\left(-(\beta_i L^2 \log N)/3\right)$$
$$\geq 1 - 4mk \exp\left(-(\beta_i \log N)/8\right).$$

since $L, k > 1$. Together with eq. (20), this completes the proof of eq. (19).

REFERENCES

[AF98]   D. Alves and J. F. Fontanari. Error threshold in finite populations. *Phys. Rev. E*, 57(6):7008 – 7013, June 1998.
[BSSD11] Rajesh Balagam, Vasantika Singh, Aparna Raju Sagi, and Narendra M. Dixit. Taking multiple infections of cells and recombination into account leads to small within-host effective-population-size estimates of HIV-1. *PLoS ONE*, 6(1):e14531, 01 2011.
[CF98]   P. R. A. Campos and J. F. Fontanari. Finite-size scaling of the quasispecies model. *Phys. Rev. E*, 58:2664–2667, 1998.
[DP09]   Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.
[DSV12]   Narendra Dixit, Piyush Srivastava, and Nisheeth K. Vishnoi. A finite population model of molecular evolution: Theory and computation. *J. Comput. Biol.*, 19(10):1176–1202, 2012.
[Eig71]   M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58:456–523, 1971.
[ES77]   M. Eigen and P. Schuster. The hypercycle, a principle of natural self-organization. Part A: Emergence of the hypercycle. *Die Naturwissenschaften*, 64:541–565, 1977.
[KAB06]   Roger D. Kouyos, Christian L. Althaus, and Sebastian Bonhoeffer. Stochastic or deterministic: What is the effective population size of HIV-1? *Trends Microbiol.*, 14(12):507 – 511, 2006.
[Mus11]   Fabio Musso. A stochastic version of Eigen's model. *Bull. Math. Biol.*, 73:151 – 180, 2011.
[MV05]   Elchanan Mossel and Eric Vigoda. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–9, Sep 2005.
[Now06]   M.A. Nowak. *Evolutionary Dynamics*. Harvard University Press, 2006.
[NS89]   M. Nowak and P. Schuster. Error thresholds of replication in finite populations-mutation frequencies and the onset of muller's ratchet. *J. Theor. Biol.*, 137:375–395, 1989.
[RPM08]   Tane S. Ray, Karl A. Payne, and L. Leo Moseley. Role of finite populations in determining evolutionary dynamics. *Phys. Rev. E*, 77(2):021909, Feb 2008.
[SRA08]   David B. Saakian, Olga Rozanova, and Andrei Akmetzhanov. Dynamics of the Eigen and the Crow-Kimura models for molecular evolution. *Phys. Rev. E*, 78(4):041908, Oct 2008.
[SV07]   Daniel Stefankovic and Eric Vigoda. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Sys. Bio.*, 56(1):113–24, Jan 2007.
[TBVD12] 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, 2012.
[Vis]   Nisheeth K. Vishnoi. Evolution without sex, drugs and Boolean functions. *http://theory.epfl.ch/vishnoi/*.
[Vis15]   Nisheeth K. Vishnoi. The speed of evolution. In *Proc. 26th Annual ACM-SIAM Symp. Discret. Algorithms (SODA)*, pages 1590–1601, 2015.
[vNCM99] Eric van Nimwegen, Japen P. Crutchfield, and Melanie Mitchell. Statistical dynamics of the Royal Road Genetic Algorithm. *Theor. Comput. Sci.*, 229:41 – 102, 1999.
[Wil05]   Claus Wilke. Quasispecies theory in the context of population genetics. *BMC Evol. Biol.*, 5(1):44, 2005.

APPENDIX A. PROOFS OMITTED FROM SECTION 3.1

Lemma A.1 (**Exponential convergence**). *Let f be a smooth contractive evolution, and let $\boldsymbol{\tau}$ and ρ be as in the conditions described in Section 3.1. Then, there exist a positive r such that for every $\boldsymbol{z} \in \Delta_m$, and every positive integer t,*

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

Proof. Let ε and k be as defined in Theorem 4.1. From the "convergence to the fixed point" condition, we know that there exists an ℓ such that for all $\boldsymbol{z} \in \Delta_m$,

$$\left\|f^\ell(\boldsymbol{z}) - \boldsymbol{\tau}\right\|_1 \leq \frac{\varepsilon}{L^k}. \tag{21}$$

Note that this implies that $f^{\ell+i}(z)$ is within distance ε of τ for $i = 0, 1, \ldots, k$, so that Theorem 4.1 can be applied to the sequence of vectors $f^\ell(z), f^{\ell+1}(z), \ldots, f^{\ell+k}(z)$ and $\tau, f(\tau) = \tau, \ldots, f^k(\tau) = \tau$ (the perturbations are simply 0). Thus, we get

$$\left\| f^{\ell+k}(z) - \tau \right\|_1 \leq \rho^k \left\| f^\ell(z) - \tau \right\|_1 \leq \frac{\rho^k \varepsilon}{L^k}.$$

Since $\rho < 1$, we see that the epoch starting at $\ell + k$ also satisfies eq. (21) and hence we can iterate this process. Using also the fact that the $1 \to 1$ norm of the Jacobian of f is at most L (which we can assume without loss of generality to be at least 1), we therefore get for every $z \in \Delta_m$, and every $i \geq 0$ and $0 \leq j < k$

$$\left\| f^{\ell+ik+j}(z) - \tau \right\|_1 \leq \rho^{ki+j} \frac{L^j}{\rho^j} \left\| f^\ell(z) - \tau \right\|_1$$

$$\leq \rho^{ki+j+\ell} \frac{L^{j+\ell}}{\rho^{j+\ell}} \left\| z - \tau \right\|_1 \leq \rho^{ki+j+\ell} \frac{L^{k+\ell}}{\rho^{k+\ell}} \left\| z - \tau \right\|_1$$

where in the last line we use the facts that $L > 1$, $\rho < 1$ 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 $z \in \Delta_m$

$$\left\| f^t(z) - \tau \right\|_1 \leq \left(\frac{L}{\rho} \right)^{k+\ell} \rho^t \left\| z - \tau \right\|_1. \tag{22}$$

Similarly, for $t < \ell$, we have, for any $z \in \Delta_m$

$$\left\| f^t(z) - \tau \right\|_1 \leq L^t \left\| z - \tau \right\|_1$$

$$\leq \left(\frac{L}{\rho} \right)^t \rho^t \left\| z - \tau \right\|_1 \leq \left(\frac{L}{\rho} \right)^\ell \rho^t \left\| z - \tau \right\|_1, \tag{23}$$

where in the last line we have again used $L > 1, \rho < 1$ and $t < \ell$. From eqs. (22) and (23), we get the claimed result with $r := \left(\frac{L}{\rho} \right)^{k+\ell}$. $\qquad \square$

APPENDIX B. PROOFS OMITTED FROM SECTION 4

Proof of Lemma 4.3. Fix a co-ordinate $j \in [m]$. Since $\boldsymbol{X}^{(i)}$ is the normalized frequency vector obtained by taking N independent samples from the distribution $f(\boldsymbol{X}^{(i-1)})$, Hoeffding's inequality yields that

$$\mathbb{P}\left[\left| X_j^{(i)} - f(\boldsymbol{X}^{(i-1)})_j \right| > N^{-\gamma} \right] \leq 2\exp\left(-N^{1-2\gamma}\right) \leq 2\exp\left(-N^\gamma\right),$$

where the last inequality holds because $\gamma \leq 1/3$. Taking a union bound over all $j \in [m]$, we therefore have that for any fixed $i \leq t$,

$$\mathbb{P}\left[\left\| \boldsymbol{X}^{(i)} - f(\boldsymbol{X}^{(i-1)}) \right\|_1 > \frac{m}{N^\gamma} \right] \leq 2m\exp\left(-N^\gamma\right). \tag{24}$$

For ease of notation let us define the quantities $s^{(i)} := \left\| \boldsymbol{X}^{(i)} - f^i(\boldsymbol{X}^{(0)}) \right\|_1$ for $0 \leq i \leq t$. Our goal then is to show that it holds with high probability that $s^{(i)} \leq \frac{(L+1)^i m}{N^\gamma}$ for all i such that $0 \leq i \leq t$.

Now, by taking an union bound over all values of i in eq. (24), we see that the following holds for all i with probability at least $1 - 2mt\exp\left(-N^\gamma\right)$:

$$s^{(i)} = \left\| \boldsymbol{X}^{(i)} - f^i(\boldsymbol{X}^{(0)}) \right\|_1 \leq \left\| \boldsymbol{X}^{(i)} - f(\boldsymbol{X}^{(i-1)}) \right\|_1 + \left\| f(\boldsymbol{X}^{(i-1)}) - f^i(\boldsymbol{X}^{(0)}) \right\|_1 \leq \frac{m}{N^\gamma} + Ls^{(i-1)}, \tag{25}$$

where the first term is estimated using the probabilistic guarantee from eq. (24) and the second using the upper bound on the $1 \to 1$ norm of the Jacobian of f. However, eq. (25) implies that $s^{(i)} \leq \frac{m(L+1)^i}{N^\gamma}$ for all $0 \leq i \leq t$, which is what we wanted to prove. To see the former claim, we proceed by induction. Since $s_0 = 0$, the claim is trivially true in the base case. Assuming the claim is true for $s^{(i)}$, we then apply eq. (25) to get

$$s^{(i+1)} \leq \frac{m}{N^\gamma} + Ls^{(i)} \leq \frac{m}{N^\gamma}\left(1 + L(L+1)^i\right) \leq \frac{m}{N^\gamma} \cdot (L+1)^{i+1}. \qquad \square$$

The following technical lemma is used in the proof of Theorem 3.6.

Lemma C.1. *Let x, α be positive real numbers less than 1 such that $\alpha < \frac{1}{e}$. Let ℓ be a positive integer, and define $y := x^{\alpha^\ell}$. Then*

$$\sum_{i=0}^{\ell} x^{\alpha^i} \leq \frac{y}{1-y}.$$

Proof. Note that since both x and α are positive and less than 1, so is y. We now have

$$\sum_{i=0}^{\ell} x^{\alpha^i} = \sum_{i=0}^{\ell} x^{\alpha^{\ell-i}} = y \sum_{i=0}^{\ell} y^{\alpha^{-i}-1}$$

$$\leq y \sum_{i=0}^{\ell} y^{i \log(1/\alpha)}, \text{ since } 0 < y \leq 1 \text{ and } \alpha^{-i} \geq 1 + i \log(1/\alpha),$$

$$\leq y \sum_{i=0}^{\ell} y^i, \text{ since } 0 < y \leq 1 \text{ and } \alpha < 1/e,$$

$$\leq \frac{y}{1-y}. \qquad \square$$