Escaping Saddle Points: from Agent-based Models to Stochastic Gradient Descent

Grant Schoenebeck[*] Fang-Yi Yu[†]

Abstract

We study a large family of stochastic processes captured by the fact that they update a limited amount in each step, e.g. agent-based models where one agent at a time updates their state or stochastic gradient descent where the step size is not too large. A key question is how this family of stochastic processes are approximated by their mean-field approximations. Prior work shows that the stochastic processed escapes repelling fixed points and saddle points in polynomial time.

We provide a tight analysis: for any non-attracting fixed point in any stochastic process in this family, we show that for a sufficiently small constant $\epsilon > 0$, the process will be ϵ-far away from the fixed point in $O(n \log n)$ time with high probability. We also show that it takes time $\Omega(n \log n)$ to escape such a fixed point with constant probability. This shows our result is optimal up to a multiplicative constant.

We leverage the above result to show that with high probability these stochastic processes are arbitrarily close to an attracting fixed point in $O(n \log n)$ time.

We show the power of our results by applying them to several settings: evolutionary game theory, opinion formation dynamics, and stochastic gradient descent.

[*] University of Michigan, schoeneb@umich.edu.

[†] University of Michigan, fayu@umich.edu.

1 Introduction

Many agent-based models are discrete time stochastic processes (SP) in Euclidean spaces, e.g., interacting particle systems, social learning, and evolutionary game theory. Of particular interest is the temporal behavior of these process: does such a process converge? What is the limit of the process? How fast does the process reach its limit? An example of such a process would be n agents, each faced with a binary set of choices, that evolves in rounds where in each round one agent updates its choice. These processes can be written:

$$\boldsymbol{X}(k+1) = \boldsymbol{X}(k) + \frac{1}{n}\left(f(\boldsymbol{X}(k)) + \boldsymbol{U}(k+1)\right) \tag{SP}$$

where $\boldsymbol{X}(k)$ represents the state at time k, n the number of agents, f the *drift* of the process, and $\boldsymbol{U}$ is the unbiased *noise*.[1]

However, such stochastic processes are hard to analyze in general. A popular simplification of such a model is hypothesizing an infinitely large population of interacting agents, $n \to \infty$, and approximating them as a deterministic continuous process modelled as a time homogeneous ordinary differential equation(ODE):

$$\frac{d}{dt}\boldsymbol{x} = f(\boldsymbol{x}). \tag{ODE}$$

Note that the function f is the same in Equations (ODE) and (SP). This approach, often called the *mean-field approximation*, encodes the process at each time as a point in Euclidean space (e.g., the fraction of agents with the blue opinion), and models the expected change as the average over a large population.

An important question is how the solutions to the mean-field approximation (ODE) relate to the stochastic process (SP), and, in particular, how fast the stochastic process (SP) can diverge from the solution to the differential equation (ODE). Such knowledge requires understanding how the two processes relate around the fixed points of the differential equation due to the following intuition: Starting at a non-attracting fixed point of the ODE, the solution to the ODE will stay at the point forever. If the deterministic process starts near but not at the same fixed point, the process can escape. However, the stochastic process (SP), due to its noisy behavior, should behave identically whether it starts exactly on a fixed point, or sufficiently close to it.

Several prior works also study this difference between a particular stochastic process (SP) and its mean-field approximation (ODE), they also show that the stochastic process escapes repelling or saddle point, but only in polynomial time. [39, 31]

Our results In contrast, we provide a tight analysis of the time for a general family of stochastic processes (SP) to escape non-attracting fixed points. Informally, given a stochastic process in the family starting at a non-attracting fixed point, there exists a constant $\epsilon > 0$ independent of the number of agents n such that a stochastic process will be ϵ-far away from the fixed point in time $O(n \log n)$ with high probability. On the other hand, we further show it takes $\Omega(n \log n)$ time to escape fixed points with a constant probability.

We leverage the above result to show that with high probability these stochastic processes are arbitrarily close to an attracting fixed point in $O(n \log n)$ time. That is, they not only locally escape fixed points quickly, but globally converge quickly.

With we apply these results on the behavior of this family of stochastic processes to three different situations: evolutionary game theory, opinion formation, and stochastic gradient descent.

[1] See preliminaries for a more precise formalism.

Evolutionary game theory: As a warm-up example, we consider the *Logit dynamics* [48, 10, 27], which is a well-studied smooth best response dynamic in evolutionary game theory. We study the symmetric coordination game with binary actions (blue and red) where half of the agents playing blue and the other half playing red is a Nash equilibrium but not a "stable" equilibrium. As a corollary of our main result, we show the Logit dynamics escape the equilibrium in $O(n \log n)$ time with high probability.

Opinion formation dynamics: We study how two mutually exclusive competing opinions evolve in networks with community structure. We model networks using a planted community model which has a long history in the sociology literature [63]. We consider a large family of dynamics called majority-like ***Node dynamics*** [61, 62]. In Node dynamics, in each round a random node updates its opinion based on the fraction of red and blue neighbors it has. The majority-like node dynamics intuitively can be thought of as in between the voter model and the iterative majority dynamics:

We prove a dichotomy theorem: for any pair of "majority-like" node dynamics and planted community model we show that either: the system quickly converges to consensus with high probability in time $\Theta(n \log(n))$; or, the system can get "stuck" and take time $2^{\Theta(n)}$ to reach consensus. We note that $O(n \log(n))$ is optimal because it takes this long for each node to even update its opinion.

Stochastic gradient descent: Our escaping non-attracting point result also adds to the recent literature on saddle-point analysis of stochastic gradient descent on a non-convex objective function. In particular, our result shows a general family of stochastic gradient descent, *bounded stochastic gradient descent* with a constant step size $1/n$, converges to a local minimal in $\Theta(n \log n)$ when n is sufficiently large, the noise is well-behaved, and the objective function has a continuous third derivative. The bounded stochastic gradient descent contains models of Ge et al. [31], Jin et al. [39] as a special case, and our analysis provides a tight convergence time to a local minima. Our analysis only applies to the case where the dimension is bounded.

2 Related Work

Mean-field and stochastic processes: There is a long line of work considering the relationship between (SP) and (ODE) which consider (SP) and (SP) short-term behavior $k = O(n)$ or limit-behavior $k \to \infty$ with fixed n [70, 7, 56]. Started by Robbins and Monro [58], another related area is called (constant step size) stochastic approximation algorithms [6, 11, 45] which use local search to find zeroes of an objective function. Several works study if the limit of (SP) with fixed n converges to a non-attracting fixed point [7, 55].

One important special cases is stochastic gradient descent on non-convex objective functions [39, 31]. These works focus on the convergence time to a local minimal of the objective function (attracting fixed point of its gradient). Informally, they consider long-term behavior of (SP) where k is a polynomial of n, and provide non-asymptotic analysis of convergence time with respect to the properties of the objective function (Lipschitz, smoothness, and dimension).

Instead of discrete time stochastic process, the convergence of stochastic differential equations are studied [45, 49] which do not have a clear analogue to long-term behavior. In particular, Mertikopoulos and Staudigl [49] also focus on Gradient-like systems.

In the literature of Markov chains, a large volume of work is devoted to bounding the hitting time of different Markov process and achieving fast convergence. The techniques typically employed are (1) showing the Markov chain has fast mixing time [50, 53], (2) reducing the dimension of the process into small set of parameters (e.g., the frequency of each opinion) and using a mean field approximation and concentration property to control the behavior of the process [4], or (3) using handcrafted potential functions [52]. Our results extend the second approach. However, the mean-field of our dynamics has unstable fixed points and does necessarily not have a nice potential function. We circumvent these challenges by exploiting the literature of dynamical systems [59, 14] and showing the existence of a potential function by analyzing the phase portrait of the flow. Additionally, we show the process leaves unstable fixed points by using the stochastic nature of our process.

Evolutionary Game theory Stability of equilibrium is one central topic in evolutionary game theory [60, 54]. One approach is showing that the limit-behavior (also called infinite-horizon) of stochastic processes does not concentrate around such fixed points. One popular notion of stability is evolutionarily stable strategy [33], which says a strategy is stable if any small enough deterministic deviation cannot push the population away from such an equilibrium [7]. Our result provides an example showing that the smooth best response dynamics can escape an equilibrium which is not an evolutionarily stable strategy.

Opinion formation Our node dynamics model extends several previously studied dynamics including the voter model, iterative majority, iterative k-majority. The voter model has been extensively studied in mathematics [17, 36, 46, 47], physics [5, 13], and even in social networks [12, 64, 66, 67, 16]. A major theme of this work is how long it takes the dynamics to reach consensus on different network topologies. Work about iterative majority dynamics [44, 9, 41, 52, 68, 71] often study when the dynamics converge and how long it takes them to do so. Doerr et al. [22] prove 3-majority reaches "stabilizing almost" consensus on the complete graph in the presence of $O(\sqrt{n})$-dynamic adversaries. Many works extend this result beyond binary opinions [18, 15, 4, 1].

Another line of related literature is about designing and analyzing algorithms for consensus on social networks. When dealing with binary opinions, these works typically study more elaborate dynamics which, in particular, include nodes having memory beyond their opinion [42, 57, 8, 51]. Another line of work deals with agents selecting an opinion from among a large (or infinite) set of options [3, 29]. There are also myriad models where the opinions space is continuous instead of discrete. Typically agents either average their neighbors' opinions [20], or a subset of their neighbors' opinions which are sufficiently aligned [35, 19]. Finally, models involving the coevolution of the opinions and the network [37, 23, 30] have been studied using simulations and heuristic arguments.

Stochastic Gradient descent Recently, there is a long line of research of stochastic gradient descent on non-convex functions, see [31, 39] and the reference therein. Searching for the minimum value of a non-convex function is in general unfeasible, and those work focus on finding local minimal efficiently which is achieved by showing that stochastic gradient decent leaves non-minimal singular points (repelling and saddle fixed points) efficiently.

3 Preliminaries

Several models capture the behavior of a large population of agents in a *phase space*, $\mathcal{X}$—a compact space—and that update in accord to some function $f : \mathcal{X} \to \mathcal{X}$.[2] We will always use $\mathcal{X} = \mathbb{R}^d$, which, technically, must be compactified by adding infinity. We will say $f \in \mathcal{C}^r(\mathbb{R}^d, \mathbb{R}^d)$ if the r-th derivative of f is continuous.

[2] Here because we only consider the set $\mathcal{X}$ is $\mathbb{R}^d$, the image of f is equal to its domain. Otherwise, if $\mathcal{X}$ is a manifold, we need to consider f maps $\boldsymbol{x} \in \mathcal{X}$ to the tangent space, $T\mathcal{X}(\boldsymbol{x})$

A function $f : \mathcal{X} \to \mathcal{X}$ defines a ordinary differential equation on $\mathcal{X}$ via

$$\frac{d}{dt}\boldsymbol{x} = f(\boldsymbol{x}). \tag{1}$$

Equation (1) has a unique solution through every point $\boldsymbol{x} \in \mathcal{X}$ (when f satisfies some smoothness condition). Hence (1) defines a *flow* $\varphi : \mathcal{X} \times \mathbb{R} \to \mathcal{X}$ such that $\varphi(\boldsymbol{x}, 0) = \boldsymbol{x}$ and $\frac{d}{dt}\varphi(\boldsymbol{x}, t) = f(\varphi(\boldsymbol{x}, t))$ for all $t \in \mathbb{R}, \boldsymbol{x} \in \mathcal{X}$. We call φ the ***flow*** induced by f if $\varphi(\boldsymbol{x}, t)$ is the position of the solution of (1) at time t starting at $\boldsymbol{x}$.

On the other hand, we consider a discrete time stochastic process with values in $\mathcal{X}$, $(\boldsymbol{X}(k))_{k\in\mathbb{N}}$ where $\mathbb{N}$ is the set of non-negative integers. We will drop the dummy index k later and use $\boldsymbol{X}$ to simplify the notation. Given $n \in \mathbb{R}_{>0}$, we call the process $\boldsymbol{X}$ a ***$1/n$-step D-bounded stochastic process*** associated with f if the following assumptions are satisfied: first, for each $k \in \mathbb{N}$, define $\boldsymbol{U}(k+1)$ as $\boldsymbol{U}(k+1) = n\left(\boldsymbol{X}(k+1) - \boldsymbol{X}(k)\right) - f(\boldsymbol{X}(k))$, and let $\mathcal{F}(k)$ be the generated filtration on the process $\boldsymbol{X}$. Then it must be the case that, for all k, $\mathbb{E}[\boldsymbol{U}(k+1) \mid \mathcal{F}(k)] = \boldsymbol{0}$ and $\|\boldsymbol{U}(k)\| \leq D$ with probability 1. Alternatively, the stochastic process $\boldsymbol{X}$ admits the representation

$$\boldsymbol{X}(k+1) = \boldsymbol{X}(k) + \frac{1}{n}\left(f(\boldsymbol{X}(k)) + \boldsymbol{U}(k+1)\right) \tag{2}$$

where $\boldsymbol{U}$ is a zero-mean martingale i.e. the conditional expectation of $\boldsymbol{U}(k+1)$ given the history $\mathcal{F}(k)$ is equal to zero.

3.1 Dynamical systems

In this section, we introduce some basic notions to understand the behavior of dynamical systems (1), which are mostly from Robinson [59]. To understand the local behavior of flows, we define fixed points and the notion of hyperbolicity. Then to help us to study global and long term behavior of flows, we define potential functions (global Lyapunov functions). We restrict our analysis to gradient-like flows, which are the family of flows with potential functions. Gradient-like flows contain gradient flows as a special case.

3.1.1 Local behavior of flows— hyperbolic fixed points

Given a flow φ with f and a point $\boldsymbol{x} \in \mathcal{X}$, the *trajectory* or *orbit* of $\boldsymbol{x}$ is the set $\mathcal{O}_{\boldsymbol{x}} = \{\varphi(\boldsymbol{x}, t) : t \in \mathbb{R}\}$. A point $\boldsymbol{x} \in \mathcal{X}$ is a ***fixed point*** if $\mathcal{O}_{\boldsymbol{x}} = \{\boldsymbol{x}\}$ that is $f(\boldsymbol{x}) = 0$, and we use Fix_f to denote the set of fixed points. We call a set $E \subseteq \mathcal{X}$ *positive invariant* if for all $\boldsymbol{x} \in E$ and $t \geq 0$, $\varphi(\boldsymbol{x}, t) \in E$, *negative invariant* if it's true for all $t \leq 0$, and *invariant* if it's true for all $t \in \mathbb{R}$.

Here we introduce some important properties of linear flow in $\mathcal{X}$. Given a matrix $A \in \mathbb{R}^{d\times d}$, the linear equation

$$\frac{d}{dt}\boldsymbol{x}(t) = A\boldsymbol{x}(t)$$

has a closed form solution $\varphi(\boldsymbol{x}, t) = \exp(At)\boldsymbol{x}$, and $\boldsymbol{0}$ is a fixed point.

The long term behavior (e.g., convergence to $\boldsymbol{0}$, divergence to infinite, or rotation) of the above system depends on the real part of eigenvalues of A. Formally, we denote the set of eigenvalues for the (real) matrix A by

$$\mathrm{eig}(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_s, \lambda_{s+1}, \ldots, \lambda_{s+u}, \lambda_{s+u+1}, \ldots, \lambda_{s+u+c}\},$$

where $\Re(\lambda_i) < 0$ for all $1 \leq i \leq s$, $\Re(\lambda_{s+i}) > 0$ for all $1 \leq i \leq u$, and $\Re(\lambda_{s+u+i}) = 0$ for all $1 \leq i \leq c$. We define the *stable, unstable, and center subspaces* of A as follows:

$$\begin{aligned}
\mathbb{E}^s =& \{v : v \text{ is a generalized eigenvector for an eigenvalue } \lambda_i, \Re(\lambda_i) < 0\}; \\
\mathbb{E}^u =& \{v : v \text{ is a generalized eigenvector for an eigenvalue } \lambda_{s+i}, \Re(\lambda_{s+i}) > 0\}; \\
\mathbb{E}^c =& \{v : v \text{ is a generalized eigenvector for an eigenvalue } \lambda_{s+u+i}, \Re(\lambda_{s+u+i}) = 0\}.
\end{aligned}$$

Definition 3.1. We say $A \in \mathbb{R}^{d\times d}$ is *hyperbolic* if $\mathbb{E}^c = \emptyset$, i.e. for all $\lambda \in \operatorname{eig}(A)$

$$\Re(\lambda) \neq 0.$$

If A is hyperbolic, we further call A *attracting* (or *repelling*) if for all $\lambda \in \operatorname{eig}(A), \Re(\lambda) < 0$, (or $\Re(\lambda) > 0$) respectively. Finally, if A is not attracting nor repelling, we call it *saddle*.

We can extend the notion of hyperbolic to fixed points of nonlinear dynamics by its local properties:

Definition 3.2 (Attracting, repelling, and saddle points)**.** Given a flow φ induced by f and a fixed point $\boldsymbol{\beta}$ we call $\boldsymbol{\beta}$ *hyperbolic* if the matrix $\nabla f(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{\beta}}$ is hyperbolic (Definition 3.1). Similarly, $\boldsymbol{\beta}$ is respectively an *attracting, repelling or saddle fixed point* if $\nabla f(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{\beta}}$ is attracting, repelling or saddle. We further call a fixed point non-attracting if it's either repelling or saddle.

3.1.2 Global behavior of flows— Lyapunov functions and gradient-like flows

However, general nonlinear flows can have complicated behavior beyond fixed points (c.f. Strogatz [65]). Here we consider a family of flows which have a relative easy global behavior and also general enough to contain several interesting dynamics.

A flow with f is called a ***gradient-flow*** if there exists a real value function $F : \mathbb{R}^d \to \mathbb{R}$ such that $f = -\nabla F$. That is the flow is a solution to:

$$\frac{d}{dt}\boldsymbol{x} = -\nabla F(\boldsymbol{x}) \tag{3}$$

However there is a more general family of dynamics called ***gradient-like flows*** that contain gradient flows. Intuitively, a dynamics is a gradient-like flows if there exists a potential (Lyapunov) function for the induced flow.

Definition 3.3. A $\mathcal{C}^1$ function $V : \mathcal{X} \to \mathbb{R}$ is called a weak global Lyapunov function for a flow φ with f if

$$\frac{d}{dt}V(\varphi(\boldsymbol{x}, t))|_{t=0} \leq 0$$

We use Lie derivative to simply the notion, $\mathcal{L}_f V(\boldsymbol{x}) \triangleq \frac{d}{dt}V(\varphi(\boldsymbol{x}, t))|_{t=0}$.

A gradient-like flows is not necessary a gradient flow but it has a strict global Lyapunov function V which is decreasing off the fixed points. Formally,

Definition 3.4. A flow φ is called *gradient-like* if there is a strict global Lyapunov function $V \in \mathcal{C}^1$ which is strictly decreasing off the fixed points:

$$\mathcal{L}_f V(\boldsymbol{x}) < 0 \text{ for all } \boldsymbol{x} \notin \operatorname{Fix}_f.$$

We further call the function f gradient-like.

A gradient flow is also a gradient-like flow as the function $F : \mathbb{R}^d \to \mathbb{R}$ can serve as the Lyapunov function. In this paper we only consider gradient-like flows with a finite number of fixed points, and all the fixed points are hyperbolic.

3.2 Notations

Analysis Here we define some common notions to characterize the function $f : \mathbb{R}^d \to \mathbb{R}^d$. Given $r_{\text{reg}} \in \mathbb{R}_{>0}$, we denote by $B(\mathbf{0}, r_{\text{reg}}) = \{\boldsymbol{x} : \|\boldsymbol{x}\| < r_{\text{reg}}\}$ an open d-sphere center at $\mathbf{0}$ with radius r_{reg}, and denote by $\bar{B}$ the closed sphere. We use $\|\cdot\|$ to denote the 2-norm in $\mathbb{R}^d$ and $\|\cdot\|_\infty$ for the ∞-norm.

Given a compact set E (a bounded and closed subset $\mathbb{R}^d$), we say f is *B_f-bounded* in E if $\|f(\boldsymbol{x})\| \leq B_f$ for all $\boldsymbol{x} \in E$. We say f is *L_f-Lipschitz* if $\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \leq L_f \|\boldsymbol{x} - \boldsymbol{x}'\|$ for all $\boldsymbol{x}$ and $\boldsymbol{x}'$ in E. We say f is *S_f-smooth* in E if $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\| \leq S_f \|\boldsymbol{x} - \boldsymbol{x}'\|$ for all $\boldsymbol{x}$ and $\boldsymbol{x}' \in E$ where we use Frobenius norm on the left hand side. We further say f is bounded (Lipschitz or smooth) if there exists a constant $M \in \mathbb{R}_{>0}$ such that f is M-bounded (M-Lipschitz or M-smooth). Note that if $f \in \mathcal{C}^1$ and E is compact set, f is automatically bounded and Lipschitz.

Stochastic process Let $\boldsymbol{X}$ be a discrete time stochastic process. A *stopping time* with respect to $\boldsymbol{X}$ is a random time such that for each $k \in \mathbb{N}$, the event $\{T = k\}$ is completely determined by (at most) the total information known up to time k, $\mathcal{F}(k)$. Given a set $B \in \mathcal{X}$, we denote *(first) exit time* of B as $T_e(B) \triangleq \min\{k \geq 0 : \boldsymbol{X}(k) \notin B\}$. We also consider the *hitting time* of B as $T_h(B) \triangleq \min\{k \geq 0 : \boldsymbol{X}(k) \in B\}$.

Linear algebra Given a hyperbolic matrix A we define $\sigma_{\max}(A) = \max\{\|A\boldsymbol{v}\| : \|\boldsymbol{v}\| = 1\}$ to be the *induced norm* of A, and $\mu_{\max}(A) = \max_{\lambda \in \text{eig}(A)} |\Re(\lambda)|$. Therefore, for all $\boldsymbol{v} \in \mathbb{R}^d$, $\|A\boldsymbol{v}\| \leq \sigma_{\max}(A)\|\boldsymbol{v}\|$, and $\boldsymbol{v}^\top A \boldsymbol{v} \leq \mu_{\max}(A)\|\boldsymbol{v}\|^2$ (Theorem 4.3.50 in Horn et al. [38]). Furthermore, we set $\mu_s(A) = \min_{\lambda_i : i \leq s} |\Re(\lambda_i)|$ to denote the smallest absolute value of the real part of stable eigenvalues, and $\mu_u(A) = \min_{\lambda_{s+i} : i \leq u} |\Re(\lambda_{s+i})|$ to denote the smallest absolute value of the real part of unstable eigenvalues.

Given two linear subspaces E_1 and E_2 in $\mathbb{R}^d$, we can define the angle between E_1 and E_2

$$arc(E_1, E_2) = \min\left\{\arccos(|\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle|) : \boldsymbol{v}_1 \in E_1, \|\boldsymbol{v}_1\| = 1, \boldsymbol{v}_2 \in E_2, \|\boldsymbol{v}_2\| = 1\right\}. \tag{4}$$

If $E_1 \cap E_2 = \{\mathbf{0}\}$, the angle $arc(E_1, E_2)$ is greater than 0. As a result, if $E_1 \oplus E_2 = \mathbb{R}^d$, for all $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\|P_1 \boldsymbol{v}\| \leq \frac{1}{\sin arc(E_1, E_2)} \|\boldsymbol{v}\| \tag{5}$$

We use $\mathbb{I}_d$ to denote the d-dimensional identity matrix. For two symmetric matrices A and B we use $A \prec B$ to denote $B - A$ is a positive definite matrix.

4 Main results

The relationship between (1) and (2) has be studied extensively and the analysis can be classified as two types: short-term and long term behavior. We illustrate the intuitive difference between short-term and long-term behavior with a thought experiment: Suppose $\boldsymbol{U}(k)$ is zero for all k and $\boldsymbol{X}(0) = \boldsymbol{x}$. Then the $1/n$-step D-bounded stochastic process with f reduces to a deterministic sequence. By standard theory in numerical methods, for all $t \in \mathbb{R}$ and sufficiently large n, $\boldsymbol{X}(nt) \approx \varphi(\boldsymbol{x}, t)$. In many cases, this can be shown to be true if t is bounded, but $n \to \infty$. However, if we exchange the roles of t and n so that n is bounded, but $t \to \infty$ this relation tends not to hold anymore. The first case describes the regime of short-term behavior where $t = O(n)$, and the later cases describes the limit-behavior with $t \to \infty$ with a fixed n.

In this section, we provide a more refined characterization of (2) around fixed points when both $t \to \infty$ and $n \to \infty$. However, first show that in the short term regime, Equations (1) and (2) behave similarly. This provides a basis for our lower bound, that it takes time $\Omega(n \log n)$ to escape a non-attracting fixed point. We then show our main result, that it takes time $O(n \log n)$ to escape a non-attracting fixed point. Finally, we show that when dynamics are started close to an attracting fixed point, the dynamics quickly converge to the attracting fixed (with high probability), and take exponential time to escape.

4.1 Short term behavior

In this section, we show that if the time is bounded by Cn, then (1) and (2) remain within ϵ with probability roughly $\exp(-c\epsilon^2 n)$. Several variations of Theorem 4.1 are proven in the literature [70, 7]. Here we provide an explicit dependency of the constant c in the exponent in terms of the process (2). This will be important for the proof of our lower bound.

Theorem 4.1. *Let $E \subseteq \mathcal{X} = \mathbb{R}^d$ be a compact space which is positive invariant for a flow induced by f defined in* (1) *and a $1/n$-step D bounded stochastic process with f defined in* (2). *Suppose f is L_f-Lipschitz and B_f-bounded* ($\max_{x \in E} \|f(x)\| \leq B_f$)*. For all $C > 0$, there exists a constant $c = \frac{2\exp\left(-2L_f C\right)}{C(B_f + D)}$ such that for all $\epsilon > 0$ and n large enough:*

$$\Pr\left[\max_{k \leq Cn} \|\boldsymbol{X}(k) - \varphi(\boldsymbol{x}, k/n)\|_\infty > \epsilon \mid \boldsymbol{X}(0) = \boldsymbol{x}\right] \leq 2d \exp(-c\epsilon^2 n)$$

for all $x \in E$.

We give the proof in the appendix. This result can be seen as a stochastic version of the Euler forward method—the discrete-time stochastic process (2) is very close to the solution of the ordinary differential equation (1). Because the noise at each step in (2) is a bounded martingale, we can use concentration bounds to show the aggregated deviation from its expectation is small.

4.2 Escaping non-attracting points

In this section we want to study how (2) behaves near the non-attracting points of f. Specifically, if $\mathbf{0}$ is a saddle point of f, we know $\varphi(\mathbf{0}, t) = \mathbf{0}$ for all $t \in \mathbb{R}_{>0}$. For large enough n does (2) stay at $\mathbf{0}$ forever as well? If not when will it escape the saddle point $\mathbf{0}$?

Theorem 4.1 easily yields a lower bound on the time required to escape a saddle point (or general fixed point). Formally, given $\epsilon > 0$ and a neighborhood of a saddle point $B(\mathbf{0}, \epsilon)$ (a d-sphere centered at $\mathbf{0}$ with radius ϵ), the time to escape $B(\mathbf{0}, \epsilon)$ with constant probability is $\Omega(n \log n)$.

Corollary 4.2 (Lower bound for fixed points)**.** *Let $\mathbf{0}$ be a fixed point of* (1)*. If E is a compact set where $\mathbf{0}$ is in the interior of E and f is bounded and Lipschitz in E, for all $0 < \epsilon$ where $\bar{B}(\mathbf{0}, \epsilon) \subset E$, there exists $\tau = \Omega(n \log n)$, such that with high probability the exit time is greater than τ*

$$\Pr\left[T_e(B(\mathbf{0}, \epsilon)) \leq \tau \mid \boldsymbol{X}(0) = \mathbf{0}\right] = o(1).$$

Note that the asymptotic notion is taken with respect to n.[3]

[3] Note that in contrast to Theorem 4.1, here we only provide asymptotic analysis for n, so the parameter of f (Lipschitz and bound), diameter of E, and dimension of space d are hidden in the big-O notation.

Now we show an upper bound: a large family of $1/n$-step D-bounded stochastic processes with f can escape saddle points or unstable fixed points in $O(n \log n)$ steps with high probability as long as the second moment of the noise is large.

Theorem 4.3 (Escaping non-attracting fixed points)**.** *Let f be defined on $\mathcal{X} = \mathbb{R}^d$. Let $\mathbf{0}$ be a hyperbolic saddle point (or repelling point) of* (1)*. Suppose there is a compact set E such that 1) $\mathbf{0}$ is in the interior of E 2) $f \in \mathcal{C}^1$ and smooth in E, 3) There exists $\alpha > 0$ such that*

$$\forall \boldsymbol{x} \in E, \frac{\alpha}{d}\mathbb{I}_d \prec \operatorname{Cov}[\boldsymbol{U}(k+1) \mid \boldsymbol{X}(k) = \boldsymbol{x}].^4$$

Then, there exist $\tau = O(n \log n)$ and positive constant $r_{\text{exit}} > 0$ such that $B(\mathbf{0}, r_{\text{exit}}) \subset E$, and

$$\Pr\left[T_e(B(\mathbf{0}, r_{\text{exit}})) \leq \tau\right] = 1 - o(1).$$

The main idea of the proof is to use the linear approximation of f at $\mathbf{0}$, and then use induction to show the process will have a large magnitude in the unstable subspace $\mathbb{E}^u$ in $O(n \log n)$ steps with high probability. Formally, let $A = \nabla f(\mathbf{0})$ defined in Section 3, and define the remainder term of f as

$$R(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - A\boldsymbol{x}. \tag{6}$$

Because f is smooth in E, by Taylor's theorem there exists a constant $H \in \mathbb{R}_{>0}$ such that for all $\boldsymbol{x}$ in E, $\|R(\boldsymbol{x})\| \leq H\|\boldsymbol{x}\|^2$, and we can rewrite the process in E as,

$$\boldsymbol{X}(k+1) - \boldsymbol{X}(k) = \frac{1}{n}(A\boldsymbol{X}(k) + R(\boldsymbol{X}(k)) + \boldsymbol{U}(k+1)). \tag{7}$$

Since $\mathbf{0}$ is hyperbolic, the process is expanding in the unstable subspace $\mathbb{E}^u$, and contracting in stable subspace $\mathbb{E}^s$ with respect to A defined in Section 3. Let P^u and P^s be the projection operators for $\mathbb{E}^s$ and $\mathbb{E}^u$ respectively. We can decompose the process $\boldsymbol{X}$, from Equation (2), into the *unstable component* $\boldsymbol{X}^u(k) = P^u\boldsymbol{X}(k)$ and the *stable component* $\boldsymbol{X}^s(k) = P^s\boldsymbol{X}(k)$ for all $k \in \mathbb{N}$,

$$\begin{aligned} \boldsymbol{X}^u(k+1) - \boldsymbol{X}^u(k) =& \frac{1}{n}\left(A\boldsymbol{X}^u(k) + R^u(\boldsymbol{X}(k)) + \boldsymbol{U}^u(k+1)\right) \in \mathbb{E}^u \\ \boldsymbol{X}^s(k+1) - \boldsymbol{X}^s(k) =& \frac{1}{n}\left(A\boldsymbol{X}^s(k) + R^s(\boldsymbol{X}(k)) + \boldsymbol{U}^s(k+1)\right) \in \mathbb{E}^s \end{aligned} \tag{8}$$

where functions $R^u \triangleq P^u R$ and $R^s \triangleq P^s R$, and the random processes $\boldsymbol{U}^s(k+1) \triangleq P^s\boldsymbol{U}(k+1) \in \mathbb{E}^s$, and $\boldsymbol{U}^u(k+1) \triangleq P^u\boldsymbol{U}(k+1) \in \mathbb{E}^u$.

For the unstable component, if we can approximate the expected movement of $\boldsymbol{X}^u$ by the first term $A\boldsymbol{X}^u$, $\|\boldsymbol{X}^u\|$ increases exponentially fast. To this end, we need to show the remainder R^u (and noise) are small. However, the magnitudes of R^u depends on $\boldsymbol{X}$, and can be much larger than $A\boldsymbol{X}^u$ if $\boldsymbol{X}^s$ is much larger than $\boldsymbol{X}^u$. The same issue holds for the stable component. To handle this, we partition the process into $O(\log n)$ phases illustrated in Figure 1 such that in each phases the difference between $\|\boldsymbol{X}^u(k)\|$ and $\|\boldsymbol{X}^s(k)\|$ is not too large. Finally we use induction to show the unstable component gets larger as the process proceeds in phases.

The proof has three parts. Lemma 4.4 shows the magnitude in $\mathbb{E}^s$ decreases rapidly after the process enter $B(\mathbf{0}, r_{\text{in}})$. Lemma 4.5 shows if the process is very close to or at $\mathbf{0}$, the noise of the process ensures the unstable part of the process will be $\Omega((\log n)^{1/3}/\sqrt{n})$ in $O(n \log n)$ time with high probability. Finally, Lemma 4.6 shows if the unstable part of the process is $\Omega((\log n)^{1/3}/\sqrt{n})$, the unstable part doubles in $O(n)$ time with probability $1 - \exp\left(-\Omega\left(\sqrt{\log n}\right)\right) = 1 - o(1/\log n)$.

[4] Note that if the noise is uniform spherical distribution, the covariance matrix is $\frac{1}{d}\mathbb{I}_d$.

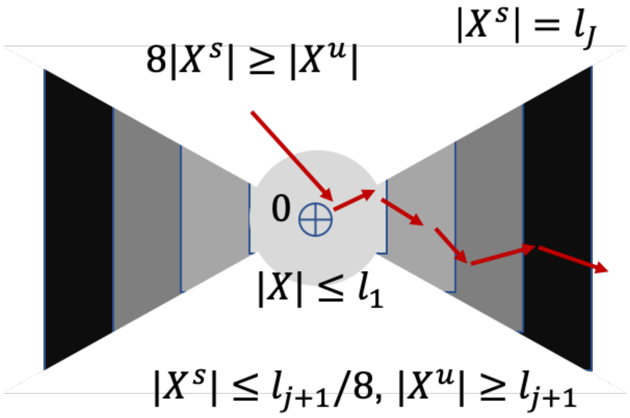


Figure 1: Here we consider the case $\mathbf{0}$ is a saddle point where the x-axis is the unstable subspace (the drift pushes outward in the x-direction), and the y-axis is the stable subspace (the drift pushes inward in the y-direction). For Theorem 4.3, we partition the process around the saddle point $\mathbf{0}$ into $O(\log n)$ phases. In phase 0 (the white region) by Lemma 4.4 the process either hits phase 1 (the lightest gray region which is the union of the ball $B(\mathbf{0}, l_1 \sin\theta_{us})$ and the region $\|\boldsymbol{x}^u\| < l_1$ and $\|\boldsymbol{x}^s\| \leq \|\boldsymbol{x}^u\|/8$) or later phases. Lemma 4.5 shows the process hits phase 2 (the second lightest region) or later, in $O(n\log n)$ with probability $1 - o(1)$. The Lemma 4.6 shows if the process is in phase j, it phase $j+1$ or the later (the next darker region) in additional $O(n)$ times.

Proof for Theorem 4.3. Here we only prove the case of saddle point; the case of repelling point is simpler. Let $\mu_u > 0$ be the minimum real part of the eigenvalues of A in $\mathbb{E}^u$ defined in Section 3.1.1. We set $r_{\text{reg}} > 0$ sufficiently small (which will be specified later), and $\tau = O(n\log n)$. We show there exists $T \leq \tau$ such that after T steps both $\|\boldsymbol{X}^s(T)\| \leq r_{\text{reg}}/16$ and $\|\boldsymbol{X}^u(T)\| \geq r_{\text{reg}}/2$, we have $\|\boldsymbol{X}(T)\| \geq \|\boldsymbol{X}^u(T)\| - \|\boldsymbol{X}^s(T)\| > r_{\text{reg}}/4$. The proof is completed by taking $r_{\text{exit}} = r_{\text{reg}}/4$.

We define a length $J = \left\lceil \log\left((r_{\text{reg}}\sqrt{n})/((\log n)^{1/3})\right)\right\rceil = O(\log n)$ sequence, (l_j),

$$l_J = r_{\text{reg}},\ l_{j-1} = \frac{1}{2} l_j \text{ for } j = 2, 3, \ldots, J, \text{ and } l_1 \in \left[\frac{(\log n)^{1/3}}{2\sqrt{n}}, \frac{(\log n)^{1/3}}{\sqrt{n}}\right]. \tag{9}$$

With the sequence (l_j), we can partition the processes in $B(\mathbf{0}, r_{\text{reg}})$ into $J + 1 = O(\log n)$ phases: We say the process is in the phase 1 if $\|\boldsymbol{X}(k)\| \leq l_1 \sin\theta_{us}$ where $\theta_{us} > 0$ is the angle between $\mathbb{E}^u$ and $\mathbb{E}^s$ defined in (4) or if $\|\boldsymbol{X}^u(k)\| < l_1$ and $\|\boldsymbol{X}^s(k)\| \leq \|\boldsymbol{X}^u(k)\|/8$. The process is in phases $j > 1$ if $l_{j-1} \leq \|\boldsymbol{X}^u(k)\| < l_j$ and $\|\boldsymbol{X}^s(k)\| \leq \|\boldsymbol{X}^u(k)\|/8$, and otherwise we call it in phase 0. [5]

First in Lemma 4.4, we show in $O(n\log n)$ time with high probability either the stable component $\|\boldsymbol{X}^s\|$ is smaller than the unstable component $\|\boldsymbol{X}^u\|$ and enters some phase $j > 0$, or $\|\boldsymbol{X}\| \leq l_1 \sin\theta_{us}$ and enters the phase 1.

Secondly, by Lemma 4.5, suppose the process is at phase 1, the process reaches phase 2 or later phases within $O(n\log n)$ steps with probability $1 - o(1)$.

Finally, by Lemma 4.6, starting at phase $j > 1$, the process reach phase $j+1$ or later phases within $\tau_j = O(n)$ steps with probability $1 - \exp\left(-\Omega\left(\sqrt{\log n}\right)\right) = 1 - o(1/\log n)$. Taking union bound on these $J = O(\log n)$ phases completes the proof. □

[5] Note that by (5) phase 1 is disjoint with phase $j > 1$, because $\|\boldsymbol{X}^u\| \leq \|\boldsymbol{X}\|/\sin\theta_{us}$.

We put the proofs of the following lemmas in the appendix. To simplify the notion, we reset the index of $\boldsymbol{X}$ to 0 in each lemma.

Lemma 4.4 (Escaping Phase 0). *If $\boldsymbol{X}(0) \in B(\mathbf{0}, r_{\text{reg}})$, in time $\tau_0 = O(n \log n)$, there exists $T_0 \leq \tau_0$ such that $\|\boldsymbol{X}^u(T_0)\| \geq 8\|\boldsymbol{X}^s(T_0)\|$ or $\|\boldsymbol{X}(T_0)\| \leq l_1 \sin\theta_{us}$ with probability $1 - o(1)$.*

Lemma 4.4 is proven using the optional stopping time theorem.

Lemma 4.5 (From Phase 1 to 2). *If $\boldsymbol{X}(0)$ is in phase 1, there are $\tau_1 = O(n \log n)$ and $T_1 \leq \tau_1$ such that $\|\boldsymbol{X}^u(T_1)\| \geq l_2$ and $\|\boldsymbol{X}^s(T_1)\| \leq \|\boldsymbol{X}^u(T_1)\|/8$ with probability at least $1 - o(1)$.*

For Lemma 4.5, because the drift of the process in phase 1 is small, we use the anti-concentration of the noise to show in expectation it can reach phase 2 after $O(n(\log n)^{2/3})$ steps. By Markov's inequality, we show it will happen in $O(n \log n)$ with probability $1 - o(1)$.

Lemma 4.6 (Phase $j > 1$). *If the process is in phase $j > 1$, $\|\boldsymbol{X}^s(0)\| \leq \frac{1}{8}\|\boldsymbol{X}^u(0)\|$ and $l_j \leq \|\boldsymbol{X}^u(0)\| \leq l_{j+1}$, there is $\tau_j = O(n)$ such that $\|\boldsymbol{X}^s(\tau_j)\| \leq \frac{1}{8}l_{j+1}$ and $\|\boldsymbol{X}^u(\tau_j)\| > l_{j+1}$ with probability $1 - \exp\left(-\Omega\left(\sqrt{\log n}\right)\right)$.*

We want to show $\|\boldsymbol{X}^u(k)\|$ in (8) increases rapidly by linear approximation which depends on two things: 1) the remainder term $R^u(\boldsymbol{X}(k))$ being small, and 2) the noise, $\boldsymbol{U}^u(k)$, being small. The remainder term $R^u(\boldsymbol{X}(k))$, however, depends both on $\boldsymbol{X}^u(k)$ and $\boldsymbol{X}^s(k)$, so we need to upper bound the value of $\|\boldsymbol{X}^s(k)\|$ as well. By (6), it is sufficient to upper bound the *quadratic* $\|\boldsymbol{X}(k)\|^2$ for all $0 \leq k \leq \tau_j$ with high probability. However, the standard Chernoff bound and union bound are not enough, so we use a more advanced tail bound for the *maximum deviation.* For the noise part, conditioned on $\|\boldsymbol{X}(k)\|^2$ being small, we can show the Doob martingale $Y_k = \mathbb{E}[\boldsymbol{X}(\tau_j) \mid \boldsymbol{X}(0), \ldots, \boldsymbol{X}(k)]$ is concentrated around $Y_0 = \mathbb{E}[\boldsymbol{X}(\tau_j)]$.

Note that in contrast to Lemmas 4.4 and 4.5 which show upper bounds for stopping times, this lemma characterizes the behavior of $\boldsymbol{X}$ at time τ_j.

4.3 Approaching attracting fixed points

Proposition 4.7. *Let $\mathbf{0}$ be an attracting fixed point of f. If $f \in \mathcal{C}^1$ is smooth in $\mathcal{X}$, there exists r_{reg}, $b > 0$ and $\tau_a = O(n \log n)$, such that*

$$\Pr\left[T_h\left(B\left(\mathbf{0}, \frac{b}{n}\right)\right) \leq \tau_a \mid \boldsymbol{X}(0) \in B(\mathbf{0}, r_{\text{reg}})\right] \geq 1 - o(1/n).$$

The proof is based on standard arguments in convex optimization [43] and is similar to the proof of Lemma 4.4. The values r_{reg} and b only depend on the function f and the parameters of the $1/n$-step D-bounded stochastic process.

Furthermore, after the process reaches $B\left(\mathbf{0}, \frac{b}{n}\right)$, the process will stay close to the fixed point exponentially long.

Proposition 4.8. *Given f and $b > 0$ defined in Proposition 4.7, for all T and constant $r_{\text{out}} > 0$,*

$$\Pr\left[T_e\left(B\left(\mathbf{0}, r_{\text{out}}\right)\right) > T \mid \boldsymbol{X}(0) \in B\left(\mathbf{0}, \frac{b}{n}\right)\right] \geq 1 - T\exp\left(-\Omega(n)\right).$$

We include the proofs of above propositions in appendix for completeness.

5 Convergence of $1/n$-step D-bounded stochastic process with a gradient-like function

In this section, we show how to use above local characterization to show the long-term behavior of (2) with a gradient-like function. At a high level, this theorem shows the process will be $\Theta(1/n)$-close to an attracting fixed point in $O(n \log n)$ time with high probability.

Theorem 5.1. *Given constants $m \in \mathbb{N}$, α, a bounded compact set E, and $f \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$, suppose*

1. *f is smooth, gradient-like, and has a constant number of fixed points $\mathrm{Fix}_f = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m\}$ which are hyperbolic,*
2. *E is positive invariant under the flow with f and the stochastic process, $\Pr[\forall k \in \mathbb{N}, \boldsymbol{X}(k) \in E \mid \boldsymbol{X}(0) \in E] = 1$.*
3. *The noise in the $1/n$-step D-bounded stochastic process with f in (2) is well-behaved. For all $k \in \mathbb{N}$, $\boldsymbol{\beta} \in \mathrm{Repel}_f \cup \mathrm{Saddle}_f$, and $\boldsymbol{x} \in B(\boldsymbol{\beta}, r_{\mathrm{reg}})$,*

$$\frac{\alpha}{d}\mathbb{I}_d \prec \mathrm{Cov}[\boldsymbol{U}(k+1) \mid \boldsymbol{X}(k) = \boldsymbol{x}].$$

There exists $\tau = O(n \log n)$ and a constant $b > 0$ such that the hitting time to a neighborhood of one of the attracting fixed points $\Gamma = \cup_{\boldsymbol{\beta} \in \mathrm{Attract}_f} B\left(\boldsymbol{\beta}, \frac{b}{n}\right)$ is smaller than τ:

$$\Pr\left[T_h\left(\Gamma\right) \leq \tau \mid \boldsymbol{X}(0) \in E\right] = 1 - o(1).$$

To show the process reaches a neighborhood of an attracting fixed point fast, we need to show two parts: locally, the process does not become stuck in any small neighborhood; globally, the process progresses without making entering cycles or having complicated recurrent behavior.

Because the flow is gradient-like, there exists a smooth complete Lyapunov function V for the flow. With this real-value function V, we can control the behavior of the reinforced random walk $\boldsymbol{X}$. Locally, for each fixed point $\boldsymbol{\beta}_i \in \mathrm{Fix}(f)$, we define a small neighborhood $N_i = B(\boldsymbol{\beta}_i, r_i)$ around it containing no additional fixed points. There are two cases: either $\boldsymbol{x} \in E \setminus (\cup_i N_i)$, and we say $\boldsymbol{x}$ is a *regular point*. In this case the complete Lyapunov function V has large (linear) decrements. Otherwise, $\boldsymbol{x} \in N_i$ for some i, we say that $\boldsymbol{x}$ is a *neighborhood point* and V decrements increasingly slowly as it approaches the fixed point $\boldsymbol{\beta}_i$.

The first lemma deals with the regular points, and shows that from them the trajectory will quickly reach a non-regular point. The proof is in appendix.

Lemma 5.2 (regular points)**.** *Given N_i, if $\boldsymbol{X}(0) \notin \cup N_i$, there exists $\boldsymbol{\beta}_i$ and $T = O(n)$ such that $X_T \in N_i$ and $V(\boldsymbol{\beta}_i) < V(\boldsymbol{X}(0))$ with probability $1 - o(1)$.*

The next lemma says that as long as $\boldsymbol{\beta}_i$ is not an attracting fixed point, then from any point in its neighborhood, the process will quickly leave the neighborhood in a manner that decreases the potential function.

Lemma 5.3 (non attracting fixed points)**.** *If $\boldsymbol{\beta}_i$ is not an attracting point, there exist $\tau_i = O(n \log n)$, constants $r_i > 0$ and $\delta > 0$, such that for all $\boldsymbol{x} \in N_i = B(\boldsymbol{\beta}_i, r_i)$ there is $T \leq \tau_i$, such that $X_T \notin N_i$, and $V(\boldsymbol{\beta}_i) > \delta + V(\boldsymbol{X}(T))$ with high probability.*

This is proved in the appendix. The proof is a similar to that of Theorem 4.3 but additionally argues that the process leaves the neighbors of saddle points in a way that decreases the potential function.

Proof of Theorem 5.1. Combining the above two characterizations, we can study the process in two alternating stages.

1. Given an initial condition $\boldsymbol{X}(0) \in E$ where E is compact and positive invariant, if $\boldsymbol{X}(0) \notin \cup_i N_i$, it converges to some N_i in $O(n)$ with high probability by Lemma 5.2.
2. If $\boldsymbol{\beta}_i$ is not an attracting point by Lemma 5.3, the process leaves the region N_i and $V(x) < V(\boldsymbol{\beta}_i) - \delta$ in $O(n \log n)$ time with high probability.
3. After leaving N_i, by Lemma 5.2, the process converges to N_j a neighborhood of another fixed point $\boldsymbol{\beta}_j$ where $V(\boldsymbol{\beta}_j) < V(\boldsymbol{\beta}_i)$ in $O(n)$ steps with high probability.
4. We can repeat these arguments until the process reaches some attracting point. The processes can never return to the neighborhood of the same fixed point twice because $V(\beta(i))$ is always decreasing. Moreover since the number fixed points are constant (and independent to the step size), the alternation between the above stages stops in constant rounds.
5. By Proposition 4.7, after the process hits a neighborhood of some attracting point, it will be $O(1/n)$-far from such point in time $O(n \log n)$ with high probability.

□

6 Warm-up Application: Escaping evolutionary unstable equilibrium

Pointed out by [33, 54], one major weakness of the Nash equilibrium is the equilibrium selection problem. For example, in a symmetric 2×2 coordination games with the following payoff matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{10}$$

There are three Nash equilibria, $(0, 0)$, $(1, 1)$ and a mixed equilibrium $(1/2, 1/2)$. However, the mixed strategy is not robust: if one player plays the strategy 1 with probability slightly greater than $1/2$, the other player will switch to strategy 1 completely. To overcome this weakness, they proposed several new solution concepts. One refinement of Nash equilibrium is *evolutionarily stable state*(ESS) which uses dynamics to study the stability of an equilibrium. We will first define our dynamics, and our result.

6.1 Models: Evolutionary Game Theory

Most of the notions are from Sandholm [60]. We consider games played by a homogeneous population with size n. Each agent chooses a pure strategy from a finite set $S = \{1, \dots, d\}$. The empirical distribution of their strategy is call a population state $\boldsymbol{x} \in \Delta(S) = \{\boldsymbol{x} \in {\mathbb{R}_{\geq 0}}^d : \|\boldsymbol{x}\|_1 = 1\}$ with $\boldsymbol{x}_j$ representing the proportion of agents choosing pure strategy j. A (single-) population game has a payoff function $F : \Delta(S) \to \mathbb{R}^d$ where $F_j(\boldsymbol{x})$ represents the payoff to strategy j when the population state is $\boldsymbol{x}$. For example in the symmetric coordination game (10), $F(\boldsymbol{x}) = A\boldsymbol{x}$.

Population state $\boldsymbol{x}^*$ is a *Nash equilibrium* of F if no agent can increase his payoff by unilaterally changing its strategy. Formally,

$$\boldsymbol{x}^*_i > 0 \Rightarrow F_i(\boldsymbol{x}) \geq F_j(\boldsymbol{x}) \text{ for all } j \in S.$$

Intuitively, $\boldsymbol{x}^*$ is an *evolutionarily stable state* (ESS) if the strategies employed in $\boldsymbol{x}^*$ remain best responses even after a small perturbation of $\boldsymbol{x}^*$. Formally, $\boldsymbol{x}^*$ is an *evolutionarily stable state* (ESS) of F if there is an $\epsilon_e > 0$ such that for all $\boldsymbol{x} \neq \boldsymbol{x}^*$ and $0 < \epsilon < \epsilon_e$

$$(\boldsymbol{x} - \boldsymbol{x}^*)^\top F(\epsilon\boldsymbol{x} + (1-\epsilon)\boldsymbol{x}^*) < 0.$$

Note that the mixed strategy for the symmetric coordination game is a Nash equilibrium but not a evolutionarily stable state.

A *revision protocol* is a map $\rho : \mathbb{R}^d \times \Delta(S) \to {\mathbb{R}_{\geq 0}}^{d\times d}$ which maps a payoff vector $\boldsymbol{\pi} = F(\boldsymbol{x})$ and a population state $\boldsymbol{x}$ to a non-negative matrix where $\rho_{ij}(\pi, \boldsymbol{x})$ is the conditional switch rate from strategy i to strategy j. At each round, one agent, chosen uniformly at random, receives a revision opportunity. If the agent is playing strategy $i \in S$, it switches to strategy $j \neq i$ with probability $\rho_{ij}(\boldsymbol{\pi}, \boldsymbol{x})$ (we will omit the input for simplicity). Therefore, a population game F on n agents with a revision protocol ρ is a Markov process on the state space $\Delta(S) \cap \frac{1}{n}\mathbb{Z}^d$ and the transition probability is

$$P(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{z}) = \begin{cases} \boldsymbol{x}_i\rho_{ij} & \text{if } \boldsymbol{z} = \frac{1}{n}(\boldsymbol{e}_j - \boldsymbol{e}_i), i, j \in S, i \neq j, \\ 1 - \sum_{i\in S}\sum_{j\neq i} \boldsymbol{x}_i\rho_{ij} & \text{if } \boldsymbol{z} = \boldsymbol{0}. \end{cases}$$

In this paper, we only consider the *logit choice rule*

$$\rho_{ij}(\boldsymbol{\pi}, \boldsymbol{x}) = \frac{\exp\left(\frac{1}{\eta}\boldsymbol{\pi}_j\right)}{\sum_{k\in S}\exp\left(\frac{1}{\eta}\boldsymbol{\pi}_k\right)}. \tag{11}$$

The parameter $\eta > 0$ is called the *noise level*: if η is large, the choice probabilities are close to uniform, but if η is near zero, choices are best-response with probability close to one. This dynamics is also call the *smooth best response dynamics* [7].

6.2 Main result: Evolutionary Game Theory

In game theory, a solution concept is a formal rule for predicting how a game will be played. The typical concept, the Nash equilibrium, predicts that the outcome of a game is any point from which no player wants to deviate. However, let us go back to the mixed equilibrium in the coordination game. Shown in Figure 2, the mixed equilibrium is *unstable* in the smooth best response dynamics. One may doubt if the mixed Nash equilibrium of the coordination game is "an outcome of this game." The population state always escapes the mixed equilibrium when the process starts not exactly at it. Here we show something stronger: even the process start at the mixed Nash equilibrium, it will escape it in a short time.

Several results show that ESS are locally stable under many evolutionary dynamics [60]. Our Theorem 6.1 shows the other direction: the mixed equilibria which is not an ESS is also not (stochastically) stable under the Logit dynamics.

Theorem 6.1 (Escaping unstable Nash equilibrium)**.** *Given the symmetric coordination game with payoff matrix A in* (10) *over n agents on the revision protocol with logit choice rule* (11) *with noise level $\eta > 0$, we have a Markov chain $\boldsymbol{X}^{(n,\eta)}(k)$ on the space $[0,1]$ which encodes the fraction of population playing strategy* 1*. Given $0 < \eta_0 < 1/2$, there exists $r > 0$ and $\tau = O(n \log n)$ such that for all $0 < \eta \leq \eta_0$*

$$\Pr[\exists T \leq \tau, \boldsymbol{X}^{(n,\eta)}(T) \notin B(1/2, r) \mid \boldsymbol{X}^{(n,\eta)}(0) = 1/2] = 1 - o(1).$$

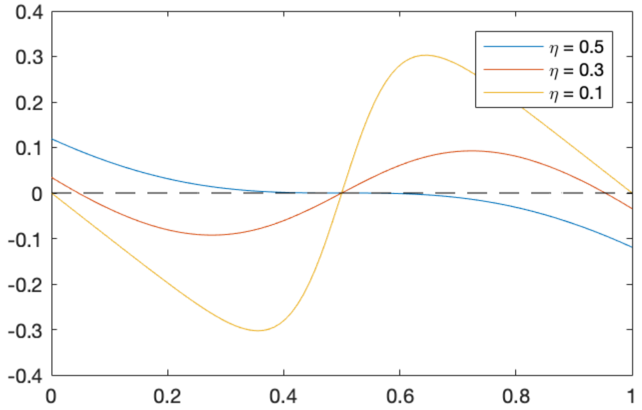


Figure 2: Here we plot $f^{(\eta)}(x)$ in (12) under different $\eta \leq 0.5$. Note that 0.5 is a repelling fixed point for $f^{(\eta)}$ as long as the noise level is smaller than 0.5.

Proof sketch. This is just a corollary of Theorem 4.3. The Markov chain $\boldsymbol{X}^{(n,\eta)}$ is a $1/n$-step 1-bounded stochastic process on $E = [0,1]$ with

$$f^{(\eta)}(\boldsymbol{x}) = \frac{\exp\left(\frac{1}{\eta}\boldsymbol{x}\right)}{\exp\left(\frac{1}{\eta}\boldsymbol{x}\right) + \exp\left(\frac{1}{\eta}(1-\boldsymbol{x})\right)} - \boldsymbol{x}. \tag{12}$$

The mixed equilibrium $\boldsymbol{x} = 1/2$ is a repelling fixed point when $(f^{(\eta)})'(1/2) = 1/(2\eta) - 1 > 0$. Therefore Theorem 6.1 follows the observation that the noise level is large enough to apply Theorem 4.3. We can use a coupling argument to handle the case where $\eta < \eta_0$. □

7 (Dis)agreement in Planted Community Networks

Opinion dynamics on networks study how a set of opinions evolve over a network. In this case, we study how two mutually exclusive competing opinions evolve in graphs with community structure: the maximum expected consensus time on a broad set of stochastic opinion formation dynamics on binary opinions called ***Node dynamics*** [61, 62] in the planted community model. Node dynamics are parameterized by an update function $f : [0,1] \to [0,1]$. In the beginning, each agent holds a binary "opinion", either red or blue. Then, in each round, an agent is uniformly chosen and updates its opinion to red with probability $f(r)$ and blue with probability $1 - f(r)$ where r is the fraction of its neighbors with the red opinion.

By changing f, one can capture many previously studied dynamics, including:

Voter Model: Update a node's opinion to that of a randomly chosen neighbor.

Iterative majority: Update a node's opinion to the majority opinion its neighbors.

Iterative k-majority: Update a node's opinion to the majority opinion of k randomly chosen (with replacement) neighbors.

ρ-noisy majority model: [25, 32] Update a node's opinion to majority opinion its neighbors with probability $1 - \rho$ and uniformly at random with probability ρ.

We model this with a planted community model where n nodes on a complete weighted graph are divided into two equal sets which we call communities. Edges within each community have weight p while edges spanning both communities have weight q. This can also be thought of a block-model which has a long history in the sociology literature [63].

Note that in the voter model with $q > 0$, the graph is not two isolated graphs and the dynamic reaches consensus in $\Theta(n^2)$ time by standard analysis. On the other hand, in the iterative majority with $p > q$, if all nodes in one community have blue opinion and the other has red opinion, the process can never reach a consensus.

7.1 Model: graph with community structure and node dynamics

In this work, we consider *Node dynamics* [61] on block models with two equal size communities:

Definition 7.1 (bi-block model [21, 69])**.** Given p and q where $1/2 \leq p \leq 1$ and $q = 1 - p$, and the set of n vertices V which can be decomposed into two equal size communities V_1 and V_2, we define a *bi-block model* $K(n,p,q) = (V,w)$ which is a weighted complete graph where

$$w(u,v) = \begin{cases} p \text{ if } u,v \text{ are in the same community;} \\ q \text{ o.w.} \end{cases} \tag{13}$$

We consider two opinions, *red* and *blue*. In each round, each agent has opinion either red or blue. A *configuration* $\boldsymbol{X}^{\text{ND}} \in [0,1]^2$ denotes the fraction of nodes having opinion red in each community: X_1^{ND} (and X_2^{ND}) encodes the fraction of node having opinion *red* in the community V_1 (and V_2). We call $\boldsymbol{X}^{\text{ND}} = (0,0)$, where everyone has the blue opinion, a blue consensus; and $(1,1)$ a red consensus. Given a configuration $\boldsymbol{X}^{\text{ND}}$ and $K(n,p,q)$, the (weighted) fractions of neighbors with red opinion in community 1 and 2 are respectively

$$p\,X_1^{\text{ND}} + q\,X_2^{\text{ND}} \text{ and } q\,X_1^{\text{ND}} + p\,X_2^{\text{ND}}.$$

Definition 7.2. An *majority-like update function* is a function $f_{\text{ND}} : [0,1] \to [0,1]$ with the following properties:

Monotone $\forall x,y \in [0,1]$, if $x < y$, then $f_{\text{ND}}(x) \leq f_{\text{ND}}(y)$.

Symmetric $\forall t \in [0,1]$, $f_{\text{ND}}(1/2+t) = 1 - f_{\text{ND}}(1/2-t)$.

Absorption $f_{\text{ND}}(0) = 0$ and $f_{\text{ND}}(1) = 1$.

Rich-get-richer for all $1/2 < x < 1$, $x < f_{\text{ND}}(x)$.

In this work, we further require the update function to have an "S" shape— $f_{\text{ND}} \in \mathcal{C}^2$ which is strictly convex in $[0,0.5]$, and strictly concave in $[0.5,1]$, and call such function a ***smooth majority-like update function***

We define node dynamics on bi-block models as follows:

Definition 7.3. Given a $K(n,p,q)$, an update function f_{ND}, and an initial configuration $\boldsymbol{X}^{\text{ND}}(0)$, a ***node dynamic*** $\text{ND}(K(n,p,q), f_{\text{ND}}, \boldsymbol{X}^{\text{ND}}(0))$ is a stochastic process over configurations, $\boldsymbol{X}^{\text{ND}}$. The dynamics proceeds in rounds. At round $k+1 \geq 1$, a node v is picked uniformly at random. If $v \in V_1$, it updates its opinion to red with probability $f_{\text{ND}}(p\,X_1^{\text{ND}} + q\,X_2^{\text{ND}})$, and blue otherwise. The case of $v \in V_2$ is defined similarly. Equivalently, at round $k+1$

$$\frac{n}{2}\left(\boldsymbol{X}^{\text{ND}}(k+1) - \boldsymbol{X}^{\text{ND}}(k)\right) = \begin{cases} (1,0) & w.p. \frac{1}{2}(1 - X_1^{\text{ND}}(k)) f_{\text{ND}}(p\,X_1^{\text{ND}} + q\,X_2^{\text{ND}}) \\ (-1,0) & w.p. \frac{1}{2} X_2^{\text{ND}}(k)(1 - f_{\text{ND}}(p\,X_1^{\text{ND}} + q\,X_2^{\text{ND}}) \\ (0,1) & w.p. \frac{1}{2}(1 - X_2^{\text{ND}}(k)) f_{\text{ND}}(q\,X_1^{\text{ND}} + p\,X_2^{\text{ND}}) \\ (0,-1) & w.p. \frac{1}{2} X_2^{\text{ND}}(k)(1 - f_{\text{ND}}(q\,X_1^{\text{ND}} + p\,X_2^{\text{ND}}) \\ 0 & o.w. \end{cases} \tag{14}$$

In this paper, we will use consensus time to study the interaction between the update function f_{ND} and the network parameters $K(n, p, q)$. The *consensus time* of a node dynamic, $\mathrm{ND}(K(n, p, q), f_{\mathrm{ND}}, \boldsymbol{X}^{\mathrm{ND}}(0))$, is a stopping time when $\boldsymbol{X}^{\mathrm{ND}}$ is either red or blue consensus. The *maximum consensus time* $\mathrm{ME}(K(n, p, q), f_{\mathrm{ND}})$ is the maximum expected consensus time over any initial configuration.

7.2 Questions and Results

First we note that the process $\boldsymbol{X}^{\mathrm{ND}}$ is a $1/n$-step 2-bounded stochastic process with a function $F_{\mathrm{ND}} : [0,1]^2 \to [0,1]^2$

$$F_{\mathrm{ND}}(x_1, x_2) \triangleq \left(f_{\mathrm{ND}}\left(p\, x_1 + q\, x_2\right) - x_1,\ f_{\mathrm{ND}}\left(p\, x_2 + q\, x_1\right) - x_2\right), \tag{15}$$

and its mean-field approximation φ_{ND} is

$$\frac{d}{dt}\varphi_{\mathrm{ND}}(\boldsymbol{x}, t) = F_{\mathrm{ND}}\left(\varphi_{\mathrm{ND}}(\boldsymbol{x}, t)\right). \tag{16}$$

Furthermore, the consensus states $(0, 0)$ and $(1, 1)$ are fixed points of F_{ND} when f_{ND} is majority-like.

As a warm-up, we consider the iterative majority dynamics, and the voter model.

For iterative majority, the update function is

$$f_{\mathrm{majority}}(x) = \begin{cases} 1 & \text{if } x > 1/2; \\ 1/2 & \text{if } x = 1/2; \\ 0 & \text{if } x < 1/2. \end{cases}$$

Thus in addition to the consensus states, $(1, 0)$ and $(0, 1)$ are attracting fixed points of F_{majority}.[6] Therefore, it is not hard to see the maximum consensus time of iterative majority dynamics is very large.

Proposition 7.4 (Iterative maority)**.** *For all $p > 1/2$ and $n \geq 2$*

$$\mathrm{ME}(K(n, p, q), f_{\mathrm{majority}}) = \infty.$$

To show this, suppose the node dynamic starts at $(1, 0)$, $\boldsymbol{X}^{\mathrm{ND}}(0) = (1, 0)$. By Definition 7.3, every node never changes its opinion.

On the other hand, for voter model, the update function is

$$f_{\mathrm{voter}}(x) = x,$$

and $F_{\mathrm{voter}}(x_1, x_2) = \left(q\left(x_2 - x_1\right), q\left(x_1 - x_2\right)\right)$, which suggests the process will converges to the line $x_1 = x_2$. Moreover, by previous results on the voter model, we know the process reaches consensus in $\Theta(n^2)$ time. Formally,

Proposition 7.5 (Voter model)**.** *For all $p < 1$, and large enough n,*

$$\mathrm{ME}(K(n, p, q), f_{\mathrm{voter}}) = \Theta(n^2).$$

Now we consider smooth majority-like function which is in between voter model or iterative majority. We will use our results in Section 5 to analyze the consensus time of $\boldsymbol{X}^{\mathrm{ND}}$ in (14).

[6] Because for all $1 - 1/(2p) < x_1 \leq 1$ and $0 \leq x_2 < 1/(2p)$, $F_{\mathrm{majority}}(x_1, x_2) = (1 - x_1,\ -x_2)$, the drift in the first coordinate is positive and negative in the second coordinate when (x_1, x_2) is around $(1, 0)$.

Theorem 7.6 (Smooth majority-like function)**.** *Given* $1/2 < p < 1$ *and a smooth majority-like function* f_{ND} *in Definition 7.2, there are three (or two) constants* δ', δ^* *and* δ'' *such that* $0 < \delta' < \delta^* \leq \delta'' < 1$

1. *If* $p - q \in (0, \delta^*) \setminus \{\delta'\}$*,* $\text{ME}(K(n,p,q), f_{\text{ND}}) = O(n \log n)$*.*
2. *If* $p - q \in (\delta^*, 1) \setminus \{\delta''\}$*,* $\text{ME}(K(n,p,q), f_{\text{ND}}) = \exp(\Omega(n))$*.*

The above theorem gives an almost comprehensive characterization of smooth majority-like dynamics on graphs with different community structure: When the community structure is weak (p is close to $1/2$) the node dynamics reach consensus fast. In contrast, when the community structure is strong (p is close to 1), there are bad initial configuration such that the node dynamics cannot reach consensus fast. We exclude these three (or two) points δ', δ^* and δ'' due to technical reasons. Specifically, if $p - q$ is in those values, some fixed points of F_{ND} may not be hyperbolic, and our analysis in Sections 4 and 5 does not apply.

7.3 Proof outline for Theorem 7.6

To show the consensus time is slow, we only need to find a bad initial configuration such that the dynamic takes a long time to escape a neighborhood of it. With Propositions 4.7 and 4.8, it is sufficient to show there is an attracting fixed point other than consensus states $(0,0)$ and $(1,1)$.

However, it is much harder to prove the node dynamics can reach consensus fast from all initial states. Notice that besides the consensus states, there are possibly many fixed points of F_{ND}. For example suppose the process starts at $(0.5, 0.5)$ which is a non-attracting fixed point, by Theorem 4.3, we may show the node dynamic can escape it in $O(n \log n)$. However, what will happen next? Will it return to $(0.5, 0.5)$ after escaping it? Can it traveling between different fixed points and never reach the consensus states? To handle these issue, we need to have some global characterization of the dynamics. A common approach is finding a potential function which decreases strictly along the trajectory of φ. However, it is not easy to find such potential function for our process, because we do not have an analytic representation of F_{ND}. We use an indirect method to show the existence of a potential function for our process. Specifically, in Theorem 7.7 we first carefully analyze the relationship between fixed points in (1), and using this show the flow with F_{ND} is gradient-like and therefore has a potential function (or a strict complete Lyapunov function).

Theorem 7.7 (Phase portrait)**.** *Given* f_{ND} *and* p, q *in the Node Dynamics defined in Theorem 7.6, there exist three constants* $0 < \delta' < \delta^* \leq \delta'' < 1$ *such that the flow with* $\bar{F}_{\text{ND}}$ *defined in* (15) *has three cases:*

1. *When* $p - q \in (0, \delta^*) \setminus \{\delta'\}$*, the flow is a gradient-like system, and the consensus states* $(0,0), (1,1)$ *are the only attracting fixed point.*
2. *When* $p - q \in (\delta^*, 1) \setminus \{\delta''\}$*,* F_{ND} *has an attracting fixed point* $\boldsymbol{\beta}_a \neq (0,0), (1,1)$*.*

A more detailed characterization of δ', δ^* *and* δ'' *is in appendix.*

In summary, we prove the fast convergence result of the Theorem 7.6 in three parts:

1. We show the flow (16) is a gradient-like system and only the consensus states are attracting fixed points (Theorem 7.7).
2. We apply Theorem 5.1 to show the process (14) converges to an arbitrary neighborhood of consensus states in $O(n \log n)$ time with high probability.

3. Finally, after the process (14) reach a small enough neighborhood of consensus states, we couple the process with birth-and-death process can show the process (14) reach a consensus state in $O(n \log n)$ time with high probability.

We defer the proof to the appendix.

8 Provable and Practical Framework for Non-Convex Problems—Bounded Stochastic Gradient Descents

Several machine learning and signal processing applications induce optimization problems with non-convex objective functions. The global optimization of a non-convex objective is an NP-hard problem in general. As a result, a much sought-after goal in applications with non-convex objectives is to find a local minimum of the objective function. One main hurdle in achieving local optimality is the presence of saddle points which can mislead local search methods by stalling their process.

Our analysis in Section 5 can be applied to these problems. Formally, given an objective function $F : \mathbb{R}^d \to \mathbb{R}$, a popular heuristic to minimize F is by stochastic gradient descent(s):

$$\boldsymbol{x}(t+1) = \boldsymbol{x}(t) - \eta \nabla F(\boldsymbol{x}(t)) + \boldsymbol{U}(t+1), \tag{SGD}$$

which can be seen as an η-step stochastic process with ∇F. (We want to emphasize there are multiple variants (SGD) with different noise term $\boldsymbol{U}$.) Stochastic gradient descent is well-studied when the objective function is convex. In this section, however, we want to study the convergence property when F is non-convex. In particular, we are interested in the time complexity for step size η.

Several works design particular stochastic gradient descents (SGD) by adding specific noise $\boldsymbol{U}(t)$ on the exact gradient at each step and then show they (SGD) escape saddle points and converge to a local minimal in polynomial time. [31, 39] However, this direction is not practical. In the typical usage of (SGD), the noise $\boldsymbol{U}(t)$ is not necessary solely from the design of the algorithm and their analyses do not apply to these (SGD). One prominent example is from the mini-batch algorithm where the gradient is estimated from batches of training examples (called a "mini-batch") at each step. Here the noise is induced from the partition of the mini-batches. 2) Even if algorithm designers have the choice to design such (SGD), provable guarantees for escaping saddle points may not be the most critical aspects of the algorithm design. Therefore, rather than offering a particular (SGD) with provable guarantees, we provide a framework to design or verify an (SGD) can escape saddle points and converge to a local minima.

8.1 Bounded stochastic gradient descent algorithm

We now state a general stochastic gradient descent with bounded martingale difference perturbation, and show such processes converge to a local minimal.

Algorithm 1: Bounded Stochastic Gradient descent Algorithm

```
Result: Finding a local minimal value
Input  : An objective function $F : \mathbb{R}^d \to \mathbb{R}$, the step length $\eta$, the running time $T$, and the
         initial point $\boldsymbol{X}^{\text{SGD}}(0)$
Output: A point $\boldsymbol{x} \in \mathbb{R}^d$
for $t$ = 0,1,. . . , $T$ do
    Sample a perturbation $\boldsymbol{U}(t+1)$ with properties defined in Theorem 8.1
    $\boldsymbol{X}^{\text{SGD}}(t+1) = \boldsymbol{X}^{\text{SGD}}(t) - \eta\left(\nabla F(\boldsymbol{X}^{\text{SGD}}(t)) + \boldsymbol{U}(t+1)\right)$
end
```

Using the same argument for Theorem 5.1, we have:

Theorem 8.1 (Bounded Stochastic Gradient Descents)**.** *Given a constant d, an objective function $F \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R}^d)$ and $\nabla^2 F$ is Lipschitz, a compact set $E \subset \mathbb{R}^d$, and constants $D, \alpha > 0$, such that*

1. *F has a constant number of critical points in E, $\{\boldsymbol{\beta} \in E : \nabla F(\boldsymbol{\beta}) = 0\}$ which are non-degenerate: $\nabla^2 F(\boldsymbol{\beta})$ is invertible.*
2. *E is positive invariant under the flow with ∇F and the process in Algorithm 1,*$\Pr[\forall k \in \mathbb{N}, \boldsymbol{X}^{\text{SGD}}(k) \in E \mid \boldsymbol{X}^{\text{SGD}}(0) \in E]$
3. *The perturbation of the process in Algorithm 1 is well-behaved*
   - *(a) For all $\boldsymbol{x} \in E$ and t, $\mathbb{E}[\boldsymbol{U}(t+1) \mid \boldsymbol{X}^{\text{SGD}}(t) = \boldsymbol{x}] = 0$*
   - *(b) For all t, $\|\boldsymbol{U}(t+1)\| \leq D$,*
   - *(c) For all t, $\boldsymbol{\beta}$ with $\nabla F(\boldsymbol{\beta}) = 0$, and $\boldsymbol{x} \in B(\boldsymbol{\beta}, \epsilon)$, $\frac{\alpha}{d}\mathbb{I}_d \prec \operatorname{Cov}[\boldsymbol{U}(t+1) \mid \boldsymbol{X}(t) = \boldsymbol{x}]$.*

There exist $T = O\left(\frac{\log 1/\eta}{\eta}\right)$ and a constant $b > 0$ such that for all initial points $\boldsymbol{X}^{\text{SGD}}(0) \in E$ $\|x_T - x^*\| \leq b\eta$ for some local minimal x^* with high probability.

8.2 Discussion and Related work

For the time complexity with respect to the step size η, this framework contains several previous results as special cases, and provides a tighter convergence time upper bound. For example, Ge et al. [31] propose the following algorithm:

Algorithm 2: Noisy Gradient Descent [31]

```
Result: Finding a local minimal value
Input  : An objective function $F : \mathbb{R}^d \to \mathbb{R}$, the step length $\eta$, the running time $T$, and the
         initial point $x_1$
Output: A point $x \in \mathbb{R}^d$
for $t$ = 1,2. . . , $T$ do
    Sample a perturbation $\boldsymbol{U}(t+1) \sim S^{d-1}$(a random point on unit sphere)
    $x_{t+1} = x_t - \eta\left(\nabla F(x_t) + \boldsymbol{U}(t+1)\right)$
end
```

They show the convergent time to constant neighborhood of some local minima is $O(1/\eta^2)$ which is weaker than Theorem 8.1 when the objective function satisfies our condition.

Similarly, Jin et al. [39] proposes a perturbed gradient descent algorithm:

Algorithm 3: Perturbed Gradient Descent [39]

```
Result: Finding a local minimal value
Input : An objective function $F : \mathbb{R}^d \to \mathbb{R}$, the step length $\eta$, the running time $T$, and the
        initial point $x_1$
Output: A point $x \in \mathbb{R}^d$
for $t$ = 1,2... do
    if $\|\nabla F(x_t)\|$ is "small" then
        $\boldsymbol{U}(t+1) \sim S^{d-1}$
    else
        $\boldsymbol{U}(t+1) = 0$
    end
    $x_{t+1} = x_t - \eta\left(\nabla F(x_t) + \boldsymbol{U}(t+1)\right)$
end
```

They show the convergent time to a constant neighborhood of some local minimal is $O((\log 1/\eta)^4/\eta)$ which is weaker than Theorem 8.1 when the objective function satisfies our condition.

Remark 8.2. Here we put some comparison between Theorem 8.1 and previous work.

1. The running time is optimal with respect to step size $1/n$, $O(n \log n)$.

2. This result applies to a larger family of stochastic gradient descent. Instead of requiring the perturbation to be independent uniform points in the unit sphere, our result only requires the noises are bounded martingale and the covariance matrix is positive definite (Theorem 8.1).

3. In gradient flow, the stable and unstable subspace are orthogonal at the saddle point (the Hessian of the function is symmetric), but it is not true for hyperbolic saddle points of non-gradient flow. Our result in Theorem 5.1 extends to reinforced random walks with non-gradient flows.

On the other hand, our result doesn't handle some aspects in Ge et al. [31], Jin et al. [39]:

1. We consider the step size η is small enough, but do not provide a closed-form upper bound.

2. We hope that our analysis can be extended to the dimension free case, but at the moment it requires constant fixed dimension.

References

[1] Mohammed Amin Abdullah and Moez Draief. Global majority consensus by local majority polling on graphs of a given degree sequence. *Discrete Applied Mathematics*, 180:1–10, 2015.

[2] Ethan Akin. *The general topology of dynamical systems*, volume 1. American Mathematical Soc., 2010.

[3] Andrea Baronchelli, Vittorio Loreto, Luca Dall'Asta, and Alain Barrat. Strategy, topology, and all that. In *The Evolution of Language: Proceedings of the 6th International Conference (EVOLANG6), Rome, Italy, 12-15 April 2006*, page 11. World Scientific, 2006.

[4] Luca Becchetti, Andrea Clementi, Emanuele Natale, Francesco Pasquale, and Luca Trevisan. Stabilizing consensus with many opinions. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 620–635. Society for Industrial and Applied Mathematics, 2016.

[5] Eli Ben-Naim, Laurent Frachebourg, and Paul L Krapivsky. Coarsening and persistence in the voter model. *Physical Review E*, 53(4):3078, 1996.

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

[7] Michel Benaïm and Jörgen W Weibull. Deterministic approximation of stochastic evolution in games. *Econometrica*, 71(3):873–903, 2003.

[8] Florence Bénézit, Patrick Thiran, and Martin Vetterli. Interval consensus: from quantized gossip to voting. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3661–3664. IEEE, 2009.

[9] Itai Benjamini, Siu-On Chan, Ryan O'Donnell, Omer Tamuz, and Li-Yang Tan. Convergence, unanimity and disagreement in majority dynamics on unimodular graphs and random graphs. *Stochastic Processes and their Applications*, 126(9):2719–2733, 2016.

[10] Lawrence E Blume. The statistical mechanics of strategic interaction. *Games and economic behavior*, 5(3):387–424, 1993.

[11] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[12] Claudio Castellano, Daniele Vilone, and Alessandro Vespignani. Incomplete ordering of the voter model on small-world networks. *EPL (Europhysics Letters)*, 63(1):153, 2003.

[13] Claudio Castellano, Vittorio Loreto, Alain Barrat, Federico Cecconi, and Domenico Parisi. Comparison of voter and glauber ordering dynamics on networks. *Physical review E*, 71(6): 066107, 2005.

[14] Charles C Conley. *Isolated invariant sets and the Morse index.* Number 38. American Mathematical Soc., 1978.

[15] Colin Cooper, Robert Elsässer, and Tomasz Radzik. The power of two choices in distributed voting. In *International Colloquium on Automata, Languages, and Programming*, pages 435–446. Springer, 2014.

[16] Colin Cooper, Tomasz Radzik, Nicolás Rivera, and Takeharu Shiraga. Fast plurality consensus in regular expanders. *arXiv preprint arXiv:1605.08403*, 2016.

[17] J Theodore Cox and David Griffeath. Diffusive clustering in the two dimensional voter model. *The Annals of Probability*, pages 347–370, 1986.

[18] James Cruise and Ayalvadi Ganesh. Probabilistic consensus via polling and majority rules. *Queueing Systems*, 78(2):99–120, 2014.

[19] Guillaume Deffuant, Frédéric Amblard, Gérard Weisbuch, and Thierry Faure. How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5(4), 2002.

[20] M.H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, pages 118–121, 1974.

[21] Paul DiMaggio. Structural analysis of organizational fields: A blockmodel approach. *Research in organizational behavior*, 1986.

[22] Benjamin Doerr, Leslie Ann Goldberg, Lorenz Minder, Thomas Sauerwald, and Christian Scheideler. Stabilizing consensus with the power of two choices. In *Proceedings of the twenty-third annual ACM symposium on Parallelism in algorithms and architectures*, pages 149–158. ACM, 2011.

[23] Richard Durrett, James P Gleeson, Alun L Lloyd, Peter J Mucha, Feng Shi, David Sivakoff, Joshua ES Socolar, and Chris Varghese. Graph fission in an evolving voter model. *Proceedings of the National Academy of Sciences*, 2012.

[24] Andreas Eberle. Markov processes. *Lecture Notes at University of Bonn*, 2009.

[25] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica: Journal of the Econometric Society*, pages 1047–1071, 1993.

[26] Xiequan Fan, Ion Grama, and Quansheng Liu. Hoeffdings inequality for supermartingales. *Stochastic Processes and their Applications*, 122(10):3545–3559, 2012.

[27] Diodato Ferraioli. Logit dynamics: a model for bounded rationality. *ACM SIGecom Exchanges*, 12(1):34–37, 2013.

[28] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[29] Jie Gao, Bo Li, Grant Schoenebeck, and Fang-Yi Yu. Engineering agreement: The naming game with asymmetric and heterogeneous agents. In *AAAI*, pages 537–543, 2017.

[30] Jie Gao, Grant Schoenebeck, and Fang-Yi Yu. The volatility of weak ties: Co-evolution of selection and influence in social networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 619–627, 2019. URL `http://dl.acm.org/citation.cfm?id=3331748`.

[31] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[32] Reza Gheissari and Anna Ben Hamou. Aimpl: Markov chain mixing times, available at http://aimpl.org/markovmixing.

[33] John C Harsanyi, Reinhard Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

[34] Thomas P Hayes. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2005.

[35] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

[36] Richard A Holley and Thomas M Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663, 1975.

[37] Petter Holme and M E J Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 74(5 Pt 2):056108, November 2006.

[38] Roger A Horn, Roger A Horn, and Charles R Johnson. *Matrix analysis*. Cambridge university press, 1990.

[39] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[40] Olav Kallenberg and Rafal Sztencel. Some dimension-free features of vector-valued martingales. *Probability Theory and Related Fields*, 88(2):215–247, 1991.

[41] Yashodhan Kanoria, Andrea Montanari, et al. Majority dynamics on trees and the dynamic cavity method. *The Annals of Applied Probability*, 21(5):1694–1748, 2011.

[42] Michael Kearns and Jinsong Tan. Biased voting and the democratic primary problem. In *International Workshop on Internet and Network Economics*, pages 639–652. Springer, 2008.

[43] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

[44] Paul L Krapivsky and Sidney Redner. Dynamics of majority rule in two-state interacting spin systems. *Physical Review Letters*, 90(23):238701, 2003.

[45] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations. *Journal of Machine Learning Research*, 20:40:1–40:47, 2019. URL `http://jmlr.org/papers/v20/17-526.html`.

[46] Thomas M Liggett. Coexistence in threshold voter models. *The Annals of Probability*, pages 764–802, 1994.

[47] Thomas M Liggett et al. Stochastic models of interacting systems. *The Annals of Probability*, 25(1):1–29, 1997.

[48] Daniel McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

[49] Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

[50] Andrea Montanari and Amin Saberi. Convergence to equilibrium in local interaction games and ising models. *arXiv preprint arXiv:0812.0198*, 2008.

[51] Elchanan Mossel and Grant Schoenebeck. Arriving at consensus in social networks. In *The First Symposium on Innovations in Computer Science (ICS 2010)*, January 2010.

[52] Elchanan Mossel, Joe Neeman, and Omer Tamuz. Majority dynamics and aggregation of information in social networks. *Autonomous Agents and Multi-Agent Systems*, 28(3):408–429, 2014.

[53] Ioannis Panageas and Nisheeth K Vishnoi. Mixing time of markov chains, dynamical systems and evolution. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[54] Christos Papadimitriou and Georgios Piliouras. Game dynamics as the meaning of a game. *ACM SIGecom Exchanges*, 16(2):53–63, 2019.

[55] Robin Pemantle et al. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

[56] Robin Pemantle et al. A survey of random processes with reinforcement. *Probability surveys*, 4:1–79, 2007.

[57] Etienne Perron, Dinkar Vasudevan, and Milan Vojnovic. Using three states for binary consensus on complete graphs. In *INFOCOM 2009, IEEE*, pages 2527–2535. IEEE, 2009.

[58] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[59] Clark Robinson. *Dynamical systems: stability, symbolic dynamics, and chaos.* CRC press, 1998.

[60] William H Sandholm. *Population games and evolutionary dynamics.* MIT press, 2010.

[61] Grant Schoenebeck and Fang-Yi Yu. Consensus of interacting particle systems on erdös-rényi graphs. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1945–1964. SIAM, 2018.

[62] Frank Schweitzer and Laxmidhar Behera. Nonlinear voter models: the transition from invasion to coexistence. *The European Physical Journal B-Condensed Matter and Complex Systems*, 67(3):301–318, 2009.

[63] John Scott. *Social network analysis.* Sage, 1988.

[64] Vishal Sood and Sidney Redner. Voter model on heterogeneous graphs. *Physical review letters*, 94(17):178701, 2005.

[65] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering.* CRC Press, 2018.

[66] Krzysztof Suchecki, Victor M Eguiluz, and Maxi San Miguel. Conservation laws for the voter model in complex networks. *EPL (Europhysics Letters)*, 69(2):228, 2005.

[67] Krzysztof Suchecki, Víctor M Eguíluz, and Maxi San Miguel. Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution. *Physical Review E*, 72(3): 036132, 2005.

[68] Omer Tamuz and Ran J Tessler. Majority dynamics and the retention of information. *Israel Journal of Mathematics*, 206(1):483–507, 2015.

[69] Harrison C White, Scott A Boorman, and Ronald L Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, 81(4):730–780, 1976.

[70] Nicholas C Wormald et al. Differential equations for random processes and random graphs. *The annals of applied probability*, 5(4):1217–1235, 1995.

[71] Ahad N. Zehmakan. Opinion forming in binomial random graph and expanders. *CoRR*, abs/1805.12172, 2018. URL `http://arxiv.org/abs/1805.12172`.

Symbols

B Ball in $\mathbb{R}^d$ centered at $\mathbf{0}$. 5, 7, 8, 9, 10, 11, 13, 19, 31, 33, 34, 36, 37, 38, 45, 47

T_e the first exit time. 6, 7, 8, 10, 33

T_h the hitting time. 6, 10, 11, 36, 45

V A complete Lyapunov function. 5, 11, 37, 38

X_1^{ND} The fraction of red opinions in community 1. 15

X_2^{ND} The fraction of red opinions in community 2. 15

$\Re$ Real part of a complex number. 4, 5, 6

$\boldsymbol{U}$ noise. 1, 4, 6, 8, 10, 11, 18, 19, 29, 31, 33, 34, 36, 45, 46

$\boldsymbol{X}^{(n,\eta)}$ Population state of smooth best response dynamics with noise level η. 13

$\boldsymbol{X}^{\mathrm{ND}}$ The fraction of red opinions in each communities. 15, 16, 45, 46, 47, 48, 49

$\boldsymbol{X}$ $1/n$-step stochastic process with f. 1, 4, 6, 7, 8, 9, 10, 11, 19, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 45

$\boldsymbol{\beta}$ a fixed point. 5, 11, 12, 17, 19, 37, 38, 45

$\boldsymbol{e}$ The standard basis. 13

$\boldsymbol{x}$ a point in $\mathcal{X}$. 1, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 16, 18, 19, 26, 27, 30, 31, 33, 35, 36, 37, 38, 45, 46, 47, 48

$\mathbb{N}$ The set of natural numbers, $0, 1, 2, \ldots$. 4, 6, 8, 11, 19, 29, 30, 47

$\mathbb{R}_{>0}$ The set of positive real numbers. 4, 5, 6, 7, 8, 29, 30

$\mathbb{R}_{\geq 0}$ The set of nonnegative real numbers. 12, 13

$\mathcal{X}$ phase space which is $\mathcal{R}^d$. 3, 4, 5, 6, 7, 10, 26, 27

μ_s stable. 6, 31, 32, 35, 36

μ_u unstable. 6, 8, 32, 35, 36

$\mu_{\max}$ maximum real part. 6, 34, 35

ρ Revision protocol. 13

φ_{ND} the induced flow. 16, 39

φ the flow with f. 4, 5, 6, 7, 17, 26, 30, 31, 37

d Dimension of space which is constant. 3, 5, 6, 7, 8, 11, 18, 19, 28, 29, 30, 33, 35

r_{reg} nice radius. 5, 8, 9, 10, 11, 31, 32, 33, 34, 35, 36, 37, 45

θ_{us} The principal angle between $\mathbb{E}^u$ and $\mathbb{E}^s$. 8, 9, 10, 31, 32, 35

A Primer of dynamical system

In this section, we want to introduce more notions for Proposition A.2 and Corollary A.4 which provide an indirect method of showing a flow (1) is gradient-like.

A.1 Fundamental theorem of dynamical system

An opposite concept to "recurrence" is transit. How we show all the non recurrent points are transit? An ideal method is to find a "potential function", $\Psi : \mathcal{X} \to \mathbb{R}$ of the system such that Ψ decrease along the trajectory of the system. To state it formally, we need to consider a more general notion of recurrence than fixed points:

Fixed point A point $\boldsymbol{x} \in \mathcal{X}$ is a *fixed point* if $\mathcal{O}_x = \{x\}$ that is $f(\boldsymbol{x}) = 0$, and we use Fix_f to denote the set of fixed points.

Periodic point A point $\boldsymbol{x} \in \mathcal{X}$ is a *periodic point* of the flow induced by f if $\exists T \geq 0$ such that $\varphi(\boldsymbol{x}, T) = \boldsymbol{x}$, and we use Per_f to denote the set of periodic points.

ω-recurrent For other non-periodic points $\boldsymbol{x} \in$, the long term behavior can be characterized as *ω-limit set* of $\boldsymbol{x}$: $\omega(\boldsymbol{x}) = \{\boldsymbol{y} : \exists t_l \to +\infty, \lim_{l\to\infty} \|\varphi(\boldsymbol{x}, t_l) - \boldsymbol{y}\| = 0\}$, and we call $\boldsymbol{x}$ *ω-recurrent* if $\boldsymbol{x} \in \omega(\boldsymbol{x})$ If we change $+\infty$ to $-\infty$ in above definition, it called *α-limit set* $\alpha(\boldsymbol{x})$ of $\boldsymbol{x}$. We call $L_f \triangleq \overline{\cup_{\boldsymbol{x}\in\mathcal{X}}\omega(\boldsymbol{x})} \cup \overline{\cup_{\boldsymbol{x}\in\mathcal{X}}\alpha(\boldsymbol{x})}$ the *limit set of f*.

Chain recurrent An ϵ-chain of length T from a point $\boldsymbol{x}$ to $\boldsymbol{y}$ is a sequence of points $(\boldsymbol{x}_\ell)_{0\leq\ell\leq n}$ and a sequence of time $(t_\ell)_{1\leq\ell\leq n}$ such that $\boldsymbol{x}_0 = \boldsymbol{x}$, $\boldsymbol{x}_n = \boldsymbol{y}$, and $\|\varphi(\boldsymbol{x}_{i-1}, t_i) - \boldsymbol{x}_i\| < \epsilon$ for $1 \leq \ell \leq n$ with $t_\ell \geq 1$ and $\sum_\ell t_\ell = T$. We define a relation $\sim_{\mathcal{CR}}$ on $\mathcal{CR}_f$. Similar to ω-limit we define $\Omega^+(\boldsymbol{x}) = \cap_{\epsilon>0,T>0}\{\boldsymbol{y} : \exists \text{an } \epsilon, T \text{ chain from } \boldsymbol{x} \text{ to } \boldsymbol{y}\}$, and a point $\boldsymbol{x}$ is said to be *chain recurrent* for the flow f if $\boldsymbol{x} \in \Omega^+(x)$. The set of chain recurrent points of f is called the *chain recurrent set* of f denoted as $\mathcal{CR}_f$. We say $\boldsymbol{x} \sim_{\mathcal{CR}} \boldsymbol{y}$ if and only if $\boldsymbol{x} \in \Omega^+(\boldsymbol{y})$ and $\boldsymbol{y} \in \Omega^+(\boldsymbol{x})$.

It is not hard to show

$$\mathrm{Fix}_f \subseteq \mathrm{Per}_f \subseteq L_f \subseteq \mathcal{CR}_f \subseteq \mathcal{X}$$

.

A.2 Morse-Smale and gradient-like dynamics

Before introducing Morse-Smale, we first define several notions.

Given a hyperbolic fixed point $\boldsymbol{x}$ for a C^r function f, and a neighborhood U of $\boldsymbol{x}$, the *local stable set/manifold* for $\boldsymbol{x}$ in the neighbor U is defined as:

$$W^s_{loc}(\boldsymbol{x}, U, f) \triangleq \{\boldsymbol{y} \in U : \varphi(\boldsymbol{y}, t) \in U, \forall t > 0 \text{ and } d(\varphi(\boldsymbol{y}, t), \boldsymbol{x}) \to 0 \text{ as } t \to \infty\}$$
$$W^u_{loc}(\boldsymbol{x}, U, f) \triangleq \{\boldsymbol{y} \in U : \varphi(\boldsymbol{y}, t) \in U, \forall t < 0 \text{ and } d(\varphi(\boldsymbol{y}, t), \boldsymbol{x}) \to 0 \text{ as } t \to -\infty\}$$

Opposite to the notion of tangency, *transversality* is a geometric notion of the intersection of manifolds. Let $\boldsymbol{x} \in \mathcal{X}$ M and N are C^r manifolds in $\mathcal{X}$. M, N are said to be *transversal* at $\boldsymbol{x}$ if $\boldsymbol{x} \notin M \cap N$; or if $\boldsymbol{x} \in M \cap N$, $T_xM + T_xN = \mathbb{R}^d$ where T_xM and T_xN denote the tangent space of M and N respectively at point x. M and N are said to be *transversal* if they are transversal at every point $\boldsymbol{x} \in \mathcal{X}$.

Definition A.1 (Morse-Smale flow)**.** Let $\varphi(\cdot, \cdot)$ be a flow on $\mathcal{X} = \mathbb{R}^d$. φ is called *Morse-Smale flow* if there are a constant collection of periodic orbits $P_1, \ldots, P_l$ such that

1. P_i is hyperbolic $i = 1, \dots, l$
2. $\mathcal{CR}_f = \mathrm{Per}_f$
3. $W^U(P_i)$ and $W^S(P_j)$ are transversal for all $1 \leq i, j \leq l$.

Furthermore, if the Morse-Smale system does not have cycle, it is further called *gradient-like.*

Here we give a sufficient condition for gradient-like flow on two dimensional manifolds.

Proposition A.2. *Let $\mathcal{X} = \mathbb{R}^2$. A vector field with $f \in \mathcal{C}^1(\mathbb{R}^2, \mathbb{R}^2)$ is a gradient-like flow if:*

1. *f has a finite number of fixed points which are all hyperbolic;*
2. *there are no saddle-connections that is an orbit whose α- and ω -limits are saddle points; and*
3. *each orbit has a unique fixed point as its α-limit and has a unique fixed point as its ω-limit.*

We further call the function f gradient-like.

Let $\{\beta_1, \dots, \beta_m\} = \mathrm{Fix}_f$ be the set of fixed point of f, and W_i^s and W_i^u be the stable and unstable manifold associated to β_i. The Morse-Smale system has the following property.

Lemma A.3. *Let f be a Morse-Smale system on $\mathcal{X}$. Let $\beta_i \succ \beta_j$ mean there is a trajectory not equal to β_i or β_j whose α-limit set is β_i and whose ω-limit set is β_j. Then $\succ$ satisfies:*

anti-reflexive *It is never true that $\beta_i \succ \beta_i$*

partial order *if $\beta_i \succ \beta_j$ and $\beta_j \succ \beta_k$ then $\beta_i \succ \beta_k$*

transversal *If $\beta_i \succ \beta_j$ then $\dim W_i^u \geq W_j^u$*

Morse-Smale systems share several properties with gradient fields: no complicated recurrent motion and existence of “potential function”— Morse function— that is decreasing along trajectories.

Corollary A.4 (Theorem 12 in Akin [2])**.** *If $f \in \mathcal{C}^2$ is a Morse-Smale system then there exists a complete Lyapunov function $V : \mathcal{X} \to \mathbb{R}$ such that*

1. *$V \in \mathcal{C}^2$ is smooth.*
2. *$\mathcal{L}_f V(\boldsymbol{x}) < 0$ for all non fixed points of f.*

B Basic Math

B.1 Markov chain

Let $\mathcal{M} = (X_t, P)$ be a discrete time-homogeneous Markov chain with a finite state space Ω and transition kernel P. For $x, a \in \Omega$, we define $T(a; x)$ to be the *hitting time* for a with initial state x:

$$T(a; x) \triangleq \min\{t \geq 0 : X_t = a, X_0 = x\},$$

and $T(Q; x)$ to be the hitting time to a set of state $Q \subseteq \Omega$—$T(Q; x) \triangleq \min\{t \geq 0 : X_t \in Q, X_0 = x\}$. We further use $\tau(a; x)$ or $\tau(Q; x)$ to denote the *expected hitting time* for a or Q from x.

Due to the memoryless property of Markov chains, sometimes it is useful to analyze its first step. Let's consider a general measurable function $w : \Omega \to \mathbb{R}$. If the Markov chain starts at state $X = x$, the next state is the random variable X', then the average change of $w(X')$ in one transition step is given by

$$(\mathcal{L}w)(x) \triangleq \mathbb{E}_{\mathcal{M}}[w(X') - w(X)|X = x] = \sum_{y \in \Omega} P_{x,y} w(y) - w(x)$$

To reduce notation we will use $\mathbb{E}_{\mathcal{M}}[w(X')|X]$ to denote the expectation of the measurable function $w(X')$ given the previous state at X.

By the Markov property, the expected hitting time $\tau(Q; x) = \mathbb{E}_{\mathcal{M}}[T(Q; x)]$ can be written as linear equations.

$$\begin{aligned} \mathcal{L}\tau(Q; x) &= -1 \text{ where } x \notin Q \\ \tau(Q; x) &= 0 \text{ where } x \in Q \end{aligned}$$

Corollary B.1 (Maximum principle [24]). *Given a Markov chain $\mathcal{M}$ with state space Ω and a set of states $Q \subsetneq \Omega$, suppose $s_Q : \Omega \to \mathbb{R}$ is a non-negative function satisfying*

$$\begin{aligned} \mathcal{L}s(Q; x) &\leq -1 \textit{ where } x \notin Q, \\ s(Q; x) &\geq 0 \textit{ where } x \in Q. \end{aligned} \tag{17}$$

Then $s(Q; x) \geq \tau(Q; x)$ for all $x \notin Q$.

B.2 Linear algebra

Corollary B.2 (Theorem 4.3.50 in Horn et al. [38].). *Let $A \in \mathbb{R}^{d \times d}$ with eigenvalues $eig(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_d\}$ with $\mu_{\min} \triangleq \min \Re(\lambda_i)$ and $\mu_{\max} \triangleq \max \Re(\lambda_i)$. For all $v \in \mathbb{R}^d$,*

$$\mu_{\min} \|v\|^2 \leq v^\top A v \leq \mu_{\max} \|v\|^2.$$

The following lemma is useful to show these two sequences are close to each other.

Lemma B.3 (Discrete Gronwall lemma). *Let $a_{k+1} \leq (1 + \frac{1}{n} L) a_k + b$ with $n > 0$, $L > 0$, $b > 0$ and $a_0 = 0$. Then*

$$a_k \leq \frac{nb}{L} \left(\exp\left(\frac{k}{n}\right) - 1 \right).$$

B.3 Martingale and concentration

In this section we will define martingales and some of its properties. Let $\mathcal{F} = (\mathcal{F}(k))_k$ be a filtration, that is an increasing sequence of σ-field. A sequence X_k is said to be adapted to $\mathcal{F}(k)$ if $X_k \in \mathcal{F}(k)$ for all k. If X_k is sequence with 1) $\mathbb{E}|X_k| < \infty$, 2) X_k is adapted to $\mathcal{F}(k)$, and 3) $\mathbb{E}[X_{k+1} \mid \mathcal{F}(k)] = X_k$ for all k, X is saied to be a *martingale* with respect to $\mathcal{F}(k)$.

We call a sequence of events $\{E_n\}_{n \in \mathbb{N}}$ happens *with high probability* if $\Pr[E_n] = 1 - o(1)$ as n increases.

Theorem B.4 (Azuma Inequality). *Let $(W_k)_{0 \leq k \leq n}$ be a martingale with c_k such that $|W_{k+1} - W_k| \leq c_k$. Then,*

$$\Pr[W_n \geq W_0 + t] \leq \exp\left(-\frac{t^2}{2 \sum c_k^2}\right).$$

To handle rare bad event, the following theorem is quite useful, and can be proved by using union bound.

Theorem B.5 (Handling bad events)**.** *Let $(W_k)_{0\leq k\leq n}$ be a martingale which is bounded, $m \leq W_n \leq M$. Let $\mathcal{B}$ be a (bad) event such that there is a sequence c_k such that $|\mathbb{E}[W_T \mid \mathcal{F}_{k-1}, W_k, \neg\mathcal{B}] - \mathbb{E}[W_T \mid \mathcal{F}_{k-1}, W'_k, \neg\mathcal{B}]| \leq c_k$. Then,*

$$\Pr[W_n \geq W_0 + t + (M-m)\Pr[\mathcal{B}]] \leq \exp\left(-\frac{2t^2}{\sum c_k^2}\right) + \Pr[\mathcal{B}].$$

The following theorem shows this concentration property is dimension free.

Theorem B.6 (Vector-valued martingale [40, 34])**.** *Let $(\boldsymbol{W}_k)$ for $k = 1, \ldots, n$ be a vector-valued martingale with filtration $\mathcal{F}_k$ such that $\mathbb{E}[\boldsymbol{W}_{k+1} \mid \mathcal{F}_k] = \boldsymbol{W}_k$ for all $k \leq n$. If $\sup \|\boldsymbol{W}_{i+1} - \boldsymbol{W}_i\| \leq c_i$ for all i. Then,*

$$\mathbb{P}\left[\|\boldsymbol{W}_n - \mathbb{E}[\boldsymbol{W}_n]\| \geq t\right] \leq 20\exp\left(-\frac{t^2}{2\sum_i c_i^2}\right).$$

The following exponential inequality for maximum of martingales can save an extra union bound.

Theorem B.7 (Maximum tail [28, 26])**.** *Let $W_0, W_1, \ldots$ be a martingale with c_k and D such that $|W_{k+1} - W_k| \leq c_k$ and $\sup_k |W_{k+1} - W_k| \leq D$. Then, for any $t \geq 0$*

$$\Pr\left[\max_{k\leq n} W_k \geq W_0 + t\right] \leq \exp\left(-\frac{t^2}{2\sum c_k^2 + Dt}\right).$$

T is called a *stopping time* for $\mathcal{F}$ if and only if $\{T = k\} \in \mathcal{F}_k, \forall k$. Intuitively, this condition means that the "decision" of whether to stop at time k must be based only on the information present at time k, not on any future information.

Theorem B.8 (Optional Stopping theorem)**.** *If (W_k) is a martingale with respect to $(\mathcal{F}(k))$ and if T is a stopping time for $(\mathcal{F}(k))$ such that W_k is bounded, T is bounded, $\mathbb{E}[T] < \infty$, and $\mathbb{E}[|W_{k+1} - W_k| \mid \mathcal{F}(k)]$ is uniformly bounded, then*

$$\mathbb{E}[W_T] = \mathbb{E}[W_0].$$

C Proof for Theorem 4.1

Note that this proof is mostly identical to the one in Benaïm and Weibull [7].

Proof. Let $\boldsymbol{e}(k)$, for $k \in \mathbb{N}$, be the local error of $\boldsymbol{X}$ at time k:

$$\boldsymbol{e}(k+1) \triangleq n\left[\boldsymbol{X}(k+1) - \boldsymbol{X}(k)\right] - f(\boldsymbol{X}(k)) \tag{18}$$

Note that

$$\|\boldsymbol{e}(k+1)\| \leq \frac{1}{n}(B_f + \|\boldsymbol{U}(k+1)\|) \leq \frac{1}{n}(B_f + D),$$

so for all $\boldsymbol{\theta} \in \mathbb{R}^d$, by Cauchy inequality we have

$$\mathbb{E}[\exp(\langle\boldsymbol{\theta}, \boldsymbol{e}(k+1)\rangle)]) \mid \mathcal{F}(k)] \leq \exp\left(\frac{1}{n}\|\boldsymbol{\theta}\| \cdot (B_f + D)\right). \tag{19}$$

Now we extend the domain of $\boldsymbol{e}$ and $\boldsymbol{X}$ to $\mathbb{R}_{>0}$ and define $\bar{\boldsymbol{e}} : \mathbb{R}_{>0} \to \mathbb{R}^d$ and $\bar{\boldsymbol{X}} : \mathbb{R}_{>0} \to \mathbb{R}^d$ such that for $\kappa \in \mathbb{R}_{>0}$ $\bar{\boldsymbol{e}}(\kappa) = \boldsymbol{e}(\lfloor \kappa \rfloor)$ which is a right-continuous step function and define $\bar{\boldsymbol{X}}$ likewise. By (2) and (18), we have for all $k \in \mathbb{N}$

$$\boldsymbol{X}(k) - \boldsymbol{X}(0) = \sum_{l=0}^{k-1} (\boldsymbol{X}(l+1) - \boldsymbol{X}(l)) = \frac{1}{n} \sum_{l=0}^{k-1} (f(\boldsymbol{X}(l)) + \boldsymbol{e}(l)) = \frac{1}{n} \int_0^k \left(f(\bar{\boldsymbol{X}}(\ell)) + \boldsymbol{e}(\ell) \right) d\ell,$$

and $\kappa \in \mathbb{R}_{>0}$

$$\bar{\boldsymbol{X}}(\kappa) - \bar{\boldsymbol{X}}(0) = \frac{1}{n} \int_0^\kappa (f(\boldsymbol{X}(\ell)) + \boldsymbol{e}(\ell)) \, d\ell,$$

On the other hand, by the Equation (1)

$$\varphi(\boldsymbol{x}, \kappa/n) - \varphi(\boldsymbol{x}, 0) = \int_0^{\kappa/n} f(\varphi(\boldsymbol{x}, \ell)) \, d\ell = \frac{1}{n} \int_0^\kappa f\left(\varphi(\boldsymbol{x}, \ell/n)\right) \, d\ell.$$

Taking the difference between above equations, we have

$$\|\varphi(\boldsymbol{x}, \kappa/n) - \boldsymbol{X}(\kappa)\|_\infty = \frac{1}{n} \left\| \int_0^\kappa f(\varphi(\boldsymbol{x}, \ell/n)) - f(\boldsymbol{X}(\ell)) - \boldsymbol{e}(\ell) \, d\ell \right\|_\infty$$

Let $\Psi(C) \triangleq \frac{1}{n} \max_{k \le Cn} \left\| \sum_{i<k} \boldsymbol{e}(i) \right\|_\infty$. Hence

$$\begin{aligned}
\|\varphi(\boldsymbol{x}, \kappa/n) - \boldsymbol{X}(\kappa)\|_\infty \le& \frac{1}{n} \left\| \int_0^\kappa f(\varphi(\boldsymbol{x}, \ell/n)) - f(\boldsymbol{X}(\ell)) \, d\ell \right\|_\infty + \Psi(C) \\
\le& \frac{L_f}{n} \int_0^\kappa \|\varphi(\boldsymbol{x}, \ell/n) - \boldsymbol{X}(\ell)\|_\infty \, d\ell + \Psi(C) && (L_f\text{-Lipschitz}) \\
\le& \Psi(C) \exp\left(\frac{L_f \kappa}{n} \right). && \text{(by Gronwall's inequality)}
\end{aligned}$$

Therefore

$$\Pr\left[\max_{k \le Cn} \|\boldsymbol{X}(k) - \varphi(\boldsymbol{x}, k/n)\|_\infty > \epsilon \mid \boldsymbol{X}(0) = \boldsymbol{x} \right] \le \Pr\left[\Psi(C) > \epsilon \exp\left(-L_f C\right) \right], \tag{20}$$

and it suffices to upper bound the probability on the right-hand side. Let

$$Z_{\boldsymbol{\theta}}(0) = 1, \text{ and } Z_{\boldsymbol{\theta}}(k) \triangleq \exp\left(\frac{1}{n} \sum_{l=0}^{k-1} \langle \boldsymbol{\theta}, \boldsymbol{e}(l) \rangle - \frac{1}{n^2} k \|\boldsymbol{\theta}\|^2 (B_f + D) \right) \text{ when } k \ge 1$$

By Equation (19), $(Z_{\boldsymbol{\theta}}(k))_{k \in \mathbb{N}}$ is a non-negative supermartingale. Thus, for all $\varepsilon > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$

$$\begin{aligned}
&\Pr\left[\frac{1}{n} \max_{k \le nC} \left\langle \boldsymbol{\theta}, \sum_{l=0}^{k-1} \boldsymbol{e}(l) \right\rangle \ge \varepsilon \right] \\
\le& \Pr\left[\max_{k \le nC} Z_{\boldsymbol{\theta}}(k) \ge \exp\left(\varepsilon - \frac{1}{n^2} nC \|\boldsymbol{\theta}\|^2 (B_f + D) \right) \right] \\
\le& \frac{\mathbb{E}\left[Z_{\boldsymbol{\theta}}(0) \right]}{\exp\left(\varepsilon - \frac{1}{n} C \|\boldsymbol{\theta}\|^2 (B_f + D) \right)} && \text{(Doob's supermartingale inequality)} \\
=& \exp\left(\frac{1}{n} C \|\boldsymbol{\theta}\|^2 (B_f + D) - \varepsilon \right)
\end{aligned}$$

Now we are ready to bound $\Psi(C)$. Let $e_1, \dots, e_m$ be the canonical basis of $\mathbb{R}^d$. Because given a vector $\boldsymbol{v} \in \mathbb{R}^d$, its infinity norm is $\|\boldsymbol{v}\|_\infty = |\max_{i \le m} \boldsymbol{v}_i\| = \max_i \{\langle e_i, \boldsymbol{v}\rangle, -\langle e_i, \boldsymbol{v}\rangle\}$. Hence, we can take $\varepsilon = \frac{2\epsilon^2 n}{B_f + D}$, and set $\boldsymbol{\theta}$ be $\frac{2\epsilon n}{B_f+D} e_i$ for some i. Then

$$\Pr\left[\frac{1}{n}\max_{k \le nC}\left\langle e_i, \sum_{l=0}^{k-1} \boldsymbol{e}(l)\right\rangle \ge \epsilon\right] = \Pr\left[\frac{1}{n}\max_{k \le nC}\left\langle \boldsymbol{\theta}, \sum_{l=0}^{k-1} \boldsymbol{e}(l)\right\rangle \ge \varepsilon\right] \le \exp\left(\frac{-2\epsilon^2 n}{C(B_f + D)}\right),$$

Therefore taking union bound on all $\boldsymbol{\theta} = \frac{2\epsilon n}{B_f+D} e_i$ and $\frac{-2\epsilon n}{B_f+D} e_i$ for all i, and we have

$$\Pr\left[\Psi(C) \ge \epsilon\right] \le 2d \exp\left(\frac{-2\epsilon^2 n}{C(B_f + D)}\right).$$

By by inequality (20), we have

$$\Pr\left[\max_{k \le Cn} \|\boldsymbol{X}(k) - \varphi(\boldsymbol{x}, k/n)\|_\infty > \epsilon \mid \boldsymbol{X}(0) = \boldsymbol{x}\right] \le 2d \exp\left(-\frac{2\exp\left(-2L_f C\right)}{C(B_f + D)} \cdot \epsilon^2 n\right)$$

which completes the proof. □

D Proofs for Sect. 4.2

Proof of Lemma 4.4. This is proved by using the optional stopping time theorem. Given $\boldsymbol{X}(0) \in B(\boldsymbol{0}, r_{\text{reg}})$, let T_0 be the stopping time such that $\|\boldsymbol{X}^u(T_0)\| \ge 8\|\boldsymbol{X}^s(T_0)\|$ or $\|\boldsymbol{X}(T_0)\| \le l_1 \sin\theta_{us}$. We consider the following random variables $W^s(k) \triangleq \left(1 - \frac{\mu_s}{2n}\right)^{-k} \|\boldsymbol{X}^s(k)\|^2$. Suppose $W^s(k)$ is a super martingale and $r_{\text{reg}} < \sin\theta_{us}$, and by optional stopping time theorem B.8 and (5),

$$\mathbb{E}[W^s(T_0)] \le W^s(0) = \|\boldsymbol{X}^s(0)\|^2 \le \left(\frac{r_{\text{reg}}}{\sin\theta_{us}}\right)^2 \le 1. \tag{21}$$

Let $\tau_0 = (2n\log n)/\mu_s$ Therefore we can upper bound $\Pr[T_0 > \tau_0]$ as follows:

$$\begin{aligned}
\mathbb{E}[W^s(T_0)] =& \mathbb{E}\left[\left(1 - \frac{\mu_s}{2n}\right)^{-T_0} \|\boldsymbol{X}^s(T_0)\|^2\right] \\
\ge& \left(1 - \frac{\mu_s}{2n}\right)^{-\tau_0} \mathbb{E}[\|\boldsymbol{X}^s(T_0)\| \mid T_0 > \tau_0] \Pr[T_0 > \tau_0] \\
\ge& n l_1^2 \Pr[T_0 > \tau_0] && (\ \|\boldsymbol{X}(T_0)\| \ge l_1 \sin\theta_{us}) \\
\ge& \frac{(\log n)^{2/3}}{4} \Pr[T_0 > \tau_0]
\end{aligned}$$

Therefore combining the equation (21) and the above, we have

$$\Pr[T_0 \le (2n\log n)/\mu_s] = o(1).$$

Now, let's use induction to show $W^s(k)$ is a supermartingale before the stopping time T_0. Let $\boldsymbol{D}^s(k) \triangleq n\left(\boldsymbol{X}^s(k+1) - \boldsymbol{X}^s(k)\right)$, and

$$\left(1 - \frac{\mu_s}{2n}\right)^{k+1} \mathbb{E}[W^s(k+1) \mid \mathcal{F}(k)] = \mathbb{E}\left[\left\|\boldsymbol{X}^s(k) + \frac{1}{n}\left(\boldsymbol{D}^s(k)\right)\right\|^2 \mid \mathcal{F}(k)\right].$$

Let $L(r) = \max_{\boldsymbol{x}\in B(\boldsymbol{0},r)} \|A\boldsymbol{x}\| + \max_{\boldsymbol{x}\in B(\boldsymbol{0},r)} \|R(\boldsymbol{x})\| + \max \|\boldsymbol{U}\|$ which is a constant depends on r. We can translate the 2 norm into inner product, and have

$$\begin{aligned}
&\left(1-\frac{\mu_s}{2n}\right)^{k+1} \mathbb{E}[W^s(k+1) \mid \mathcal{F}(k)] \\
\le & \|\boldsymbol{X}^s(k)\|^2 + \frac{1}{n}\mathbb{E}\left[\langle \boldsymbol{X}^s(k), \boldsymbol{D}^s(k)\rangle \mid \mathcal{F}(k)\right] + \frac{L(r_{\text{reg}})}{n^2} && \text{(by (7))} \\
\le & \|\boldsymbol{X}^s(k)\|^2 + \frac{1}{n}\left(\boldsymbol{X}^s(k)^\top A \boldsymbol{X}^s(k) + \frac{H}{\sin\theta_{us}}\|\boldsymbol{X}(k)\|^3\right) + \frac{L(r_{\text{reg}})}{n^2} && \text{(by (6) and (5))} \\
\le & \left(1-\frac{\mu_s}{n}\right)\|\boldsymbol{X}^s(k)\|^2 + \frac{H}{n\sin\theta_{us}}\|\boldsymbol{X}(k)\|^3 + \frac{L(r_{\text{reg}})}{n^2} && \text{(by Corollary B.2)}
\end{aligned}$$

Because $\boldsymbol{X}(k)\| \ge l_1 \sin\theta_{us}/2 = \omega(1/\sqrt{n})$, if r_{reg} is small enough and n large enough, we have $H\|\boldsymbol{X}(k)\|^3/\sin\theta_{us} + L(r_{\text{reg}})/n \le \mu_s\|\boldsymbol{X}^s(k)\|^2/2$, and

$$\left(1-\frac{\mu_s}{2n}\right)^{k+1} \mathbb{E}[W^s(k+1) \mid \mathcal{F}(k)] \le \left(1-\frac{\mu_s}{2n}\right)\|\boldsymbol{X}^s(k)\|^2 = \left(1-\frac{\mu_s}{2n}\right)^{k+1} W^s(k).$$

This completes the proof. □

Proof of Lemma 4.5. Note that if the process is in the phase 1 then $\|\boldsymbol{X}(0)\| \le l_1/\sin\theta_{us}$.

Set $\tau_1 = n\log n)$. Let T_1 be the stopping time that $\|\boldsymbol{X}(T_1)\| \ge 4l_2/\sin\theta_{us}$. We first show the expectation of T_1 is much smaller than τ_1. Then we show the stable component $\|\boldsymbol{X}^s(k)\|$ is smaller than l_2 for all $k \le \tau_1$. By union bound on these two event, we show with high probability there exists $T_1 \le \tau_1$ such that $\|\boldsymbol{X}^u(T_1)\|$ is larger than $3l_2$ and $\|\boldsymbol{X}^s(T_1)\|$ is smaller than l_2. This completes the proof.

By (5), to lower bound $\|\boldsymbol{X}\|$ it is sufficient to lower bound the magnitude of unstable component, $\|\boldsymbol{X}^u\|$. Let $a^u_{\text{noise}} \triangleq \frac{\alpha}{2d}\operatorname{Tr}((P^u)^\top P^u) > 0$ and $W^u(k) \triangleq \|\boldsymbol{X}^u(k)\|^2 - a^u_{\text{noise}}k/n^2$. If $W^u(k)$ is a submartingale, by optional stopping theorem (Theorem B.8) $\mathbb{E}[W^u(T_1) \mid \mathcal{F}(0)] \ge \mathbb{E}[W^u(0)] \ge 0$ and

$$\mathbb{E}[\|\boldsymbol{X}^u(T_1)\|^2] \ge a^u_{\text{noise}}\frac{\mathbb{E}[T_1]}{n^2}. \tag{22}$$

Therefore by (5) and (22),

$$\mathbb{E}[T_1] \le \frac{n^2}{a^u_{\text{noise}}}\mathbb{E}[\|\boldsymbol{X}^u(T_1)\|^2] \le \frac{(n\sin\theta_{us})^2}{a^u_{\text{noise}}}\mathbb{E}[\|\boldsymbol{X}(T_1)\|^2] \le \frac{(4nl_2)^2}{a^u_{\text{noise}}} = O(n\log^{2/3} n).$$

By Markov inequality $\Pr[T_1 \le \tau_1] = 1 - 1/(\log n)^{1/3} = 1 - o(1)$.

Now, let's show $W^u(k)$ is a submartingale with respect to $\mathcal{F}(k)$ before stopping time T_1. Let $\boldsymbol{D}^u(k) \triangleq \boldsymbol{X}^u(k+1) - \boldsymbol{X}^u(k)$.

$$\begin{aligned}
\mathbb{E}[W^u(k+1) \mid \mathcal{F}(k)] =& \mathbb{E}\left[\|\boldsymbol{X}^u(k+1)\|^2 - \frac{a^u_{\text{noise}}}{n^2}(k+1) \mid \mathcal{F}(k)\right] \\
=& \mathbb{E}\left[\langle \boldsymbol{X}^u(k) + \boldsymbol{D}^u(k), \boldsymbol{X}^u(k) + \boldsymbol{D}^u(k)\rangle \mid \mathcal{F}(k)\right] - \frac{a^u_{\text{noise}}}{n^2}(k+1) \\
=& W^u(k) + 2\mathbb{E}\left[\langle \boldsymbol{X}^u(k), \boldsymbol{D}^u(k)\rangle \mid \mathcal{F}(k)\right] + \mathbb{E}\left[\langle \boldsymbol{D}^u(k), \boldsymbol{D}^u(k)\rangle \mid \mathcal{F}(k)\right] - \frac{a^u_{\text{noise}}}{n^2}.
\end{aligned}$$

Thus, it is sufficient to show the following two claims:

$$2\mathbb{E}\left[\langle \boldsymbol{X}^u(k), \boldsymbol{D}^u(k)\rangle | \mathcal{F}(k)\right] + \frac{a^u_{\text{noise}}}{n^2} \ge 0 \tag{23}$$

$$\mathbb{E}\left[\langle \boldsymbol{D}^u(k), \boldsymbol{D}^u(k)\rangle | \mathcal{F}(k)\right] - \frac{2a^u_{\text{noise}}}{n^2} \ge 0 \tag{24}$$

For (23), we need to use the fact that A is expanding is subspace of $\mathbb{E}^u$ before stopping time,

$$\begin{aligned}
&2\mathbb{E}\left[\langle \boldsymbol{X}^u(k), \boldsymbol{D}^u(k)\rangle | \mathcal{F}(k)\right] + \frac{a^u_{\text{noise}}}{n^2} \\
=&\frac{2}{n}\langle \boldsymbol{X}^u(k), A\boldsymbol{X}^u(k) + R^u(\boldsymbol{X}(k))\rangle + \frac{a^u_{\text{noise}}}{n^2} \\
\geq&\frac{2}{n}\left(\mu_u\|\boldsymbol{X}^u(k)\|^2 - \|\boldsymbol{X}(k)\|^3/\sin\theta_{us}\right) + \frac{a^u_{\text{noise}}}{n^2} && \text{(by Corollary B.2)} \\
>&\frac{1}{n}O(\|\boldsymbol{X}(k)\|^3) + \frac{a^u_{\text{noise}}}{n^2} \geq 0. && (\ \|\boldsymbol{X}(k)\| = O(l_1) \text{ and } n \text{ large enough})
\end{aligned}$$

For (23), we use the variance of $\boldsymbol{U}$ is bounded below by some constant

$$\begin{aligned}
&\mathbb{E}\left[\langle \boldsymbol{D}^u(k), \boldsymbol{D}^u(k)\rangle | \mathcal{F}(k)\right] \\
=&\frac{1}{n^2}\|A\boldsymbol{X}^u(k) + R^u(\boldsymbol{X}(k))\|^2 + \frac{1}{n^2}\mathbb{E}\left[\|\boldsymbol{U}^u(k)\|^2\right] \\
\geq&\frac{1}{n^2}\mathbb{E}\left[\|\boldsymbol{U}^u(k)\|^2\right] \\
\geq&\frac{1}{n^2}\frac{\alpha}{d}\operatorname{Tr}((P^u)^\top P^u) = \frac{2}{n^2}a^u_{\text{noise}} && \text{(by Lemma D.1 and definition of } a^u_{\text{noise}})
\end{aligned}$$

For the second part, we can use similar argument in Lemma 4.4 and union bound to show it's true with high probability. □

Lemma D.1 (projected noise)**.** *Given matrices $P, S \in \mathbb{R}^{d\times d}$, constants $0 < \alpha$, and a d-dimensional random vector $\boldsymbol{X} \in \mathbb{R}^d$, if P is not the zero matrix, S is positive definite matrix with $\frac{\alpha}{d}\mathbb{I}_d \prec S$, and $\mathbb{E}[\boldsymbol{X}] = 0$, $\operatorname{Cov}[\boldsymbol{X}] = S$,*

$$0 < \frac{\alpha}{d}\operatorname{Tr}(P^\top P) < \mathbb{E}\left[\|P\boldsymbol{X}\|^2\right].$$

Proof. First we observe that

$$\begin{aligned}
\mathbb{E}\left[\|P\boldsymbol{X}\|^2\right] =&\mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{X}^\top P^\top P\boldsymbol{X}\right)\right] \\
=&\mathbb{E}\left[\operatorname{Tr}\left(P^\top P\boldsymbol{X}\boldsymbol{X}^\top\right)\right] \\
=&\operatorname{Tr}\left(P^\top P\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right]\right) && \text{(linearity of trace)} \\
=&\operatorname{Tr}\left(P^\top PS\right) > 0
\end{aligned}$$

The last ineqality holds because S is positive definite and $P^\top P$ is positive semi-definite and not the zero matrix. With this inequality we have

$$\mathbb{E}\left[\|P\boldsymbol{X}\|^2\right] - \frac{\alpha}{d}\operatorname{Tr}(P^\top P) = \operatorname{Tr}\left(P^\top PS\right) - \frac{\alpha}{d}\operatorname{Tr}(P^\top P) = \operatorname{Tr}\left(P^\top P\left(S - \frac{\alpha}{d}\mathbb{I}_d\right)\right) > 0.$$

The last one is true, since $S - \frac{\alpha}{d}\mathbb{I}_d$ is positive definite. □

Proof of Lemma 4.6. Let $\tau_j = C_j n$ for some C_j and r_{reg} small enough such that $B(\boldsymbol{0}, \sqrt{r_{\text{reg}}}) \in E$. Let T_j be the exit time, $T_j = T_e\left(B(\boldsymbol{0}, \sqrt{r_{\text{reg}}})\right)$ given $\boldsymbol{X}(0) = \boldsymbol{x}$ in the phase j defined in the statement. Here we abuse the notation and define $\boldsymbol{X}(k)$ as a new process by Equation (7) and couple it with the original process until T_j. Therefore, the lemma can be proved with the following are three equations:

1. With very high probability the stopping time T_j is greater than τ_j ,
$$\Pr[T_j > \tau_j] = 1 - o(1/\log n); \tag{25}$$
2. The expectation at time τ_j, $\mathbb{E}[\boldsymbol{X}(\tau_j)]$, is nice,
$$\|\mathbb{E}[\boldsymbol{X}^s(\tau_j)]\| \leq l_j/16 \text{ and } 4l_j \leq \|\mathbb{E}[\boldsymbol{X}^u(\tau_j)]\|; and \tag{26}$$
3. $\boldsymbol{X}(\tau_j)$ is concentrated
$$\Pr\left[l_{j+1} \geq 8\|\boldsymbol{X}^s(\tau_j)\| \text{ and } \|\boldsymbol{X}^u(\tau_j)\| > l_{j+1}\right] = 1 - o(1/\log n). \tag{27}$$

Before proving these, let's do some computation to gain some intuition. To compute the $\mathbb{E}[\boldsymbol{X}(\tau_j)]$ suppose $T_j > \tau_j$ we can use the linear function $A\boldsymbol{X}(k)$ to approximate $f(\boldsymbol{X}(k))$ for all $k \leq \tau_j$ and tower property of expectation:

$$\begin{aligned}\mathbb{E}[\boldsymbol{X}(k+1)] =&\mathbb{E}[\mathbb{E}[\boldsymbol{X}(k+1) \mid \mathcal{F}(k)]]\\ =&\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{X}(k) + \frac{1}{n}\left(A\boldsymbol{X}(k) + R(\boldsymbol{X}(k)) + \boldsymbol{U}(k)\right) \mid \mathcal{F}(k)\right]\right] && \text{(by Equation (7))}\\ =&\left(1 + \frac{1}{n}A\right)\mathbb{E}\left[\boldsymbol{X}(k)\right] + \frac{1}{n}\left[R(\boldsymbol{X}(k))\right].\end{aligned}$$

For each k, let
$$e(k) \triangleq \left\|\mathbb{E}[\boldsymbol{X}(k)] - \left(1 + \frac{1}{n}A\right)^k \boldsymbol{X}(0)\right\|$$
which is the 2-norm error between $\mathbb{E}[\boldsymbol{X}(k)]$ and $\left(1 + \frac{1}{n}A\right)^k \boldsymbol{X}(0)$ By triangle inequality, $e(k+1) \leq \left(1 + \frac{\sigma_{\max}(A)}{n}\right) e(k) + \frac{1}{n}\|\mathbb{E}[R(\boldsymbol{X}(k))]\|$ where $\sigma_{\max}(A)$ is the induced norm of A. Additionally by (6), $\|\mathbb{E}[R(\boldsymbol{X}(k))]\| \leq \mathbb{E}[\|R(\boldsymbol{X}(k))\|] \leq H \cdot \|\boldsymbol{X}(k)\|^2$ for all $\boldsymbol{X}(k) \in B(\mathbf{0}, \sqrt{r_{\text{reg}}}) \subseteq E$.

Therefore with Gronwall theorem, we can bound $e(k)$ as
$$e(\tau_j) \leq \sigma_{\max}(A)^{-1}H \cdot \max_k \|\boldsymbol{X}(k)\|^2 \exp\left(\tau_j/n\right) - 1 \leq \sigma_{\max}(A)^{-1}He^{C_j} \cdot \max_k \|\boldsymbol{X}(k)\|^2 \tag{28}$$

Therefore, suppose the norm $\|\boldsymbol{X}(k)\|^2$ are small for all $0 \leq k < \tau_j$, the value $\mathbb{E}[\boldsymbol{X}(\tau_j)]$ can be approximated by the linear term, $\left(1 + \frac{1}{n}A\right)^{\tau_j} \boldsymbol{X}(0)$. Specifically, we want to show for all constant $\epsilon > 0$,
$$\max_{0 \leq k < \tau_j} \|\boldsymbol{X}(k)\|^2 \leq \epsilon\|\boldsymbol{X}(0)\| \tag{29}$$
which implies
$$\left\|\mathbb{E}[\boldsymbol{X}(k)] - \left(1 + \frac{1}{n}A\right)^k \boldsymbol{X}(0)\right\| \leq \sigma_{\max}(A)^{-1}He^{C_j}\epsilon. \tag{30}$$

Equation (25)**:** We define $W_k \triangleq \left(1 + \frac{2\mu_{\max}}{n}\right)^{-k} \|\boldsymbol{X}(k)\|^2$ where $\mu_{\max}$ is the maximum real part of eigenvalues of A. By Corollary B.2 and similar argument in Lemma 4.4, W_k is a supermartingale such that $\mathbb{E}[W_{k+1} \mid \mathcal{F}(k)] \leq W_k$.

Let's apply Theorem B.7 on (W_k). Because for all $k \le C_j n$ $|\|\boldsymbol{X}(k+1)\|^2 - \|\boldsymbol{X}(k)\|^2| = O\left(\frac{1}{n}\right)$ uniformly. By Theorem B.7, we let $D = O(1/n)$, $c_k = |W_{k+1} - W_k| = O\left(\left(1 + \frac{2\mu_{\max}}{n}\right)^{-k} \frac{1}{n}\right)$, $\sum c_i^2 = O\left(\frac{1}{n}\right)$, and $\delta = \frac{(\log n)^{1/4}}{\sqrt{n}}$, so we have

$$\Pr\left[\max_{k \le \tau_j} W_k \ge W_0 + \delta\right] \le \exp\left(-\frac{\delta^2}{2\sum_{k\le\tau_j} c_k^2 + D\delta}\right) = \exp\left(-\Omega\left(\sqrt{\log n}\right)\right).$$

Let $\mathcal{E}$ be the good event that $\max_{k\le\tau_j} W_k < W_0 + \delta$. Note that condition on $\mathcal{E}$, with probability $\Pr[\mathcal{E}] = 1 - \exp\left(-\Omega\left(\sqrt{\log n}\right)\right)$ we have Equation (29) for all $0 \le k \le \tau_j$

$$\|\boldsymbol{X}(k)\|^2 \le \left(1 + \frac{2\mu_{\max}(A)}{n}\right)^k \left(\|\boldsymbol{X}(0)\|^2 + \delta\right) \le 2\|\boldsymbol{X}(0)\| \exp\left(2\sigma_{\max}(A) C_j\right) \cdot \|\boldsymbol{X}(0)\|. \tag{31}$$

Given $\epsilon, C_j, \mu_{\max}(A) > 0$, we can take r_{reg} small enough such that $\|\boldsymbol{X}(0)\| \le r_{\text{reg}}/\sin\theta_{us}$ is small and proves Equation (25).

Equation (26) **and** (30)**:** Now we are ready to prove the first part. By Equation (28) and (31), let $\mathcal{E}$ be the event defined in (29) we have for arbitrary small $\varepsilon > 0$ as n large enough and ϵ small enough

$$\begin{aligned}
&\left\|\mathbb{E}[\boldsymbol{X}(\tau_j)] - \left(1 + \frac{1}{n}A\right)^{\tau_j} \boldsymbol{X}(0)\right\| \\
\le& \frac{He^{C_j}}{\sigma_{\max}(A)} \max \|\boldsymbol{X}(k)\| && \text{(by (28))} \\
\le& \frac{H\epsilon e^{C_j}}{\sigma_{\max}(A)} \|\boldsymbol{X}(0)\| \Pr[\mathcal{E}] + O(\Pr[\neg\mathcal{E}]) && (\boldsymbol{X}(\tau_j) \in E \text{ which is bounded}) \\
\le& \varepsilon l_j. && \text{(by (29))}
\end{aligned}$$

This proves inequality (30). As a result, for the unstable component and stable component we have

$$\begin{aligned}
\|\mathbb{E}[\boldsymbol{X}^u(\tau_j)]\| &\ge \left(1 + \frac{\mu_u}{n}\right)^{\tau_j} \|\boldsymbol{X}^u(0)\| - \frac{\varepsilon}{\sin\theta_{us}} l_j \ge \left(e^{\mu_u C_j} + \frac{\varepsilon}{\sin\theta_{us}}\right) l_j \\
\|\mathbb{E}[\boldsymbol{X}^s(\tau_j)]\| &\le \left(1 - \frac{\mu_s}{n}\right)^{\tau_j} \|\boldsymbol{X}^s(0)\| + \frac{\varepsilon}{\sin\theta_{us}} l_j \le \left(\frac{e^{-\mu_s C_j}}{8} + \frac{\varepsilon}{\sin\theta_{us}}\right) l_j.
\end{aligned}$$

This proves Equation (26) by taking C_j large enough and ε small enough.

Equation (27)**:** We define a vector-valued Doob martingale,

$$\boldsymbol{Y}_k = \boldsymbol{Y}_k(\boldsymbol{X}(0), \ldots, \boldsymbol{X}(k)) = \mathbb{E}[\boldsymbol{X}(\tau_j) | \boldsymbol{X}(0), \ldots, \boldsymbol{X}(k)] \in \mathbb{R}^d. \tag{32}$$

and prove Equation (27) by using concentration property of vector-valued martingale $\boldsymbol{Y}_k$ (Theorem B.6 and B.5). With good event $\mathcal{E}$, we want to bound $\{c_k\}_{0\le k\le\tau_j}$ the "variability" of each variable $\boldsymbol{X}(0), \ldots, \boldsymbol{X}(\tau_j)$ on the martingale Y_k condition on this good event defined in (32),

$$c_k = \sup\left\{\left\|\mathbb{E}[\boldsymbol{X}(\tau_j) | \mathcal{F}(k-1), \boldsymbol{X}(k) = \boldsymbol{x}, \mathcal{E}] - \mathbb{E}[\boldsymbol{X}(\tau_j) | \mathcal{F}(k-1), \boldsymbol{X}(k) = \boldsymbol{x}', \mathcal{E}]\right\|\right\}.$$

Equivalently, c_k is the 2-norm error with initial difference at step k be $\|\boldsymbol{x} - \boldsymbol{x}'\| = O(1/n)$. Formally by (26) and $\mathcal{E}$, we have $c_k = O(1/n)$ for all $k \le \tau_j$ and $\sum_{k=0}^{\tau_j} c_k^2 = O(1/n)$. By concentration property of vector-valued martingale Y_k (Theorem B.6), for any constant $D' > 0$

$$\Pr\left[\|\boldsymbol{X}(\tau_j) - \mathbb{E}[\boldsymbol{X}(\tau_j)]\| \ge \frac{l_j}{16D'}\right] \le O\left(\exp\left(-\Omega(nl_j^2)\right)\right) + \Pr[\neg\mathcal{E}] = \exp\left(-\Omega\left(\sqrt{\log n}\right)\right) \tag{33}$$

Therefore, by Equations (26), and (33), with probability $1 - \exp\left(-\Omega\left(\sqrt{\log n}\right)\right) = 1 - o(1/\log n)$ we have,

$$\|\boldsymbol{X}^u(\tau_j)\| \geq \|\mathbb{E}[\boldsymbol{X}^u(\tau_j)]\| - \frac{l_j}{16D'} \geq \left(4 - \frac{1}{16D'}\right) l_j \geq 2l_j = l_{j+1}.$$

This the last inequality can be true by first take D' large, C large. The stable component can be upper bounded as follows

$$\|\boldsymbol{X}^s(\tau_j)\| \leq \|\mathbb{E}[\boldsymbol{X}^s(\tau_j)]\| + \frac{l_j}{16D'} \leq \left(\frac{1}{2} + \frac{1}{2D'}\right) \frac{l_j}{8} \leq \frac{1}{8} l_j \leq \frac{1}{8} l_{j+1}.$$

which proves Equation (27). □

E Proofs in Section 4.3

Proof for Proposition 4.7. The proof is basically identical to Lemma 4.4. This is proved by using optional stopping time theorem. Given $\boldsymbol{X}(0) \in B(\mathbf{0}, \frac{b}{n} r_{\text{reg}})$, let $T_a \triangleq T_h\left(B\left(\mathbf{0}, \frac{b}{n}\right)\right)$ be the hitting time to the set $B\left(\mathbf{0}, \frac{b}{n}\right)$. Since $\mathbf{0}$ is an attracting fixed point, there is $\mu_u > 0$. We consider the following random variables $W(k) \triangleq \left(1 - \frac{\mu_s}{2n}\right)^{-k} \|\boldsymbol{X}(k)\|^2$. Suppose $W(k)$ is a super martingale and $r_{\text{reg}} < 1$. By optional stopping time theorem B.8, $\mathbb{E}[W(T_a)] \leq W(0) \leq 1$.

Let $\tau_a = (6n \log n)/\mu_s$ Therefore we can upper bound $\Pr[T_a > \tau_a]$ as follows:

$$\begin{aligned}
\mathbb{E}[W(T_a)] =& \mathbb{E}\left[\left(1 - \frac{\mu_s}{2n}\right)^{-T_a} \|\boldsymbol{X}(T_a)\|^2\right] \\
\geq& \left(1 - \frac{\mu_s}{2n}\right)^{-\tau_0} \mathbb{E}[\|\boldsymbol{X}^s(T_0)\| \mid T_0 > \tau_0] \Pr[T_0 > \tau_0] \\
\geq& n^3 \left(\frac{b}{n}\right)^2 \Pr[T_a > \tau_a]
\end{aligned}$$

Therefore combining these two inequalities, we have

$$\Pr[T_0 \leq (6n \log n)/\mu_s] = o(1/n).$$

Now, let's use induction to show $W(k)$ is a supermartingale before the stopping time T_a:

$$\left(1 - \frac{\mu_s}{2n}\right)^{k+1} \mathbb{E}[W(k+1) \mid \mathcal{F}(k)] = \mathbb{E}\left[\left\|\boldsymbol{X}(k) + \frac{1}{n}\left(A\boldsymbol{X}(k) + R(\boldsymbol{X}(k)) + \boldsymbol{U}(k+1)\right)\right\|^2 \mid \mathcal{F}(k)\right].$$

Let $L(r) = \max_{\boldsymbol{x} \in B(\mathbf{0}, r)} \|A\boldsymbol{x}\| + \max_{\boldsymbol{x} \in B(\mathbf{0}, r)} \|R(\boldsymbol{x})\| + \max \|\boldsymbol{U}\|$ which is a constant depends on r. We can translate the 2 norm into inner product, and have

$$\begin{aligned}
&\left(1 - \frac{\mu_s}{2n}\right)^{k+1} \mathbb{E}[W(k+1) \mid \mathcal{F}(k)] \\
\leq& \|\boldsymbol{X}(k)\|^2 + \frac{1}{n}\mathbb{E}\left[\langle \boldsymbol{X}(k), A\boldsymbol{X}(k) + R(\boldsymbol{X}(k)) + \boldsymbol{U}(k+1)\rangle \mid \mathcal{F}(k)\right] + \frac{L(r_{\text{reg}})}{n^2} && \text{(by (7))} \\
\leq& \left(1 - \frac{\mu_s}{n}\right) \|\boldsymbol{X}(k)\|^2 + H\|\boldsymbol{X}(k)\|^3 + \frac{L(r_{\text{reg}})}{n^2}
\end{aligned}$$

Because for all $k < T_a$, $\boldsymbol{X}(k)\| \geq b/n$, if r_{reg} is small enough, b is large enough, and n large enough, we have $H\|\boldsymbol{X}(k)\|^3 + L(r_{\text{reg}})/n \leq \mu_s \|\boldsymbol{X}(k)\|^2/2$, and

$$\left(1 - \frac{\mu_s}{2n}\right)^{k+1} \mathbb{E}[W(k+1) \mid \mathcal{F}(k)] \leq \left(1 - \frac{\mu_s}{2n}\right) \|\boldsymbol{X}(k)\|^2 = \left(1 - \frac{\mu_s}{2n}\right)^{k+1} W(k).$$

This completes the proof. □

Proof of Proposition 4.8. The proof is quite straightforward. To the end, we want to show there exists a constant $r_{\text{reg}} > 0$ such that it is very hard to escape $Q_o \triangleq B(\mathbf{0}, 2r_{\text{reg}})$ from $Q_i \triangleq B(\mathbf{0}, r_{\text{reg}})$. The proof follows with the following three observations: 1) Because the step size is bounded by $O(1/n)$ it takes at least cn for some constant c to escape. 2) There is a (local) potential function that decreases by $\Theta(1/n)$ in each step in $Q_o \setminus Q_i$. 3) Combining the above two, to escape Q_o there is a time interval with length at least cn such that the deviation from expectation is $\Omega(1)$ which finishes the proof by Azuma's inequality.

The first part is trivial because $f \in \mathcal{C}^1$ and Q_o is bounded.

For the second part, because $\mathbf{0}$ is an attracting fixed point, all the eigenvalues of $A \triangleq \nabla f(\mathbf{0})$ have negative real part, and it is called a stable matrix (or sometimes Hurwitz matrix), and by Lyapunov theorem there exists a positive definite matrix P such that $PA + A^\top P = -\mathbb{I}_d$. We define $V(\boldsymbol{x}) \triangleq \boldsymbol{x}^\top P \boldsymbol{x}$. Let $L(r) = \sigma_{\max}(A)(H + \max_{\boldsymbol{x} \in B(\mathbf{0},r)} \|A\boldsymbol{x}\| + \max_{\boldsymbol{x} \in B(\mathbf{0},r)} \|R(\boldsymbol{x})\| + D)$ which is a constant depending r.

Let r_{reg} be a positive constant which will be specified later. For all $\boldsymbol{X}(k) \in Q_o$ we have

$$\begin{aligned}\mathbb{E}[V(\boldsymbol{X}(k+1)) \mid \mathcal{F}(k)] = \leq V(\boldsymbol{X}(k)) + \frac{1}{n}\boldsymbol{X}(k)^\top(PA + A^\top P)\boldsymbol{X}(k) + \frac{L(4r_{\text{reg}})}{n}\|\boldsymbol{X}(k)\|^3 \\ \leq V(\boldsymbol{X}(k)) - \frac{1}{n}\|\boldsymbol{X}(k)\|^2 + \frac{L(4r_{\text{reg}})}{n}\|\boldsymbol{X}(k)\|^3 \qquad (PA + A^\top P = -\mathbb{I}_d)\end{aligned}$$

Therefore the value $V(X_k)$ is a super martingale and there exists $r_{\text{reg}} > 0$ such that

$$\mathbb{E}[V(\boldsymbol{X}(k+1)) \mid \mathcal{F}(k)] - V(\boldsymbol{X}(k)) \leq -r_{\text{reg}}/(2n)$$

for all $\boldsymbol{X}(k) \in Q_o \setminus Q_i$. Furthermore, because P is positive definite the potential value has constant separation: there exists a constant $h > 0$ such that

$$h < \min\{V(\boldsymbol{x}) : \boldsymbol{x} \notin Q_o\} - \max\{V(\boldsymbol{x}) : \boldsymbol{x} \in Q_i\}.$$

Finally, suppose there exists $0 \leq l \leq T$ is the exit time such that $\boldsymbol{X}(l) \notin Q_o$. Because $\boldsymbol{X}(0) \in Q_i$, there exists an interval of time from k to l such that $\boldsymbol{X}(k) \in Q_i$, $X_l \notin Q_o$ and $X_\ell \in Q_o \setminus Q_i$ for all $k < \ell < l$, we define this event as E_l which happens with probability

$$\Pr[\boldsymbol{X}(l) \notin Q_o \mid \boldsymbol{X}(0) \in Q_i] \leq \Pr[E_l] \leq \exp(-\Omega(n))$$

by Azuma's inequality. The proof is finished by taking union bound on $l \leq T$. □

F Proofs for Section 5

Proof of Lemma 5.2. Since (1) is a gradient-like system and V is a complete Lyapunov function, for all $\boldsymbol{x}$ we know there exists a fixed point $\boldsymbol{\beta}_i \in \text{Fix}_f$ such that $\liminf_{t\to\infty} \|\varphi(\boldsymbol{x}, t) - \boldsymbol{\beta}_i\| = 0$ and $V(\boldsymbol{\beta}_i) \leq V(\boldsymbol{x})$. Therefore, given $r_i > 0$ a neighborhood of $\boldsymbol{\beta}_i$, $B(\boldsymbol{\beta}_i, r_i)$, there is a constant t such that $\varphi(\boldsymbol{x}, t) \in B(\boldsymbol{\beta}_i, r_i)$.

Moreover by Theorem 4.1, $\boldsymbol{X}$ converges to $B(\boldsymbol{\beta}_i, 2r_i)$ in $O(n)$ steps with high probability, and we finish the proof. □

The proof of Lemma 5.3 has two parts: we first show the process is constant away from the fixed point $\boldsymbol{\beta}_i$ within time $T_1 = O(n \log n)$ with high probability in Theorem 4.3, and we use the property of complete Lyapunov function, and show the value of $V(X_{T_1})$ is not much bigger than

$V(\boldsymbol{\beta}_i)$. In the second part, we run the process for extra $T_2 = O(n)$ steps. Because the process is far from fixed point, the decrease rate of V is large and $V(X_{T_1+T_2})$ is constantly smaller than $V(\boldsymbol{\beta}_i)$.

To define this two parts formally, We first define several neighborhoods of $\boldsymbol{\beta}_i$: $N_i \subset B(\boldsymbol{\beta}_i, r/2) \subset B(\boldsymbol{\beta}_i, 3r/4) \subset B(\boldsymbol{\beta}_i, r)$ where $B(\boldsymbol{\beta}_i, r)$ is the open ball with radius $\boldsymbol{\beta}_i$ and centered at $\boldsymbol{\beta}_i$. Lemma 5.3 keeps track of the process when it enter the region N_i and stop after leaving $B(\boldsymbol{\beta}_i, r)$. Taking r small enough such that $\bar{B}(\boldsymbol{\beta}_i, r)$ only has a single fixed point $\boldsymbol{\beta}_i$. Because the complete Lyapunov function $V \in \mathcal{C}^1$ and $\mathcal{L}_f V(\boldsymbol{x}) < 0$ for all $x \in \bar{B}(\boldsymbol{\beta}_i, r) \setminus B(\boldsymbol{\beta}_i, r/2)$ which is a compact set, there exists $\kappa > 0$ such that

$$\forall x \in \bar{B}(\boldsymbol{\beta}_i, r) \setminus B(\boldsymbol{\beta}_i, r/2),\ \mathcal{L}_f V(\boldsymbol{x}) < -\kappa. \tag{34}$$

Fixing r with κ, because f is smooth, there exists D' such that $D' = \max \|f(x)\| + D$ for all $x \in \bar{B}(\boldsymbol{\beta}_i, r)$ which is an upper bound for the movement of the process in $\bar{B}(\boldsymbol{\beta}_i, r)$. Finally we can take N_i small enough such that

$$\forall \boldsymbol{x} \in N_i,\ \|V(\boldsymbol{x}) - V(\boldsymbol{\beta}_i)\| \le \frac{\kappa r}{32 D'}. \tag{35}$$

Proof of Lemma 5.3. Suppose the process starting in $\in N_i$. Let $V(k) \triangleq V(\boldsymbol{X}(k))$, by Equation (35),

$$V(0) \le V(\boldsymbol{\beta}_i) + \frac{\kappa r}{32 D'}$$

By Theorem 4.3, we know there exists some r such that in $T_1 = O(n \log n)$ steps the process starting at N_i leaves $\boldsymbol{\beta}_i$: $X_{T_1} \in B(\boldsymbol{\beta}_i, 3r/4) \setminus B(\boldsymbol{\beta}_i, r/2)$ with high probability .

Because by direct computation the value of complete Lyapunov function V is a almost a supermartingale , $\mathbb{E}[V(X_{k+1})] \le V(\boldsymbol{X}(k)) + O\left(\frac{1}{n^2}\right)$, by Azuma's inequality(Theorem B.4), with high probability,

$$V(T_1) \le V(0) + \frac{\kappa r}{32 D'} \le V(\boldsymbol{\beta}_i) + \frac{\kappa r}{16 D'}.$$

By Equation (34), $\mathcal{L}_f V(x) \le -\kappa$ for all $x \in \bar{B}(\boldsymbol{\beta}_i, r) \setminus B(\boldsymbol{\beta}_i, r/2)$, we run the process for additional $T_2 = \frac{rn}{4D'}$ steps then

$$\begin{aligned}
V(T_1 + T_2) =& V(T_1) + \sum_{k=T_1}^{T_1+T_2} V(k+1) - V(k) \\
=& V(T_1) + \sum_{k=T_1}^{T_1+T_2} \left(\frac{d}{dt} V(\boldsymbol{X}(k)) + O(\frac{1}{n^2}) \right) \frac{1}{n} \\
\le& V(T_1) + \sum_{k=T_1}^{T_1+T_2} \left(-\kappa + O(\frac{1}{n^2}) \right) \frac{1}{n} \\
\le& V(T_1) - \frac{\kappa r}{4D'} + O\left(\frac{1}{n^2}\right) \\
\le& V(\boldsymbol{\beta}_i) - \frac{\kappa r}{8D'}
\end{aligned}$$

which shows the process leaves the neighborhood N_i in $O(n \log n / \rho)$ time with high probability. □

G Phase portrait: Theorem 7.7

In this section, we prove Theorem 7.7 (which will follow immediately from Theorem G.1), by analyzing the fixed points of the function F_{ND} defined in (15) and apply Proposition A.2 and Corollary A.4. We can classify the fixed points into three types: symmetric, anti-symmetric and eccentric. Lemma G.2 characterizes the property of symmetric fixed points; Lemma G.3, anti-symmetric fixed points; and Lemma G.4, eccentric fixed points. The following section introduces the symmetry property of the flow on F_{ND} and Theorem G.1 is proved in the next one.

G.1 Setup and examples

The fixed points of the system x^{ND} are the zeroes of F_{ND} which can be parameterized by $\delta \triangleq p-q$:

$$\begin{aligned} 0 &= f_{\mathrm{ND}}\left(p\,x_1 + q\,x_2\right) - x_1, \\ 0 &= f_{\mathrm{ND}}\left(p\,x_2 + q\,x_1\right) - x_2. \end{aligned} \tag{36}$$

Denote the solutions of equation (36) as

$$\begin{aligned} \gamma_1 &= \left\{(x_1^{(1)}, x_2^{(1)}) \in [0,1]^2 : x_1^{(1)} = f_{\mathrm{ND}}\left(p\,x_1^{(1)} + q\,x_2^{(1)}\right)\right\} \\ \gamma_2 &= \left\{(x_1^{(2)}, x_2^{(2)}) \in [0,1]^2 : x_2^{(2)} = f_{\mathrm{ND}}\left(p\,x_2^{(2)} + q\,x_1^{(2)}\right)\right\} \end{aligned}. \tag{37}$$

Note that the system of Equation 36 is symmetric with respect to two axes $x_1 = x_2$ and $x_1 + x_2 = 1$, so we define four disjoint regions of $[0,1]^2$:

$$\begin{aligned} R_1 =& \{(x_1, x_2) \in [0,1]^2 : x_1 < x_2 \text{ and } x_1 + x_2 < 1\}, \\ R_2 =& \{(x_1, x_2) \in [0,1]^2 : x_1 < x_2 \text{ and } x_1 + x_2 > 1\}, \\ R_3 =& \{(x_1, x_2) \in [0,1]^2 : x_1 > x_2 \text{ and } x_1 + x_2 < 1\}, \text{ and} \\ R_4 =& \{(x_1, x_2) \in [0,1]^2 : x_1 > x_2 \text{ and } x_1 + x_2 < 1\}. \end{aligned}$$

With this symmetric property, we classify the fixed points of (36) into three types:

- symmetric fixed points: $(x_1^{(s)}, x_2^{(s)})$ such $x_1^{(s)} = x_2^{(s)}$,
- anti-symmetric fixed points:$(x_1^{(a)}, x_2^{(a)})$ such $x_1^{(a)} + x_2^{(a)} = 1$,
- eccentric fixed points:$(x_1^{(e)}, x_2^{(e)})$ such $x_1^{(e)} + x_2^{(e)} > 1$ and $x_1^{(e)} < x_2^{(e)}$.

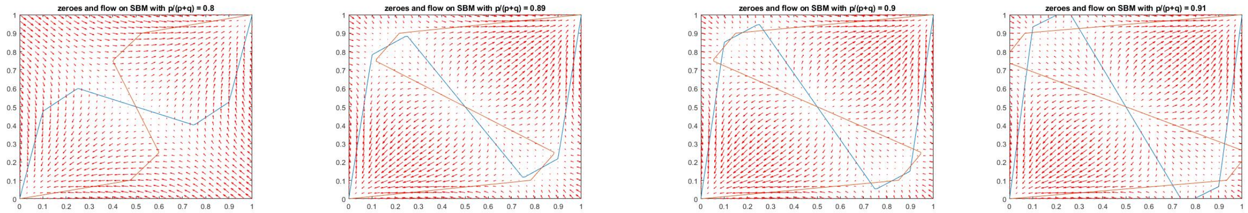
Figure 3 shows some examples of a dynamic with different p, q.

To consider the dynamic φ_{ND} as a flow, there is a caveat: the function F_{ND} only has domain in $[0,1]^2$ instead of $\mathbb{R}^2$, and the set $[0,1]^2$ is not invariant since the $x^{\mathrm{ND}}(t)$ leaves $[0,1]$ if we reverse the time t. Fortunately, it's not hard to extend the domain of F_{ND} without changing the structure: let $m_1 = \lim_{x\to 1^-} f'_{\mathrm{ND}}(x)$ and $m_0 = \lim_{x\to 0^+} f'_{\mathrm{ND}}(x)$

$$\bar{f}_{\mathrm{ND}}(x) = \begin{cases} m_1 x \text{ if } x < 0 \\ f_{\mathrm{ND}}(x) \text{ if } x \in [0,1] \\ m_1(x-1) + 1 \text{ if } x > 1 \end{cases}.$$

We can have $\bar{F}_{\mathrm{ND}}$ by using $\bar{f}_{\mathrm{ND}}$ in (15) instead of f_{ND}.[7]

[7] To make $\bar{f}_{\mathrm{ND}} \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$, we can consider $\epsilon > 0$ and set $f''(x) = 0$ if $x < -\epsilon$ and set the intermediate value in $[-\epsilon, 0]$ smoothly. Then we have an $\mathcal{C}^2$ function moreover it can be arbitrary close to the above definition if we take ϵ small enough.



(a) $p - q < \delta_{\text{symm}}$ (b) $\delta_{\text{symm}} < p - q < \delta_{\text{ecce}}$ (c) $\delta_{\text{ecce}} < p - q < \delta_{\text{anti}}$ (d) $\delta_{\text{anti}} < p - q$

Figure 3: In Theorem G.1 there are three critical values $\delta_{\text{symm}}, \delta_{\text{ecce}}$ and δ_{anti}. In the case (a), the difference $p - q$ is smaller than $\delta_{\text{symm}} = 1/f'_{\text{ND}}(1/2)$, and there are only three fixed points characterized in Lemma G.3. In case (b), the $p - q$ is bigger such that there are two extra saddle anti-symmetric fixed points. For some specific update function f_{ND} there is case (c) such that there are two extra eccentric fixed points but the antisymmetric fixed points are saddle which is discussed in Lemma G.4. Finally in case (d), the $p - q$ is big enough such that the antisymmetric fixed points become attracting which is characterized in Lemma G.3.

G.2 Proof of Theorem 7.7

The following theorem is a detailed characterization of the flow x^{ND} with F_{ND}, and Theorem 7.7 is an corollary of it. In the first case, we take $(\delta', \delta^*, \delta'') = (\delta_{\text{symm}}, \delta_{\text{ecce}}, \delta_{\text{anti}})$ and $(\delta_{\text{symm}}, \delta_{\text{anti}}, \delta_{\text{anti}})$ in the second case.

Theorem G.1 (Phase portrait). *Fix the flow x^{ND} with p, q and $\bar{F}_{\text{ND}}$ defined in* (15)*, depending on the property of f_{ND} there are two situations*

1. *If there exists δ_e such that equation* (36) *with $p_e = (1 + \delta_e)/2$ has an eccentric fixed point $(x_1^{(e)}, x_2^{(e)})$ where $x_1^{(e)} + x_2^{(e)} > 1$ and $x_1^{(e)} < x_2^{(e)}$ there are three constants $\delta_{\text{symm}} < \delta_{\text{ecce}} < \delta_{\text{anti}}$ where $\delta_{\text{anti}} = 1/f'_{\text{ND}}(1/2)$ is defined in Lemma G.2 and δ_{anti} is defined in Lemma G.3 and δ_{ecce} defined in Lemma G.4 such that there are three cases:*

   (a) When $p - q < \delta_{\text{symm}}$, there are only three fixed points $(0, 0), (0.5, 0.5), (1, 1)$. The system is a gradient-like system, and the consensus states $(0, 0), (1, 1)$ are the only attracting fixed point.

   (b) When $\delta_{\text{anti}} < p - q < \delta_{\text{ecce}}$, there are five fixed points, $(0, 0), (0.5, 0.5), (1, 1)$ and two anti-symmetric saddle points. The system is a gradient-like system and the consensus states $(0, 0), (1, 1)$ are the only attracting fixed point.

   (c) When $\delta_{\text{ecce}} < p - q < \delta_{\text{anti}}$ or $\delta_{\text{anti}} < p - q$, there exists an attracting fixed point $\beta \neq (0, 0), (1, 1)$.

2. *Otherwise, there are two constants $\delta_{\text{symm}} < \delta_{\text{anti}}$ where $\delta_{\text{symm}} = 1/f'_{\text{ND}}(1/2)$ is defined in Lemma G.2 and δ_{anti} is defined in Lemma G.3 such that the following three cases:*

   (a) When $p - q < \delta_{\text{symm}}$, there are only three fixed points $(0, 0), (0.5, 0.5), (1, 1)$. The system is a gradient-like system, and the consensus states $(0, 0), (1, 1)$ are the only attracting fixed point.

   (b) When $\delta_{\text{symm}} < p - q < \delta_{\text{anti}}$, there are five fixed points, $(0, 0), (0.5, 0.5), (1, 1)$ and two anti-symmetric saddle points. The system is a gradient-like system and the consensus states $(0, 0), (1, 1)$ are the only attracting fixed point.

(c) When $\delta_{\text{anti}} < p - q$, there exists an attracting fixed point $\beta \neq (0,0),(1,1)$.

We will use two lemmas to proof Theorem G.1.

Lemma G.2 (symmetric fixed points)**.** *Given F_{ND} with p, q and f_{ND}, let $0 < \delta_{\text{symm}} \triangleq 1/f'_{\text{ND}}(1/2)$. There are three symmetric fixed points: $(0,0),(1,1)$ are attracting points, and $(0.5, 0.5)$ which is a saddle point if $(p-q) < \delta_{\text{symm}}$ and a repelling point when $(p-q) > \delta_{\text{symm}}$. Moreover, when $(p-q) < \delta_{\text{symm}}$, the system in (36) only has the above three fixed points.*

Lemma G.3 (anti-symmetric fixed points)**.** *Given F_{ND} with p, q and f_{ND} and δ_{symm} in Lemma G.2, there exists $\delta_{\text{anti}} > \delta_{\text{symm}}$ such that there are two cases for the anti-symmetric fixed points in Equation (36) depending on the value of $p - q$:*

saddle *If $\delta_{\text{symm}} < p - q < \delta_{\text{anti}}$, there are anti-symmetric fixed points which are saddle.*

attracting *If $\delta_{\text{anti}} < p - q$, there are anti-symmetric fixed points which are stable.*

With Lemma G.3, one might guess the systems only have consensus as stable fixed points when $p - q < \delta_{\text{anti}}$, and have two extra stable fixed points when $p - q > \delta_{\text{anti}}$. However, as $p - q$ increases there is some f_{ND} such that the system has extra stable eccentric fixed points before the anti-symmetric fixed points become stable, e.g. Figure 3. Though we can use simulation to estimate the phase space, the following lemma shows: Given f_{ND} suppose there exists $\delta_e < \delta_{\text{anti}}$ such that the system with $\delta_e = p_e - q_e$ in Equation (36) has an eccentric fixed point. Then there exists $\delta_{\text{ecce}} < \delta_{\text{anti}}$ such that for all p'_e such that $\delta_{\text{ecce}} < p'_e - q'_e < \delta_{\text{anti}}$ the system (36) has attracting eccentric stable fixed points fixed points. By symmetry, we only state the result in R_2.

Lemma G.4 (eccentric fixed points)**.** *Given F_{ND} with p, q, f_{ND}, δ_{symm} and δ_{anti} in Lemma G.2, G.3, if there exists $\delta_e < \delta_{\text{anti}}$ such that equation (36) with $p_e = (1+\delta_e)/2$ has an eccentric fixed point $(x_1^{(e)}, x_2^{(e)}) \in R_2$, then for all $\delta_e < \delta'_e < \delta_{\text{anti}}$ the system in (36) with p'_e has an eccentric fixed point $(x_1^{(e)'}, x_2^{(e)'}) \in R_2$ which is a stable fixed point.*

We call $\delta_{\text{ecce}} = \min\{\delta_e\}$ which is the smallest δ_e such that the there exists a eccentric fixed point and anti-symmetric saddle points.

Now we are ready to prove Theorem G.1.

Proof of Theorem G.1. The main statement of theorem is proved by Lemma G.3 and G.4. Now we prove the case 1 and 2 are indeed gradient-like. Because it's only a two dimensional system, by Proposition A.2, we only need to show 1) the system only have constant hyperbolic fixed points, 2) there is no saddle connections 3) there is no cycle.

For the first case, by Lemma G.3, the system have constant hyperbolic fixed points and no saddle connections. By symmetric and positive invariant property of $[0,1]^2$, suppose there is cycle in the system, it should contained in one of the triangles, R_1, R_2, R_3 or R_4. However, it is impossible, since there is no fixed point within those four region.

For the second case, by Lemma G.3 and G.4, the system only have 5 fixed points. Secondly, the saddle point have stable subspace in $\{(x_1, x_2) : x_1 + x_2 = 1\}$, so there is no saddle connection. No limit cycle argument is similar to the first case. □

G.3 Proofs for Lemmas for Phase Portrait

Proof of Lemma G.2. We first show there is no fixed point outside $[0,1]^2$, that is the curve γ_1 and γ_2 do not have intersection outside.

Let $(x_1, x_2) \in \gamma_1 \cap \gamma_2$. When $m_0 = f'_{\text{ND}}(0) = 0$, if $p\,x_1 + q\,x_2 \leq 0$ or $p\,x_2 + q\,x_1 \leq$ by the definition of $\bar{f}_{\text{ND}}$ and γ_1, $(x_1, x_2) = (0, 0)$. On the other hand, when $m_0 = f'_{\text{ND}}(0) > 0$, $\bar{f}_{\text{ND}}$ is monotone, the above solution curve can be rewritten with respect to

$$g(z) \triangleq \frac{1}{q}\left(f_{\text{ND}}^{-1}(z) - pz\right) \tag{38}$$

$$\begin{aligned} \gamma_1 &= \left\{(x_1, x_2) \in [0,1]^2 : x_2 = g(x_1)\right\} \\ \gamma_2 &= \left\{(x_1, x_2) \in [0,1]^2 : x_1 = g(x_2)\right\} \end{aligned} \tag{39}$$

For $x_1 < 0$, because $(x_1, x_2) \in \gamma_1$, $x_2 < x_1$, and because $(x_1, x_2) \in \gamma_2$, $x_2 > x_1$. Therefore there is no fixed point out side $[0,1]^2$.

If $\delta_{\text{symm}} = 1/f'_{\text{ND}}(1/2)$, we want to show $(0,0), (1,1)$ and $(0.5, 0.5)$ are the only intersections between γ_1 and γ_2 in $[0,1]^2$ which by symmetry is enough to show the curve γ_1 is in $R_1 \cup R_3 \cup \{(0,0), (1,1), (0.5, 0.5)\}$. By Definition 7.2, $f_{\text{ND}}(0) = 0$, $f_{\text{ND}}(1/2) = 1/2$, and f_{ND} is strictly convex in $[0.0.5]$, $g(0) = 0$, $g(0.5) = 0.5$, and g is strictly concave in $[0, 0.5]$, so for all $x_1 \in (0, 0.5)$,

$$g(x_1) = g\left((1 - 2x_1) \cdot 0 + 2x_1 \cdot \frac{1}{2}\right) > \left((1 - 2x_1) \cdot g(0) + 2x_1 \cdot g\left(\frac{1}{2}\right)\right) = x_1, \tag{40}$$

and we show γ_1 is above $x_1 = x_2$.

On the other hand, since g is strictly concave and $\mathcal{C}^2$ in $[0, 0.5]$, $g'(x_1) > g'(0.5)$, and $g'(0.5) = \frac{1}{q}\left(\frac{1}{f'(0.5)} - p\right) > -1$, since $p - q < \delta_{\text{symm}} = 1/f'_{\text{ND}}(0.5)$. Thus we have

$$g(x_1) = g(0.5) + \int_{0.5}^{x_1} g'(s)\, ds = 0.5 - \int_{x_1}^{0.5} g'(s)\, ds < 0.5 + (0.5 - x_1),$$

and show

$$g(x_1) + x_1 < 1. \tag{41}$$

Combining equations (40) and (41) we prove the number of fixed points is exactly 3.

For the property of these three fixed points for all p and q. By Definition 3.2 it is sufficient to study the linear approximation of the dynamics at these points:

$$\nabla \bar{F}_{\text{ND}}|_{(x_1,x_2)} = \begin{bmatrix} -1 + p\bar{f}'_{\text{ND}}\,(p\,x_1 + q\,x_2) & q\bar{f}'_{\text{ND}}\,(p\,x_1 + q\,x_2) \\ q\bar{f}'_{\text{ND}}\,(q\,x_1 + p\,x_2) & -1 + p\bar{f}'_{\text{ND}}\,(q\,x_1 + p\,x_2) \end{bmatrix} \tag{42}$$

When $(x_1, x_2) = (0, 0)$, $\nabla \bar{F}_{\text{ND}}|_{(0,0)} = \begin{bmatrix} -1 + p\bar{f}'_{\text{ND}}(0) & q\bar{f}'_{\text{ND}}(0) \\ q\bar{f}'_{\text{ND}}(0) & -1 + p\bar{f}'_{\text{ND}}(0) \end{bmatrix}$ has trace $2(p\bar{f}'_{\text{ND}}(0) - 1)$ and determinant $((p - q)\bar{f}'_{\text{ND}}(0) - 1)(\bar{f}'_{\text{ND}}(0) - 1)$. Thus $\nabla \bar{F}_{\text{ND}}|_{(0,0)}$ has two negative real eigenvalues since $f'_{\text{ND}}(0) < 1$.

Similarly there are two cases for the fixed point $(0.5, 0.5)$: if $1 < f'_{\text{ND}}(0.5) < 1/(p - q)$, the determinant is negative $((p - q)\bar{f}'_{\text{ND}}(0) - 1)(\bar{f}'_{\text{ND}}(0) - 1) < 0$, so $(0.5, 0.5)$ is a saddle point. On the other hand if $f'_{\text{ND}}(0.5) > 1/(p - q)$, $(0.5, 0.5)$ is a repelling point. □

Proof of Lemma G.3. We first show the number of anti-symmetric fixed points is two, than analyze the property of those fixed points.

Because $p - q > \delta_{\text{symm}}$, we have $g'(0.5) = \frac{1}{q}\left(\frac{1}{f'(0.5)} - p\right) < -1$, so the curve γ_1 overlaps with R_2. Therefore there exists a non-symmetric intersection between γ_1 and the line $x_1 + x_2 = 1$,

$(x_1^{(a)}, x_2^{(a)})$ with $x_1^{(a)} \neq x_2^{(a)}$ which is also in the intersection of γ_1 and γ_2 due to the symmetry.

$$\begin{cases} x_1^{(a)} = & \bar{f}_{\text{ND}}\left(p\, x_1^{(a)} + q\, x_2^{(a)}\right) \\ x_2^{(a)} = & \bar{f}_{\text{ND}}\left(p\, x_2^{(a)} + q\, x_1^{(a)}\right) \\ 1 = & x_1^{(a)} + x_2^{(a)} \text{ and } x_1^{(a)} < x_2^{(a)} \end{cases} \tag{43}$$

Because f is convex in $[0, 0.5]$, the system only has two anti-symmetric fixed points $(x_1^{(a)}, x_2^{(a)})$ and $(1 - x_1^{(a)}, 1 - x_2^{(a)})$.

Now we want to show the property of these fixed points. Let $\delta = p - q$ and $s^{(a)} = p\, x_1^{(a)} + q\, x_2^{(a)}$ and $t^{(a)} = p\, x_2^{(a)} + q\, x_1^{(a)}$. Rearrange the above equations we have,

$$1 = f_{\text{ND}}(s^{(a)}) + f_{\text{ND}}(t^{(a)}) \tag{44}$$

$$\frac{p+q}{p-q} = \frac{f_{\text{ND}}(s^{(a)}) - f_{\text{ND}}(t^{(a)})}{s^{(a)} - t^{(a)}} \tag{45}$$

$$1 = s^{(a)} + t^{(a)} \text{ and } s^{(a)} > t^{(a)} \tag{46}$$

Because $1 = x_1^{(a)} + x_2^{(a)}$ and the symmetry of f_{ND}, we have $\bar{f}'_{\text{ND}}\left(s^{(a)}\right) = \bar{f}'_{\text{ND}}\left(t^{(a)}\right)$ and call it $m^{(a)}(\delta)$. By Equation (45) and the convexity of f_{ND}, as δ increases, the derivative at $s^{(a)}$, $m^{(a)}(\delta)$, decreases. By the monotone property, there exists $\delta_{\text{anti}} > \delta_{\text{symm}}$ such that $m^{(a)}(\delta) < 1$ for all $\delta = p - q < \delta_{\text{anti}}$, and $m^{(a)}(\delta) > 1$ for all $\delta < \delta_{\text{anti}}$.

Using Equation (42) the matrix $\nabla \bar{F}_{\text{ND}}|_{(x_1^{(a)}, x_2^{(a)})}$ has the trace $2(pm^{(a)}(\delta) - 1)$ and the determinant $((p-q)m^{(a)}(\delta) - 1)(m^{(a)}(\delta) - 1)$, so

attracting Both eigenvalues are negative, when $m^{(a)}(\delta) < 1$.

saddle One positive and negative eigenvalues, when $\frac{1}{p-q} < m^{(a)}(\delta) < 1$.

Note it is impossible that $\frac{1}{p-q} > m^{(a)}(\delta)$; otherwise, $g'(x_1^{(a)}) < -1$ and implies there are more than two anti-symmetric fixed points contradicting the property of f_{ND}. □

Proof of Lemma G.4. Let $(x_1^{(a)}, x_2^{(a)})$ be the anti-symmetric fixed point defined in (43). Given p_e, q_e and $\delta_e < \delta_{\text{anti}}$, let $(x_1^{(e)}, x_2^{(e)}) \in R_2$ be the eccentric fixed point such that $x_1^{(e)}$ is the smallest value that greater than $x_1^{(a)}$.

We first characterize the local behavior of $(x_1^{(e)}, x_2^{(e)})$. Because f_{ND} is a $\mathcal{C}^2$ function by implicit function theorem, we can parametrize curves (37) as $(x_1^{(1)}, x_2^{(1)})$ and $(x_1^{(2)}, x_2^{(2)})$ of γ_1, and γ_2 respectively. Given $\delta_e < \delta_{\text{anti}}$, by Lemma G.3 $(x_1^{(a)}, x_2^{(a)})$ is a saddle point,

$$m^{(a)}(\delta_e) = \left.\frac{dx_2^{(1)}}{dx_1^{(1)}}\right|_{(x_1^{(a)}, x_2^{(a)})} < 1 < \left.\frac{dx_2^{(2)}}{dx_1^{(2)}}\right|_{(x_1^{(a)}, x_2^{(a)})} = \frac{1}{m^{(a)}(\delta_e)}.$$

By convexity of f_{ND} and definition of $(x_1^{(e)}, x_2^{(e)})$ we have

$$\left.\frac{dx_2^{(2)}}{dx_1^{(2)}}\right|_{(x_1^{(e)}, x_2^{(e)})} \leq \left.\frac{dx_2^{(1)}}{dx_1^{(1)}}\right|_{(x_1^{(e)}, x_2^{(e)})} < m^{(a)}(\delta_e) < 1 \tag{47}$$

Let $I \subseteq (\delta_e, \delta_{\text{anti}})$ be the set of δ such that the system (39) has eccentric fixed points. We want to show the system has an eccentric fixed point when δ is between δ_e and δ_{anti}— $I = (\delta_e, \delta_{\text{anti}})$. Since $(\delta_e, \delta_{\text{anti}})$ is connected, it is sufficient to show the set I is relative open and closed. By the continuity of system (39), we know the set I is closed. To show I is open, without loss of generality, we show there is a neighborhood of δ_e contained in I. Given $(x_1^{(e)}, x_2^{(e)})$ with δ_e, fixing $x_1 = x_1^{(e)}$, let's consider and the movement of $x_2^{(1)}(\delta)$ and $x_2^{(2)}(\delta)$ as δ changes around δ_e where $x_2^{(1)}(\delta)$ (and $x_2^{(2)}(\delta)$) is the highest intersection between $x_1 = x_1^{(e)}$ and γ_1 (γ_2 respectively).

$$\frac{d}{d\delta}\left(x_2^{(1)} - x_2^{(2)}\right) > 0. \tag{48}$$

Informally, by Equation (39), as δ changes, the curve γ_1 is stretched vertically (x_2 direction) and the movement is proportional to the change rate of δ. On the other hand, γ_2 is stretched horizontally (x_1 direction), and by Equation (47) the slope is smaller than 1, so the vertically increment rate is smaller than the rate of δ. Therefore the $x_2^{(1)}(\delta)$ should increase faster than $x_2^{(2)}(\delta)$ in x_2. Now let give a formal argument. Through direct computation on Equation (39),

$$\frac{dx_2^{(1)}}{d\delta} = \frac{1}{2(1-\delta)}(x_2^{(1)} - x_1^{(1)}) = \frac{1}{2(1-\delta)}(x_2^{(e)} - x_1^{(e)}).$$

Similarly,

$$\left(1 + \frac{1}{1-\delta}\left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)}))} - 1\right)\right)\frac{dx_2^{(2)}}{d\delta} = \frac{1}{2(1-\delta)}(x_2^{(e)} - x_1^{(e)})$$

Therefore, to prove Equation (48), it is sufficient to show

$$\left(1 + \frac{1}{1-\delta}\left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)}))} - 1\right)\right) > 1. \tag{49}$$

This can be proved by taking derivative at Equation (39) with respect to $x_1^{(2)}$ and applying Equation (47),

$$1 = \left(1 + \frac{1}{1-\delta}\left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)}))} - 1\right)\right)\frac{dx_2^{(2)}}{dx_1^{(2)}} < \left(1 + \frac{1}{1-\delta}\left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)}))} - 1\right)\right).$$

Now, let's prove the eccentric fixed point is stable. Note that by (47) and (48), for all $\delta > \delta_e$,

$$0 < \left.\frac{dx_2^{(2)}}{dx_1^{(2)}}\right|_{(x_1^{(e)}, x_2^{(e)})} < \left.\frac{dx_2^{(1)}}{dx_1^{(1)}}\right|_{(x_1^{(e)}, x_2^{(e)})} < 1. \tag{50}$$

Rewrite the above inequality in terms of f_{ND} we have,

$$1 > \frac{1}{1-\delta}\left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_1^{(e)}))} - \delta\right) > \left[\frac{1}{1-\delta}\left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)}))} - \delta\right)\right]^{-1} > 0.$$

By Equation (42), the matrix $\nabla \bar{F}_{\text{ND}}|_{(x_1^{(e)}, x_2^{(e)})}$ is

$$\begin{bmatrix} -1 + p\bar{f}'_{\text{ND}}(f_{\text{ND}}^{-1}(x_1^{(e)})) & q\bar{f}'_{\text{ND}}(f_{\text{ND}}^{-1}(x_1^{(e)})) \\ q\bar{f}'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)})) & -1 + p\bar{f}'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)})) \end{bmatrix}.$$

The trace is negative, because $f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_1^{(e)})) < 1$ and $f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)})) < 1/\delta$. The determinant is positive, because $\left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_1^{(e)}))} - \delta\right) \cdot \left(\frac{1}{f'_{\text{ND}}(f_{\text{ND}}^{-1}(x_2^{(e)}))} - \delta\right) > (1-\delta)^2$. Therefore, the $(x_1^{(e)}, x_2^{(e)})$ is a stable fixed point.

□

H Proof for Theorem 7.6

To prove the first part, our proof has two steps: Let $E = [0,1]^2$ given $r > 0$ a neighborhood of consensus states $Q(r) = (B(\mathbf{0}, r) \cup B(\mathbf{1}, r)) \cap E$, the Markov chain $\boldsymbol{X}^{\text{ND}}$ reaches $Q(r)$ in $O(n \log n)$ with high probability, and it hits the consensus states in $O(n \log n)$ with constant probability when $Q(r)$ small enough. The first one is proved in Lemma H.1 and the second part is proved in Lemma H.3.

Lemma H.1 (Reaching neighborhood Q)**.** *In case* 1 *of Theorem 7.6, given any* $r > 0$ *and* $Q(r) = (B(\mathbf{0}, r) \cup B(\mathbf{1}, r)) \cap E$ *(we omit r later), there is* $\tau_Q = O(n \log n)$ *such that the hitting time of* $\boldsymbol{X}^{\text{ND}}$ *to set* $Q(r)$ *from any initial states is smaller than* τ_Q *with high probability*

$$\Pr[T_h(Q(r)) \leq \tau_Q \mid \boldsymbol{X}(0) \in E] = 1 - o(1).$$

Proof of Lemma H.1. With Theorem 7.7, and 5.1, $\boldsymbol{X}^{\text{ND}}$ reaches a fixed neighborhood of consensus states $(0,0), (1,1)$, Q in $O(n \log n)$ with high probability if the noise is well-behaved:

$$\exists \alpha > 0, \forall \boldsymbol{x} \in E \setminus Q(r), \alpha \mathbb{I}_d \prec \operatorname{Cov}[\boldsymbol{U}(k+1) \mid \boldsymbol{X}^{\text{ND}}(k) = \boldsymbol{x}]. \tag{51}$$

which is proved in Lemma H.2.

□

Lemma H.2 (Well-behaved noise)**.** *Given* $\boldsymbol{X}^{\text{ND}}$ *defined in* (2)*, there exist* α*, for all* $\boldsymbol{x} \in E \setminus Q(r)$*, and k*

$$\alpha \mathbb{I}_d \prec \operatorname{Cov}[\boldsymbol{U}(k+1) \mid \boldsymbol{X}^{\text{ND}}(k) = \boldsymbol{x}].$$

Lemma H.3 (Reaching consensus)**.** *In the first case of Theorem 7.6, let* $C = \{(0,0), (1,1)\}$ *be the set of consensus states. There exist* $\tau_C = O(n \log n)$*, and* $r_{\text{reg}} > 0$*, such that for all* $\boldsymbol{x} \in Q(r_{\text{reg}})$

$$\Pr[T_h(C) \leq \tau_C] \geq 1/6.$$

With above lemmas, we are ready to prove the Theorem 7.6

Proof of Theorem 7.6. For the first part, let $r_{\text{reg}} > 0$ be defined in Lemma H.3, by Lemma H.1 $\boldsymbol{X}^{\text{ND}}$ reaches, $Q(r_{\text{reg}})$ in $\tau_Q = O(n \log n)$ with high probability. By Lemma H.3, the process further hits consensus states C in $\tau_C = O(n \log n)$ with probability at least 1/7. Therefore

$$\Pr[T_h(C) \leq \tau_C + \tau_Q \mid \boldsymbol{X}^{\text{ND}}(0) \in E] \geq 1/7. \tag{52}$$

Because the $\boldsymbol{X}^{\text{ND}}$ is a Markov chain bounds are independent of different time intervals with length $\tau_C + \tau_Q$, so $\text{ME}(K(n,p,q), f_{\text{ND}}) = O(n \log n)$.

For the second part, by Theorem 7.7 there is an extra attracting fixed point $\boldsymbol{\beta}_a$ of $\bar{F}_{\text{ND}}$. By Proposition 4.8, there exists neighborhoods of $\boldsymbol{\beta}_a$, r_{in} and r_{out} such that for any σ_0 with $\boldsymbol{X}^{\text{ND}}(0) \in B(\boldsymbol{\beta}_a, r_{\text{in}})$ and $T \geq 1$, we have $\Pr[X_T \in B(\boldsymbol{\beta}_a, r_{\text{out}})] \geq 1 - T \exp(-\Omega(n))$. Therefore, with initial state $\boldsymbol{X}^{\text{ND}}(0) = \boldsymbol{\beta}_a$

$$\Pr[T_h(C) \geq k \mid \boldsymbol{X}^{\text{ND}}(0) = \boldsymbol{\beta}_a] \geq \Pr[\boldsymbol{X}^{\text{ND}}(k) \in B(\boldsymbol{\beta}_a, r_{\text{out}}) \mid \boldsymbol{X}^{\text{ND}}(0) = \boldsymbol{\beta}_a] \geq 1 - k \exp(-\Omega(n))$$

Because the hitting time is a non-negative random variable

$$\mathbb{E}[T_h(C) \mid \boldsymbol{X}^{\mathrm{ND}}(0) = \boldsymbol{\beta}_a] = \sum_k \Pr[T_h(C) \geq k] \geq \sum_k 1 - k\exp(-\Omega(n)) = \exp(\Omega(n)).$$

□

Proof of Lemma H.2. Since $\boldsymbol{X}^{\mathrm{ND}}$ is a Markov chain, given $\boldsymbol{X}^{\mathrm{ND}}(k) = \boldsymbol{x} = (x_1, x_2) \in \Omega_X \setminus Q$, the difference to be $\boldsymbol{Y} \triangleq n(\boldsymbol{X}^{\mathrm{ND}}(k+1) - \boldsymbol{X}^{\mathrm{ND}}(k))$ which is independent to the index k, and $\boldsymbol{Y} = (Y_1, Y_2) \in \{(0,0), (1,0), (-1,0), (0,1), (0,-1)\}$ only have these five possible outcomes, and we can compute these directly:

$$\begin{aligned}
p_1^+(\boldsymbol{x}) \triangleq& \Pr[\boldsymbol{Y} = (1,0) \mid \boldsymbol{X} = \boldsymbol{x}] = \frac{1-x_1}{2} f_{\mathrm{ND}}(px_1 + qx_2),\\
p_1^-(\boldsymbol{x}) \triangleq& \Pr[\boldsymbol{Y} = (-1,0) \mid \boldsymbol{X} = \boldsymbol{x}] = \frac{x_1}{2}\left(1 - f_{\mathrm{ND}}(px_1 + qx_2)\right),\\
p_2^+(\boldsymbol{x}) \triangleq& \Pr[\boldsymbol{Y} = (0,1) \mid \boldsymbol{X} = \boldsymbol{x}] = \frac{1-x_2}{2}\left(f_{\mathrm{ND}}(qx_1 + px_2)\right),\\
p_2^-(\boldsymbol{x}) \triangleq& \Pr[\boldsymbol{Y} = (0,-1) \mid \boldsymbol{X} = \boldsymbol{x}] = \frac{x_1}{2}\left(1 - f_{\mathrm{ND}}(qx_1 + px_2)\right).
\end{aligned}$$

We omit $\boldsymbol{x}$ when it is clear. We define $\boldsymbol{U}(\boldsymbol{x})$ be the noise $\boldsymbol{U}(1)$ condition on $\boldsymbol{X}^{\mathrm{ND}}(0) = \boldsymbol{x}$ which is well-define because $\boldsymbol{X}^{\mathrm{ND}}$ is a Markov chain. By the definition of $\boldsymbol{U}(x)$ and Y,

$$\mathrm{Cov}[\boldsymbol{U}(\boldsymbol{x})] = \mathrm{Cov}[\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}] = \begin{bmatrix} p_1^+ + p_1^- - (p_1^+ - p_1^-)^2 & -(p_1^+ - p_1^-)(p_2^+ - p_2^-) \\ -(p_1^+ - p_1^-)(p_2^+ - p_2^-) & p_2^+ + p_2^- - (p_2^+ - p_2^-)^2 \end{bmatrix}.$$

Let $S_1 = p_1^+ + p_1^-$, $S_2 = p_2^+ + p_2^-$, $D_1 = p_1^+ - p_1^-$, and $D_2 = p_2^+ - p_2^-$, and $\mathrm{Cov}[\boldsymbol{U}(\boldsymbol{x})]$ can be simplified as,

$$\mathrm{Cov}[\boldsymbol{U}(\boldsymbol{x})] = \begin{bmatrix} S_1 - D_1^2 & -D_1 D_2 \\ -D_1 D_2 & S_2 - D_2^2 \end{bmatrix}. \tag{53}$$

Because $\mathrm{Cov}[\boldsymbol{U}(\boldsymbol{x})]$ is symmetric, the eigenvalues are real. By Gershgorin circle theorem and (53), the eigenvalues are upper bounded by

$$\max\left\{S_1 - D_1^2 + |D_1 D_2|, S_2 - D_2^2 + |D_1 D_2|\right\} \leq 1,$$

and lower bounded by

$$\min\left\{S_1 - D_1^2 - |D_1 D_2|, S_2 - D_2^2 - |D_1 D_2|\right\}, \tag{54}$$

so to find d_1 it is sufficient to lower bound Equation (54).

By the definition of $Q(r)$, there exists constant $r > 0$ such that 1-norm balls $\{\boldsymbol{U}(\boldsymbol{x}) \in E : \|x\|_1 \leq r\}$ and $\{\boldsymbol{U}(\boldsymbol{x}) \in E : \|x - (1,1)\|_1 \leq r\}$ are insides $Q(r)$. Thus, if $(x_1, x_2) \in E \setminus Q(r)$, $px_1 + qx_2$, $qx_1 + \bar{b}x_2$ are in $[qr, p(1-r)]$, so

$$0 < f_{\mathrm{ND}}(qr) \leq f_{\mathrm{ND}}(px_1 + qx_2) \text{ and } f_{\mathrm{ND}}(qx_1 + px_2) \leq f_{\mathrm{ND}}(p(1-r)) < 1 \tag{55}$$

As a result, p_1^+, p_1^-, p_2^+ and p_2^- are smaller or equal to $\frac{1}{2} f_{\mathrm{ND}}(p(1-r))$, and $|D_1|, |D_2| \leq \frac{1}{2} f_{\mathrm{ND}}(p(1-r))$. Moreover,

$$\begin{aligned}
(53) \geq& \min\left\{S_1 - f_{\mathrm{ND}}(p(1-r))|D_1|, S_2 - f_{\mathrm{ND}}(p(1-r))|D_2|\right\}\\
\geq& (1 - f_{\mathrm{ND}}(p(1-r))) \min\left\{S_1, S_2\right\}.
\end{aligned}$$

Because $S_1 = p_1^+ + p_1^-$ is a convex combination of $f_{\mathrm{ND}}(px_1 + qx_2)/2$ and $(1 - f_{\mathrm{ND}}(px_1 + qx_2))/2$, and $S_2 = p_2^+ + p_2^-$ is a convex combination of $f_{\mathrm{ND}}(qx_1 + px_2)/2$ and $(1 - f_{\mathrm{ND}}(qx_1 + px_2))/2$, by (55), $\min\{S_1, S_2\} \geq \frac{1}{2}\min\{f_{\mathrm{ND}}(qr), 1 - f_{\mathrm{ND}}(p(1-r))\}$,

$$(53) \geq (1 - f_{\mathrm{ND}}(p(1-r))) \cdot \frac{1}{2}\min\{f_{\mathrm{ND}}(qr), 1 - f_{\mathrm{ND}}(p(1-r))\} > 0$$

Therefore, we can take $0 < \alpha < \frac{1}{2}(1 - f_{\mathrm{ND}}(p(1-r))) \cdot \min\{f_{\mathrm{ND}}(qr), 1 - f_{\mathrm{ND}}(p(1-r))\}$ which completes the proof. □

Proof of Lemma H.4. Let $\psi(k) = \sum_{1 \leq \ell \leq k} d_\ell$ and $\psi(0) = 0$. By direct computation, for all $0 < k < m$

$$\begin{aligned}
\mathcal{L}\psi(k) =& p^+(k)\left(\psi(k+1) - \psi(k)\right) - p^-(k)\left(\psi(k) - \psi(k-1)\right) \\
=& p^+(k)d_{k+1} - p^-(k)d_k && \text{(definition of } \psi\text{)} \\
\leq& -1 && \text{(definition of } d_k\text{)}
\end{aligned}$$

Finally, $\mathcal{L}\psi(m) = -p^-(k)\left(\psi(k) - \psi(k-1)\right) - p^-(k)d_k \leq -1$. Therefore $\psi(m)$ is a upper bound for the maximum expected hitting time by Corollary B.1. □

H.1 From neighborhood of attracting fixed points to fixed points

In this section, we want to prove Lemma H.3: once the process $\boldsymbol{X}^{\mathrm{ND}}$ hits the set Q defined in Lemma H.1 process reaches consensus states with constant probability within $O(n \log n)$ time. We achieve this by coupling the process with a birth-and-death chain. In Lemma H.4, we give a simple upper bound for hitting time of birth-and-death chain. In Lemma H.5, a uniform bound for (56) is given for our process.

Lemma H.4 (Hitting time of birth-and-death chains)**.** *Let discrete time Markov chain W_k be a birth-and-death chain on space $\Omega = \{0, 1, \ldots, m\}$ such that in each transition the state can increase or decrease by at most 1 where*

$$\begin{aligned}
\Pr[W' = W + 1 \mid W = \ell] =& \quad p^+(\ell) \\
\Pr[W' = W \mid W = \ell] =& \quad 1 - p^+(\ell) - p^-(\ell) \\
\Pr[W' = W - 1 \mid W = \ell] =& \quad p^-(\ell)
\end{aligned}$$

Let $d_1, \ldots, d_m$ be a positive sequence such that

$$d_m \geq \frac{1}{p^-(m)} \text{ and } d_{l-1} \geq \frac{1}{p^-(\ell-1)} + \left(\frac{p^+(\ell+1)}{p^-(\ell-1)}\right) d_l \tag{56}$$

Then the maximum expected hitting time from state ℓ to 0 can be bounded as follows:

$$\max_{\ell \in \Omega} \mathbb{E}[T_0(x)] \leq \sum_{0 < \ell \leq m} d_\ell$$

where $T_0(x)$ denotes the hitting time from state x to state 0.

To simplify the notions we use $\boldsymbol{X}$ to represent $\boldsymbol{X}^{\mathrm{ND}}(k)$ and $\boldsymbol{X'} = \boldsymbol{X}^{\mathrm{ND}}(k+1)$ where the index does not mater because it is a Markov chain. We also apply this notion to other Markov chains.

Lemma H.5. *Let $h(\boldsymbol{x}) \triangleq n(x_1 + x_2)/2$. There exist positive constants α, γ and ϵ, such that for all $\boldsymbol{x}$ with $h(\boldsymbol{x}) \leq \epsilon n$,*

$$\Pr\left[h(\boldsymbol{X}') = h(\boldsymbol{X}) - 1 \mid \boldsymbol{X} = \boldsymbol{x}\right] \geq \gamma h(\boldsymbol{x})/n, \tag{57}$$

and

$$\frac{\Pr\left[h(\boldsymbol{X}') = h(\boldsymbol{X}) + 1\right]}{\Pr\left[h(\boldsymbol{X}') = h(\boldsymbol{X}) - 1\right]} \leq 1 - \alpha. \tag{58}$$

Proof of Lemma H.3. Without loss of generality, we consider $\boldsymbol{x} \in B(\mathbf{0}, \epsilon)$. Let $V(k) = h(\boldsymbol{X}^{\text{ND}}(k))$ is a stochastic process on $\mathbb{N}$ and the process $\boldsymbol{X}^{\text{ND}}$ reaches $(0,0)$ if and only $h(\boldsymbol{X}^{\text{ND}}) = 0$. We define $m_0(\epsilon) = \max\{h(\boldsymbol{x}) : \boldsymbol{x} \in B(\mathbf{0}, \epsilon)\} = \Theta(n)$.

To show the process hits $(0,0)$ in $O(n \log n)$ with probability 1/6, the proof has two steps: we first upper bound the expected optional stopping time, $T = \min\{k : V(k) = 0 \vee V(k) \geq 2m_0\}$,

$$\mathbb{E}[T] = \tau' = O(n \log n) \tag{59}$$

Then show

$$\Pr[V(T) = 0] \geq \Pr[V(T) \geq 2m_0(\epsilon)] \tag{60}$$

With the above two equations, we have

$$\begin{aligned} \Pr[T \leq 3\tau'] \geq & \Pr[T \leq 3\tau' \wedge V(T) = 0] \\ \geq & 1 - \Pr[V(T) \neq 0] - \Pr[T \geq 3\tau'] && \text{(union bound)} \\ \geq & 1/2 - 1/3 = 1/6 && \text{(by Markov inequality and (60))} \end{aligned}$$

Now let's prove the Equation (59) and (60). For Equation (59) we couple the process $V(k)$ with a birth-and-death chain $W(k)$ as follows: $W(k)$ is a Markov chain on space $\{0, 1, \ldots, 2m_0\}$, one step the state can increase or decrease by at most 1 such that for all $0 < \ell < 2m_0$

$$\begin{aligned} \Pr[W' = W + 1 \mid W = \ell] = & \quad \max\nolimits_{\boldsymbol{x}:h(\boldsymbol{x})=\ell} \Pr[V' = V + 1 \mid V = h(\boldsymbol{x})] \\ \Pr[W' = W - 1 \mid W = \ell] = & \quad \min\nolimits_{\boldsymbol{x}:h(\boldsymbol{x})=\ell} \Pr[V' = V - 1 \mid V = h(\boldsymbol{x})] \end{aligned} \tag{61}$$

recalled that we use W' to denote state of single transition of a discrete time Markov chain starting at W. For the boundary states 0 and $2m_0$, we set $\Pr[W' = W + 1 \mid W = 2m_0] = 0$ and $\Pr[W' = W - 1 \mid W = 0] = 0$.

By Lemma H.5 and H.4, the expected hitting time of W to state 0 is upper bounded by $\sum_{\ell \leq 2m_0} d_\ell$ where d_ℓ is defined in Lemma H.4. By Lemma H.5, we can set $d_{2m_0} = \frac{n}{\gamma 2m_0} = O(1)$, for all $1 \leq \ell < 2m_0$, $d_\ell = \frac{1}{\gamma \ell} + (1-\alpha)d_{\ell+1}$. By induction there exists C such that $d_\ell \leq \frac{Cn}{\ell}$ for all $1 \leq \ell \leq 2m_0$. Therefore

$$\mathbb{E}[\min\{k : W(k) = 0\}] \leq \sum d_\ell = O(n \log n).$$

By the definition of $W(k)$, we can couple these two process $V(k)$ and $W(k)$ before the process hits the boundary such that $W(k) \geq V(k)$ for all $k \leq \tau$. Therefore, we can upper bound $\mathbb{E}[\tau] \leq \mathbb{E}[\min\{k : W(k) = 0\}] = O(n \log n)$.

Finally Equation (60) is true, because $V(k)$ is a supermartingale, $\mathbb{E}[V(k+1) \mid \boldsymbol{X}^{\text{ND}}(k)] \leq V(k)$ by Lemma H.5. □

Proof of Lemma H.5. This Lemma shows if the fraction of opinion 1 in V_1 and V_2 is smaller than α, the number of 1 opinion decrease fast. Given configuration $\boldsymbol{X}^{\text{ND}}(k)$, let a_k, b_k be the number of 1 opinion in V_1, V_2 at time k. Note that the update function f_{ND} is smooth and strictly concave in $[0.5, 1]$ and $f_{\text{ND}}(1) = 1, f_{\text{ND}}(0.5) = 0.5$, there exists m_1 such that $f'_{\text{ND}}(1) < m_1 < 1$ and for all $0 < 1 - x < \epsilon$

$$f_{\text{ND}}(x) \leq 1 + m_1(x - 1). \tag{62}$$

Similarly there exists m_0 such that $f'_{\text{ND}}(0) < m_0 < 1$ and for all $0 < x < \epsilon$

$$f_{\text{ND}}(x) \geq m_0 x. \tag{63}$$

Let $V(k) = h(\boldsymbol{X}^{\text{ND}}(k))$. We first prove (57). The event that $V(k+1) = V(k) - 1$ is equal at time $k+1$ a node with opinion 1 is chosen and updates its opinion to 0,

$$\begin{aligned}
&\Pr[V(k+1) = V(k) - 1 \mid \boldsymbol{X}^{\text{ND}}(k)] \\
=&\frac{a_k}{n} \Pr[v_1 \in V_1 \text{ updates to } 0] + \frac{b_k}{n} \Pr[v_2 \in V_2 \text{ updates to } 0] \\
=&\frac{a_k}{n} \left(1 - f_{\text{ND}}\left(p\frac{2a_k}{n} + q\frac{2b_k}{n}\right)\right) + \frac{b_k}{n}\left(1 - f_{\text{ND}}\left(q\frac{2a_k}{n} + p\frac{2b_k}{n}\right)\right) \\
\geq&\frac{a_k}{n} m_1 \left(1 - p\frac{2a_k}{n} - q\frac{2b_k}{n}\right) + \frac{b_k}{n} m_1 \left(1 - q\frac{2a_k}{n} - p\frac{2b_k}{n}\right) && \text{(by (62))} \\
\geq&\frac{a_k + b_k}{n} m_0 \left(1 - 2\epsilon\right) \\
\geq&\frac{m_1}{2n}(a_k + b_k) = \frac{m_1}{2} V(k)/n && \text{(if } \epsilon \text{ smaller than } 1/4)
\end{aligned}$$

Therefore this proves (57) by taking $0 < \gamma < \frac{m_1}{2}$

For the (58), with (57), it is sufficient to show there exists δ such that $\Pr\left[V(k+1) = V(k) - 1\right]$ minus $\Pr\left[V(k+1) = V(k) + 1\right]$ is greater than $\delta V(k)/n$. This can be done by computation

$$\begin{aligned}
&\Pr\left[V(k+1) = V(k) - 1\right] - \Pr\left[V(k+1) = V(k) + 1\right] \\
=&\mathbb{E}[V(k+1)] - V(k) \\
=&\mathbb{E}[a_{k+1} + b_{k+1}] - a_k - b_k \\
=&f_{\text{ND}}(pa_k/n + qb_k/n) + f_{\text{ND}}(qa_k/n + pb_k/n) \\
\geq&m_0(pa_k/n + qb_k/n) + m_0(qa_k/n + pb_k/n) && \text{(by (63))} \\
\geq&m_0(a_k/n + b_k/n) = m_0 V(k)/n
\end{aligned}$$

, and these complete the proof for (58). □