Abstract

We study opinion dynamics on networks with two communities. Each node has one of two opinions and updates its opinion as a “majority-like” function of the frequency of opinions among its neighbors. The networks we consider are weighted graphs comprised of two equally sized communities where intracommunity edges have weight $p$, and intercommunity edges have weight $q$. Thus $q$ and $p$ parameterize the connectivity between the two communities.

We prove a dichotomy theorem about the interaction of the two parameters: 1) the “majority-like” update function, and 2) the level of intercommunity connectivity. For each setting of parameters, 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.

Technically, we achieve this the fast convergence result by exploiting the connection between a family of reinforced random walks and dynamical systems literature. Our main result shows if the systems are a reinforced random walk with a gradient-like function, it converges to an arbitrary neighborhood of a local attracting point in $O(n \log n)$ time with high probability. This result adds to the recent literature on saddle-point analysis and shows a large family of stochastic gradient descent algorithm converges to local minimal in $O(n \log n)$ when the step size $O(1/n)$.

Our opinion dynamics model captures a broad range of systems, sometimes called interacting particle systems, exemplified by the voter model, iterative majority, and iterative $k-$majority processes—which have found use in many disciplines including distributed systems, statistical physics, social networks, and Markov chain theory.
1 Introduction

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. This general model captures important settings in sociology (competing beliefs or practices), economics (competing technologies/brands), statistical physics (e.g., the Ising Model), distributed computer systems (consensus), and evolutionary biology (genetic inheritances).

We study the maximum expected consensus time on a general set of stochastic process called Node dynamics. Node dynamics are parameterized by an update function $f : [0, 1] \rightarrow [0, 1]$. 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.

**Iterative $\rho$-noisy majority model:** Update a node’s opinion to majority opinion its neighbors with probability $1 - \rho$ and uniformly at random with probability $\rho$.

We focus on a specific regime of $f$ that we call “majority-like” (definition 6). In particular $f$ should be monotone, symmetric, twice continuously differentiable, strictly convex in $[0, .5]$, and have $f(0) = 0$. Though node dynamics with majority-like function do not contain the iterative majority (non-smooth) nor voter model (not strictly convex), this still captures rich get richer property and yields a very general class of functions including $k$-majority. Such functions promote consensus within a single homogeneous community. This result is intuitive because once symmetry is broken, the majority should retain its advantage. Here we study whether such dynamics also quickly reach consensus in models with community structure from arbitrary initial states.

While previous work on this general model only considered Erdős-Rényi graphs, we study what happens when community structure is present. 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.

Our Contributions We prove a dichotomy theorem about the interaction of the update function and the level of intercommunity connectivity. For each “majority-like” function we show a threshold such that if the communities (the difference between $p$ and $q$) are more connected than this threshold value, they will converge to consensus in time $\Theta(n \log(n))$ from arbitrary initial states. However, if they are more isolated than this threshold value, an attracting, non-convergence fixed point will emerge which can delay consensus to $2^{\Omega(n)}$ steps. For technical reasons, there may exist a single point in each region (above and below the threshold) that we cannot classify.

There are two challenges for showing fast consensus from worst case initial state: breaking symmetry and measuring progress. For breaking symmetry, if initially precisely half of the agents have the red opinion in both each community, our process does not move toward consensus in expectation. Thus, we show the randomness in the process can break this symmetry and help
escaping such singular states. On the other hand, we also need to measure the global progress of the dynamics to prevent it making cycle or having other complicated recurrent behavior.

Our analysis is mostly decoupled from the particular problem we are solving, but instead relies on two properties: the mean-field dynamics is a gradient-like flow (which implies the existence of a potential function for us to measure the progress), and the dynamics are reinforcing random walks (which enables us to break symmetry). Thus we believe there will be other applications of it in the future.

To break symmetry, our main technical lemma shows that a dynamics with the two properties mentioned above can quickly escape from non-attracting fixed points. This result adds to the recent literature on saddle-point analysis. In particular, the process studied in Theorem 10 greatly generalize Ge et al. [25], Jin et al. [32], and we prove the convergence time is $O(\ell \log \ell)$ as long as the noise is well-behaved, and the objective function has a continuous third derivative.

Finally, our work has applications to ideological polarization [8]. The threshold behavior implies that even when the dynamics are very polarized, a small change in the network or processes can lead to large-scale consensus. Conversely, if the dynamics are far from the threshold, small measures may yield no effect at all.

1.1 Related Work

The most closely related work is the rigorous treatment of node dynamics on dense Erdős-Rényi graphs [45] which shows that it converges in time $O(n \log(n))$. The technical difficulties dealt with by these works are largely orthogonal. The difficulty with Erdős-Rényi graphs is that the exponentially many configurations are all slightly different. Thus the potential function must be custom designed for each function $f$ but done in an automated fashion. However, the current treatment assumes a complete graph with weighted edges, and so the state can be succinctly represented. Here the difficulty in constructing a nice potential function comes from the fact that there may be more than one non-attracting fixed point.

As mentioned, our model extends several previously studied dynamics including the voter model, iterative majority, iterative $k$-majority. The voter model has been extensively studied in mathematics [13, 29, 37, 38], physics [5, 10], and even in social networks [9, 48, 49, 50, 12]. A major theme of this work is how long it takes the dynamics to reach consensus on different network topologies. Works about iterative majority dynamics [36, 7, 34, 42, 51, 54] often study when the dynamics converge and how long it takes them to do so. Another interest question, orthogonal to those explored here, is whether the dynamics converge to the original majority opinion—that is, successfully aggregate the original opinion. Doerr et al. [18] 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 [14, 11, 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 [35, 43, 6, 41]. Another line of work deals with agents selecting an opinion from among a large (or infinite) set of options [3, 24]. There are also myriad models where the opinions space is continuous instead of discrete. Typically agents either average their neighbors’ opinions [16], or a subset of their neighbors’ opinions which are sufficiently aligned [28, 15]. Finally, models involving the coevolution of the opinions and the network [30, 19] have been studied using simulations and heuristic arguments.

A large volume of literature 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 [40], (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 [42]. Our results extend the second approach. We map our high dimensional process into a process on a low dimensional space ($\mathbb{R}^2$). This new process is a reinforced random walk with small step size which is closely related to the solution of an ordinary differential system which can be seen as a mean-field approximation of our random walk. 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 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.

Recently, there is a long line of research of stochastic gradient descent on non-convex functions, see [25, 32] 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.

2 Preliminaries

2.1 Flows, mappings, and reinforced random walks

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

1. **Vector fields** or ordinary differential equation with $f$ solve for:

   \[
   \frac{d}{dt} x = f(x). \tag{1}
   \]

   The result is the continuous function $\varphi : \mathcal{X} \times \mathbb{R} \to \mathcal{X}$ such that $\varphi(x, 0) = x$ and $\frac{d}{dt} \varphi(x, t) = f(\varphi(x, t))$ for all $t \in \mathbb{R}, x \in \mathcal{X}$.

2. **Maps** or difference equations with $f$ are discrete time processes

   \[
   x_{k+1} = x_k + \frac{1}{n} f(x_k), \tag{2}
   \]

   and the range of change of each update, $\| \frac{1}{n} f(x_k) \|$, is bounded by $1/n$ for some large $n \in \mathbb{N}$ when the process is in some compact set $B \subset \mathbb{R}^d$.

3. **Reinforced random walks with $f$** consider the evolution of a process subject to an unbiased stochastic perturbation. Let $(X_k, F_k)$ be a random process in $\mathcal{X}$ with filtration $\mathcal{F}$ which can be composed of a predictable part $f(X_k)$ and noise part $U_{k+1}$:

   \[
   X_{k+1} = X_k + \frac{1}{n} (f(X_k) + U_{k+1}) \tag{3}
   \]

   such that for all $x \in \mathbb{R}^d$, $\mathbb{E}[U_{k+1} | F_k] = 0$. 


2.2 Dynamical systems

First let’s define some basic notions which are mostly from Robinson [44]. Let $\mathcal{X}$ be $\mathbb{R}^d$. A $C^r$-flow $\varphi$ is defined to be a $C^r$-function, $\varphi : \mathcal{X} \times \mathbb{R} \to \mathcal{X}$ with the property that $\forall x_0 \in \mathcal{X}, t_1, t_2 \in \mathbb{R}$,

$$\varphi(x_0, 0) = x_0; \quad \varphi(x_0, t_1 + t_2) = \varphi(\varphi(x_0, t_1), t_2).$$

Given function $f \in C^r$, initial condition $x \in \mathcal{X}$, and time $t \in \mathbb{R}$, the solution of $\varphi(x, t; f)$ forms a $C^r$-flow $\varphi(x, t; f)$ called the flow with $f$. We call a set $B \subseteq \mathcal{X}$ positive invariant if and only if for all $x \in B$ and $t \geq 0$, $\varphi(x, t) \in B$, negative invariant iff it’s true for all $t \leq 0$, and invariant iff it’s true for all $t \in \mathbb{R}$.

The trajectory or orbit of a point $x \in \mathcal{X}$ is the set $O_x = \{\varphi(x, t; f) : t \in \mathbb{R}\}$. A point $x \in \mathcal{X}$ is a fixed point if $O_x = \{x\}$ that is $f(x) = 0$, and we use $\text{Fix}_f$ to denote the set of fixed points. The $\omega$-limit set of $x$ is the set of “limit points” such that $\omega(x) = \{y : \exists \{t_i\}_i \to +\infty, \lim_{t \to +\infty} d(\varphi(x, t_i), y; f) = 0\}$ and $\alpha$-limit is defined similarly with $t \to -\infty$.

This section has two parts. We first introduce linear flows and linear mappings, then talk about gradient-like flows which contain gradient flows as a special case.

2.2.1 Linear dynamics

Here we introduce some important properties of linear flow (and mapping) in $\mathbb{R}^d$. Given a matrix $A \in \mathbb{R}^{d \times d}$,

$$\frac{d}{dt} x(t) = Ax(t) \quad (\text{and } x_{k+1} = Ax_k)$$

which has a closed form solution $\varphi(x_0, t; A) = \exp(At)x_0$ and $\varphi(x_0, k; A) = A^kx_0$ respectively.

The long term behavior (e.g., converges to 0, diverges to infinite, or rotating) of the above systems both depend on the real part of eigenvalues of $A$. For linear flow, we denote the set of eigenvalues for the (real) matrix $A$ by

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

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/center eigenspace of $A$,

$$E^s = \{v : v \text{ is a generalized eigenvector for an eigenvalue } \lambda_i, \Re(\lambda_i) < 0\};$$

$$E^u = \{v : v \text{ is a generalized eigenvector for an eigenvalue } \lambda_{s+i}, \Re(\lambda_{s+i}) > 0\};$$

$$E^c = \{v : v \text{ is a generalized eigenvector for an eigenvalue } \lambda_{s+u+i}, \Re(\lambda_{s+u+i}) = 0\}.$$

Moreover the subspaces $E^s$, $E^u$ and $E^c$ are invariant under the flow and map with $A$.

Definition 1. We say $A \in \mathbb{R}^{d \times d}$ is hyperbolic if $E^c = \emptyset$, i.e. for all $\lambda \in \rho(A)$

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

A hyperbolic $A$ is called attracting (or repelling) if for all $\lambda \in \rho(A), \Re(\lambda) < 0$, (or $\Re(\lambda) > 0$) respectively. Finally, if $A$ is not attracting nor repelling, we call it saddle.

Therefore, there is a hyperbolic splitting of $\mathbb{R}^d$ such that $\mathbb{R}^d = E^s \oplus E^u$, and two positive real numbers $\lambda_s$ and $\lambda_u$ such that

$$\max_{1 \leq i \leq s} \Re(\lambda_i) = -\lambda_s < 0 \quad \text{and} \quad \min_{1 \leq i \leq u} \Re(\lambda_{s+i}) = \lambda_u > 0.$$
2.3 Nonlinear dynamics: Morse-Smale, gradient-like, and gradient flow

For nonlinear dynamics we first characterize some local property of fixed points.

**Definition 2 (Attracting, repelling, and saddle points).** Given a fixed point \( x \in \mathcal{X} \), with the linear approximation matrix \( A = \nabla f|_x \). A fixed point \( x \) is called *hyperbolic* (Definition 1). Similarly, a fixed point \( x \) is respectively an *attracting, repelling or saddle fixed point* if \( A \) is attracting, repelling or saddle.

We use \( \text{Attract}_f \), \( \text{Repel}_f \), and \( \text{Saddle}_f \) to denote the sets of attracting, repelling and saddle fixed points respectively. If all the fixed points are hyperbolic, \( \text{Fix}_f = \text{Attract}_f \cup \text{Repel}_f \cup \text{Saddle}_f \).

A flow with \( f \) is called a *gradient flow* if there exists a real value function \( V : \mathbb{R}^d \to \mathbb{R} \) such that \( f = -\nabla V \). However there is a more general family of dynamics called *gradient-like flow* that contains the gradient flows with mild restriction which is discussed at Proposition 28. We give a formal definition of gradient-like flow in appendix, and here only gives a sufficient condition for gradient-like flow on two dimensional manifolds.

**Proposition 3.** Let \( \mathcal{X} = \mathbb{R}^2 \). A vector field with \( f \in C^r(\mathbb{R}^2, \mathbb{R}^2) \) 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 \( \alpha \)- and \( \omega \)-limits are saddle points; and
3. each orbit has a unique fixed point as its \( \alpha \)-limit and has a unique fixed point as its \( \omega \)-limit.

We further call the function \( f \) gradient-like.

Moreover, there is a more general family of dynamic system called *Morse-Smale flows* which allows the \( \omega \)- and \( \alpha \)-limit to be cycles which is introduced in Definition 26. Gradient-like systems share several properties with gradient flows: no complicated recurrent motion and the existence of “potential function” that is decreasing along trajectories. Furthermore, by the *Fundamental theorem of dynamical system* (Theorem 25) defined in the appendix, the gradient-like flows have complete Lyapunov functions.

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

1. \( V \in C^2 \) which is smooth, and
2. \( \frac{d}{dt}V(\phi(x_0, t))|_{t=0} < 0 \) for all non fixed points of \( f \).

We use Lie derivative to simply the notion, \( \mathcal{L}_f V(x_0) \triangleq \frac{d}{dt}V(\phi(x_0, t))|_{t=0} \).

By the third condition in Proposition 3 and second property of Corollary 4, if \( V \) is a complete Lyapunov function of a gradient-like system with \( f \in C^2 \), the following are true:

1. \( \beta^* \) is a local minimal of \( V \) if and only if \( \beta^* \) is a attracting point of the flow with \( f \).
2. \( \beta^* \) is a local maximal of \( V \) if and only if \( \beta^* \) is a repelling point of the flow with \( f \).
3. There exists an partial order over the fixed points of \( f \), \( (\text{Fix}_f, \succ) \) which respect the complete Lyapunov function \( V \): if \( \beta_i \succ \beta_j \), \( V(\beta_i) > V(\beta_j) \).

\footnote{For Morse Smale system, we have a stronger notion of potential function \( \xi \)-function \[39\]. However, it often requires the flow to be smooth}
2.4 Graph with community structure and node dynamics

In this work, we consider blockmodels with two communities:

**Definition 5** (bi-blockmodel [17, 52]). Given \( p > q > 0 \), and the set of \( n \) vertices \( V \) which can be decomposed into two equal size communities \( V_1 \) and \( V_2 \), we define a weighted complete graph \( K(n, p, q) = (V, w) \) a Graph where

\[
w(u, v) = \begin{cases} p & \text{if } u, v \text{ are in the same community;} \\ q & \text{otherwise}. \end{cases}
\]

A configuration \( \sigma^{(G)} : V \to \{0, 1\} \) assigns the “color” of each node \( v \in G \) to be \( \sigma^{(G)}(v) \), equivalently \( \sigma^{(G)} \in \{0, 1\}^n \). We will usually suppress the superscript when it is clear. Moreover in a configuration \( \sigma \) we say \( v \) is red if \( \sigma(v) = 1 \) and is blue if \( \sigma(v) = 0 \). We then write the set of red vertices as \( \sigma^{-1}(1) \). We say that a configuration \( \sigma \) is in consensus if \( \sigma(\cdot) \) is the constant function (so all nodes are red or all nodes are blue), and call these two states consensus states.

Given a node \( v \) in configuration \( \sigma \) we define

\[
r_{\sigma}(v) = \frac{\|w(v, \sigma^{-1}(1))\|}{\|w(v, V)\|} = \frac{\sum_{u \in V : \sigma(u) = 1} w(v, u)}{\sum_{u \in V} w(v, u)}
\]

(5) to be its fractional weight of the red neighbors \( \sigma^{-1}(1) \).

**Definition 6.** An update function is a \( C^2 \) function \( f_{ND} : [0, 1] \to [0, 1] \) with the following properties:

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

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

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

In this work, we further assume the update function has an “S” shape— \( f \) is strictly convex in \([0,0.5] \), and strictly concave in \([0.5,1] \), and called such function smooth majority-like update function.

We define node dynamics as follows:

**Definition 7.** Given a undirected edge-weighted graph \( G = (V, w) \), an update function \( f_{ND} \) and an initial configuration \( \sigma_0 \), a node dynamic \( \text{ND}(G, f_{ND}, \sigma_0) \) is a stochastic process over configurations, \( \{S_k^{ND}\}_{k \geq 0} \) where \( S_0^{ND} = \sigma_0 \) is the initial configuration. The dynamics proceeds in rounds. At round \( k + 1 \), a node \( v \) is picked uniformly random, \( v \) updates its opinion

\[
S_{k+1}^{ND}(v) = \begin{cases} 1 & \text{with probability } f_{ND}(r_{S_k^{ND}}(v)), \\ 0 & \text{otherwise} \end{cases}
\]

where \( r_{S_k^{ND}}(v) \) is the fractional weight of the red neighbors with configuration \( S_k^{ND} \) defined in Equation (5), and we further define \( S_k = S_k^{ND} \) and \( r_k \triangleq r_{S_k^{ND}}(v) \) in the later discussion.

In this paper, we will use consensus time to study on the interaction between update function \( f \) in Definition 6 and community structure of \( G \) in definition 5. Note that we can assume \( K(n, p, q) \) with \( p + q = 1 \) which does not change the node dynamics.

**Definition 8.** The consensus time of a node dynamic \( \text{ND}(G, f_{ND}, \sigma_0) \) is a stopping time \( T(G, f_{ND}, \sigma_0) \) denoting the first time step that ND is in a consensus configuration. The maximum consensus time \( \text{ME}(G, f_{ND}) \) is the maximum consensus time over any initial configuration, \( \text{ME}(G, f_{ND}) = \max_{\sigma_0} \mathbb{E}[T(G, f_{ND}, \sigma_0)] \).
3 Main results

Theorem 9. Given a smooth majority-like function $f_{\text{ND}}$ in Definition 6, let $(S_{k}^{\text{ND}})_{k \geq 0} = \text{ND}(G, f_{\text{ND}}, \sigma_0)$ be a node dynamic over $K(n, p, q)$ where $p > q > 0$ and $p + q = 1$. There are three constants $\delta', \delta^*$ and $\delta''$ such that $0 < \delta' < \delta^* \leq \delta'' < 1$

1. If $p - q \in (0, \delta^*) \setminus \{\delta'\}$, the maximum expected consensus time
   
   $\text{ME}(K(n, p, q), f_{\text{ND}}) = O(n \log n)$.

2. If $p - q \in (\delta^*, 1) \setminus \{\delta''\}$, the maximum expected consensus time
   
   $\text{ME}(K(n, p, q), f_{\text{ND}}) = \exp(\Omega(n))$.

We prove the first part of the Theorem 9, fast convergence result, in three parts:

1. We first construct a function $\phi$ and show both the process $\phi(S_{k}^{\text{ND}})$ is a reinforced random walk with a gradient-like function and only the images of consensus states are the fixed points of the gradient-like flow (Theorem 11).

2. We next show a general theorem that a family of reinforced random walks with a gradient-like function reaches an arbitrary neighborhood of some attracting fixed point in $O(n \log n)$ with high probability under mild conditions on the perturbation (Theorem 10). Combining these we can show our process $S_{k}^{\text{ND}}$ gets close to the consensus states in $O(n \log n)$ with high probability. In section E we additionally show the process indeed hits the consensus states after arriving in the neighborhoods of consensus states.

The second part is relatively straightforward, and proved in Section E.

3.1 Fast convergence result of reinforced random walk

Informally, Theorem 10 shows if the Markov chain $S_k$ of interest can be mapped into $X_k \triangleq \phi(S_k)$ such that $X_k$ is a reinforced random walk in $\mathbb{R}^d$ with a gradient-like function $f$, then Theorem 10 shows that the behavior of the reinforced random walk with $f$ is closely related to its mean field—the flow with $f$. By the definition of the gradient-like flow with $f$, the flow (mean field) converges to the (repelling, attracting, and saddle) fixed points of $f$. The theorem, on the other hand, shows the process $X_k$ converges to an arbitrary neighborhood of an attracting fixed point fast as long as the noise around repelling and saddle points is sufficiently large. Intuitively, this noise allows the process $X_k$ to quickly escape from any non-attracting fixed point, this is unlike analogous the flow (mean field).

Theorem 10 (Hitting time of reinforced random walk). Let $S_k$ be a time homogeneous Markov chain on state space $\Omega$. If there exist constants $d \in \mathbb{N}$, $D, d_1, d_2 \in \mathbb{R}_+$, a function $\phi : \Omega \to \mathbb{R}^d$, a compact set $B \subset \mathbb{R}^d$, and $f \in C^2(\mathbb{R}^d, \mathbb{R}^d)$ with the set of fixed points $\text{Fix}_f = \{\beta_1, \ldots, \beta_m\}$ for any $\{N_i\}_{1 \leq i \leq m}$ which is a set of open sets in $\mathbb{R}^d$ with $\beta_i \in N_i$, such that

1. the flow with $f$ in (1) is a gradient-like system and $B$ is positive invariant under the flow with $f$,

2. $\{X_k\}_{k \in \mathbb{N}} = \{\phi(S_k)\}_{k \in \mathbb{N}}$ is a function of the Markov chain in $\mathbb{R}^d$, and
   
   $X_{k+1} = X_k + \frac{1}{n} (f(X_k) + U(S_k))$

   such that $X_k \in B$ with probability 1 and the noise is well-behaved: let $U_{k+1} \triangleq U(S_k)$.


(a) For all $X_k \in B$, $\mathbb{E}[U_{k+1} | S_k] = 0$,
(b) For all $X_k \in B$, $\|U_{k+1}\| \leq D$
(c) For all $X_k \in \cup_{i: \beta_i \in \text{Repel} \cup \text{Saddle}} N_i$,
\[
d_1 \mathbb{I}_d \prec \text{Cov}[U_{k+1} | S_k] \prec d_2 \mathbb{I}_d
\]
there exists $\tau = O(n \log n)$ for all $X_0 \in B$, $X_T \in \cup_{i: \beta_i \in \text{Attract}} N_i$ for some $T \leq \tau$ with high probability.

The proof is in Section 4. Note that when the reinforced function is gradient-like, by Corollary 4, there is a complete Lyapunov function for the flow, and we can use it to control the global movement of the reinforced random walk. However, the technical difficulty is how to control the behavior of the reinforced random walk around fixed points, and this is carefully analyzed in Section 4.

In Theorem 22, we show the connection between stochastic gradient descent on non-convex functions and Theorem 10. Informally Theorem 10 ensures that the reinforced random walk with the gradient of a non-convex function converges to a local minimal (attracting fixed point) in $O(n \log n)$.

### 3.2 Phase portrait

To apply Theorem 10, we need to show our node dynamics on bi-blockmodel are time homogeneous Markov chains which can be mapped to $\mathbb{R}^2$ as a reinforced random walk with gradient-like function. In this section, we define such a mapping $\phi$, and show $(\phi(S^\text{ND}_k))$ itself is a Markov chain which is stronger than reinforced random walk defined in Equation (3).

Because of the symmetry of $S^\text{ND}_k$ and bi-blockmodel defined in Definition 5, we define a mapping from $\sigma = \{0, 1\}^d$ to the numbers of red nodes in those two communities, $\text{Pos} : \{0, 1\}^n \rightarrow [n/2] \times [n/2]$ and $\phi = \text{pos} = 2\text{Pos}/n$ such that
\[
\text{Pos}(\sigma) = (\|V_1 \cap \sigma^{-1}(1)\|, \|V_2 \cap \sigma^{-1}(1)\|) = \left(\sum_{v \in V_1} \sigma(v), \sum_{u \in V_2} \sigma(u)\right).
\]

Then we define the process: for all $k \in \mathbb{N}_0$
\[
X^\text{ND}_k \triangleq \phi(\sigma_k),
\]
which is a function of Markov chain $S^\text{ND}_k$, and it is a two-dimensional reinforced random walk, as defined in [3], with $F_{ND} : [0, 1]^2 \rightarrow \mathbb{R}^2$ defined as follows:
\[
F_{ND}(x_1, x_2) \triangleq (f_{ND}(p x_1 + q x_2) - x_1, f_{ND}(p x_2 + q x_1) - x_2).
\]

Moreover, $X^\text{ND}_k$ itself is also a Markov chain and the consensus states $0^n$ and $1^n$ is mapped to $(0, 0)$ and $(1, 1)$ respectively. As a result, we will use $X^\text{ND}$ and $S^\text{ND}$ interchangeably as our node dynamics.

As mentioned in section 2.1, we summarize three closely related dynamics here:

1. $\{S^\text{ND}_k\}_{k \in \mathbb{N}_0}$ the stochastic sequence on state space $\Omega = \{0, 1\}^n$, the dynamics in theorem 9, $\text{ND}(G, f_{ND}, \sigma_0)$ with $G = K(n, p, q)$,
2. $\{X^\text{ND}_k\}_{k \in \mathbb{N}_0}$ the stochastic sequence on state space $\{0, 2/n, 4/n, \ldots, 1\}^2 \subset \mathbb{R}^2$ defined in Equation (7) which is a function of Markov chain and a reinforced random walk with $F_{ND}$.
3. \( \{ x^{\text{ND}}(t) \}_{t \in \mathbb{R}} \) the deterministic flow in \( \mathbb{R}^2 \) associated with \( F_{\text{ND}} \).

Markov chain \( \{0,1\}^n \) Reinforced random walk in \( \{0, 2/n, 4/n, \ldots, 1\}^2 \) Flow in \( \mathbb{R}^2 \)

\[
\begin{align*}
S_k^{\text{ND}} & \quad \phi = \text{pos} \quad X_k^{\text{ND}} \quad \text{Theorem} \, 10 \quad x^{\text{ND}}(t)
\end{align*}
\]

**Theorem 11** (Phase portrait). Given \( f_{\text{ND}} \) and \( p, q \) in the Node Dynamics defined in Theorem 9, there exist three constants \( 0 < \delta' < \delta^* \leq \delta'' < 1 \) such that the flow with \( F_{\text{ND}} \) defined in \( (8) \) 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 \( p - q \in (\delta^*, 1) \setminus \{\delta''\} \), \( F_{\text{ND}} \) has an attracting fixed point \( \beta_a \neq (0,0), (1,1) \).

A more detailed characterization of \( \delta', \delta^* \) and \( \delta'' \) is in Theorem 18.

### 3.3 From neighborhood of fixed points to the fixed points

In Section E, we complete the proof of Theorem 9. For the first part of Theorem 9, we want to show the process indeed reach a consensus state \( 0^n \) or \( 1^n \) fast. Theorem 10 ensures our process reach a neighborhood of those consensus states which is not enough. In section E, we show after reaching arbitrary neighborhoods of consensus states the process can reach a consensus state in \( O(n \log n) \) steps with constant probability. We achieve this by constructing a coupling between our process and a birth-and-death process in Lemma 45 and an upper bound for expected hitting time of birth-and-death in Lemma 47 and 48.

Finally, for the second part, Theorem 11 shows the existence of an attracting fixed point \( \beta_a \) other than consensus states. By a standard argument (Lemma 46) if the process starts at \( \beta_a \), the probability of leaving \( Q' \) a neighborhood of \( \beta_a \) in \( n \) steps is exponentially small. Therefore the expected time of reach consensus states is \( \exp(\Omega(n)) \).

### 4 Reinforced random walks of gradient-like function and the proof of Theorem 10

This section is concerned with proving Theorem 10. To show the process reach a neighborhood of an attracting fixed point fast, we need to show two parts: locally, the process does not stuck at any small neighborhood; globally, the process to progress without making circle or having complicated recurrent behavior.

For global characterization, because the flow is gradient-like, by Corollary 4 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 \( X_k \). Locally, for each fixed point \( \beta_i \in \text{Fix}(f) \), we define a small neighborhood \( N_i \) around it containing no additional fixed points, and we set the index such that \( \beta_i \in N_i \). There are two cases: either \( x \in \mathcal{X} \setminus (\cup_i N_i) \), and we say \( x \) is a regular point. In this case the complete Lyapunov function \( V \) has large (linear) decrements. Otherwise,

\(^2\)We need to extend the domain of \( F_{\text{ND}} \) into \( \mathbb{R}^2 \) which is defined in Section 5 and call it as \( \tilde{F}_{\text{ND}} \).
$x \in N_i$ for some $i$, we say that $x$ is a \textit{neighborhood point} and $V$ decrements increasingly slowly as it approaches the fixed point $\beta_i$.

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.

\textbf{Lemma 12} (regular points). If $X_0 \notin \bigcup N_i$, there exists $i$ and $T = O(n)$ such that $X_T \in N_i$ and $V(\beta_i) < V(X_0)$ with probability $1 - o(1)$.

The next lemma says that as long as $\beta_i$ 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.

\textbf{Lemma 13} (non attracting fixed points). If $X_0 \in N_i$ and $\beta_i$ is not an attracting point, there exists $\delta > 0$ such that $\tau = O(n \log n)$, $X_T \notin N_i$, and $V(\beta_i) > \delta + V(X_0)$ for some $T \leq \tau$ with high probability.

This is proved in the appendix. The proof relies heavily on our main technical lemma, Lemma 14, which shows that the processes leaves saddle points (or unstable fixed points). Lemma 14 is proved in Section 4.1.

\textit{Proof of Theorem 10}. Combining the above two characterizations, we can study the process in two alternating stages.

1. Given an initial condition $x_0 \in B$ where $B$ is compact and positive invariant, if $x_0 \notin \bigcup_i N_i$, it converges to some $N_i$ in $O(n)$ with high probability by Lemma 12.

2. If $\beta_i$ is not an attracting point by Lemma 13, the process leaves the region $N_i$ and $V(x) < V(\beta_i) - \delta$ in $O(n \log n)$ time with high probability.

3. After leaving $N_i$, by Lemma 12, the process converges to $N_j$ a neighborhood of another fixed point $\beta_j$ where $V(\beta_j) < V(\beta_i)$ 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.

\hfill $\square$

\section*{4.1 Escaping local saddle points}

In this section we state and prove our main technical lemma, which shows that our process will quickly leave the neighborhood of a saddle point or unstable fixed point.

\textbf{Lemma 14} (Leaving non-attracting fixed point). Given the setup in Theorem 10, there are a sufficiently small constant $r > 0$ and $\tau_i = O(n \log n)$, such that if the process starting in $N_i$ for some non-attracting fixed points, after $T_i \leq \tau_i$,

$$\Pr[X_{T_i} \in B(\beta_i, 3r/4) \setminus B(\beta_i, r/2)] = 1 - o(1).$$
Figure 1: The solid blue circle represents $N_i$, dash circles are different phases of the process, non-grid region are those $\|Z^u\| \geq 8\|Z^s\|$, and the other solid circle is $B(\beta_i, 3r/4)$. For Lemma 14, we partition the process around saddle point $\beta_i$ into $O(\log n)$ phases: In phase 0 the process hits $N_i$ (the solid blue circle), and Lemma 15 shows the process hits $\|Z\| \leq l_1$ (the smallest dash circle), in $O(n \log n)$ with probability $1 - o(1)$. In the phase 1, by Lemma 16, after hitting $\|Z\| \leq l_1$, the process enters $\|Z_u\| \geq l_1$ (the non-grid region of the bigger dash circle) in $O(n \log n)$. The Lemma 17 shows if $\|Z^u_0\| \geq l_j$ the process will enter the $\|Z^u\| \geq l_{j+1}$ (the non-grid region of the next bigger dash circle) in additional $O(n)$ times. Finally, the process leaves the region $B(\beta_i, 3r/4)$.

Roughly, around the saddle point $\beta_i$ the dynamics can be approximated by a linear flow with $A = \nabla f(\beta_i)$ defined in section 2.2.1. That is the process is expanding in the subspace $E^u$, and contracting in subspace $E^s$ with respect to $A$. However, because of the non-linearity of the process, there is a quadratic error term $O(\|Z_k\|^2)$. To handle this, we partition the process into $O(\log n)$ phases illustrated in Figure 1 such that as long as the difference between $\|Z^u_k\|$ and $\|Z^s_k\|$ is not too large, the errors are comparable (and small).

The proof has three parts. Intuitively, Lemma 15 shows the magnitude in contracting subspace decrease rapidly. Lemma 16 shows if the process is very close to or at $\beta_i$, the noise of the process can ensure the unstable part of the process can be $\Omega((\log n)^{1/3}/\sqrt{n})$ far away from $\beta_i$ in $O(n \log n)$ times. Finally, Lemma 17 shows if the unstable part of the process is $\Omega((\log n)^{1/3}/\sqrt{n})$ away from $\beta_i$, the unstable part double in $O(n)$ time with probability $1 - \exp(-\Omega(\sqrt{\log n})) = 1 - o(1/\log n)$.

**Proof.** Because the fixed points of the Gradient-like system are hyperbolic we can rewrite the process around $B(\beta_i, r)$ as,

$$Z_{k+1} - Z_k = \frac{1}{n}(AZ_k + O(\|Z_k\|^2) + \text{noise}) \tag{9}$$

where $A = \nabla f(\beta_i)$ is hyperbolic and $Z_k = X_k - \beta_i$. Note that here we use $O(\|Z_k\|^2)$ to denote an error vector such that each coordinate of this vector is $O(\|Z_k\|^2)$. Furthermore, given the matrix $A$, we can decompose the tangent space $\mathbb{R}^d$ into the stable and the unstable subspaces $E^s$ and $E^u$ with respect to $A$ (Section 2.2.1). Let $P^u$ and $P^s$ be the projection operators for $E^s$ and $E^u$ respectively. Without loss of generality, we consider $\beta_i$ to be a saddle point.
We can consider the following two (correlated) processes which are updated by the original process decomposed into operating on the (not necessarily orthogonal) spaces $E^s$ and $E^u$:

\begin{align}
Z^u_{k+1} - Z^u_k &= \frac{1}{n} (AZ^u_k + O(\|Z_k\|^2) + \text{noise}^u) \in E^u \tag{10} \\
Z^s_{k+1} - Z^s_k &= \frac{1}{n} (AZ^s_k + O(\|Z_k\|^2) + \text{noise}^s) \in E^s \tag{11}
\end{align}

where $\text{noise}^u \triangleq P^u U_{k+1} \in E^u$ and $\text{noise}^s \triangleq P^s U_{k+1} \in E^s$. We call $Z^u_k$ and $Z^s_k$ unstable component and stable component of the process $Z_k$ respectively.

If we can show after $\tau_1 = O(n \log n)$ steps both in the stable manifold (11) we have $\|Z^s_{\tau_1}\| \leq r/4$ and in unstable manifold (10) we have $\|Z^u_{\tau_1}\| \geq 3r/4$, the $\|Z_{\tau_1}\| \geq \|Z^u_{\tau_1}\| - \|Z^s_{\tau_1}\| = r/2$ which completes the proof.\footnote{Although the process $Z_k$ may even leave $B(\beta, r)$ before $\tau_1$ such that Equation (9) does not hold anymore, we can define another process by Equation (9) and couple it with the original process when the process is in $B(\beta, 3r/4)$. We analyze the new process instead and show it leaves $B(\beta, 3r/4)$ with high probability. Therefore the original process also leave it with high probability.}

Let $\lambda_n = \min \{ \Re(\lambda_i) \} > 0$ which is minimum real part of eigenvalue of $A$ in $E^u$. We define a length $J = O(\log n)$ sequence

\begin{equation}
l_1 = \frac{(\log n)^{1/3}}{\sqrt{n}}, \quad l_{j+1} = 2l_j \text{ for } j = 1, 2, \ldots, J - 1, \text{ and } l_J = 3r/4. \tag{12}
\end{equation}

With the sequence $(l_j)$, we partition the processes in $B(\beta_i, 3r/4)$ into $O(\log n)$ phases, and say the process $Z_k^u$ is in phases $j$ if and only if $l_{j-1} \leq \|Z_k^u\| < l_j$ and $\|Z_k^u\| \leq \|Z_k^u\|/8$.

First in Lemma 15 we show either the stable component $\|Z^s\|$ is smaller than the unstable component $\|Z^u\|$ or enters the phase 0, $\|Z\| \leq l_1$ in $O(n \log n)$ time with high probability.

Secondly, by Lemma 16 suppose the process is at phase 0, $\|Z_0\| \leq l_1$, the process reach phase 1 within $O(n \log n)$ steps with probability $1 - o(1)$.

Finally, by Lemma 17 starting at phase $j$, the process reach phase $j + 1$ with in $O(n)$ steps with probability $1 - \exp(-\Omega(\sqrt{\log n})) = 1 - o(1/ \log n)$. Thus the proof completes by taking union bound on these $J = O(\log n)$ phases.

Due to the space constrain, we put all of the proofs of the following lemmas in to the appendix.

4.1.1 Phase 0: decreasing the stable component

Lemma 15 (Phase 0). If $X_0 \in N_i$, in time $\tau_0 = O(n \log n)$, there exists $T_0 \leq \tau_0$ such that $\|Z^u_{T_0}\| \geq 8\|Z^s_{T_0}\|$ or $\|Z_{T_0}\| \leq l_1$ with probability $1 - o(1)$.

4.1.2 Phase 1: leaving the fixed point

For Lemma 16, because the drift of the process is too small, we use the anti-concentration of noise (Lemma 42) to show in expectation it can reach $l_1 = \Omega((\log n)^{1/3}/\sqrt{n})$ after $O(n \log n)^{2/3})$ steps. By Markov inequality, we show it will happen in $O(n \log n)$ with probability $1 - o(1)$.

Lemma 16 (Phase 1). If $\|Z_0\| \leq l_1$, there are $\tau_1 = O(n \log n)$ and $T_1 \leq \tau_1$ such that $\|Z^s_{T_1}\| \geq 2l_1$ and $\|Z^u_{T_1}\| = o(l_1)$ with probability at least $1 - o(1)$.\footnote{Although the process $Z_k$ may even leave $B(\beta, r)$ before $\tau_1$ such that Equation (9) does not hold anymore, we can define another process by Equation (9) and couple it with the original process when the process is in $B(\beta, 3r/4)$. We analyze the new process instead and show it leaves $B(\beta, 3r/4)$ with high probability. Therefore the original process also leave it with high probability.}
4.1.3 Phase $j$: amplifying the unstable component

To the end we want to show $\|Z_k^u\|$ in (10) increases rapidly which depends on three things: the linear part $AZ_k^u$ is large, the nonlinear term $O(\|Z_k\|^2)$ is small and the noise, noise$^u$, is small. However, $O(\|Z_k\|^2)$ depends both on $Z_k^u$ and $Z_k^s$, so we need to upper bound the value of $\|Z_k^s\|$ as well. Therefore in contrast to Lemma 15 to prove the $\|Z_k^u\|$ reach large value fast, we use induction because to control the process multiple quantities $\|Z_k^u\|/\|Z_k^s\|$, $\|Z_k\|$, and $\|Z_k^u\|$, and it requires more delicate argument than optional stopping time theorem.

For non-linearity because $f \in C^2$ is smooth we can upper bound the quadratic values $\|Z_k\|^2$ by $\|Z_0\| = o(l_j)$ for all $0 \leq k \leq T$ with high probability. However, the standard Chernoff bound and union bound are not enough, so use a more advanced tail bound for the maximum deviation (Theorem [39]). For the noise part, condition on $\|Z_k\|^2$ being small we use linear approximation of $f$ to study two aspect to the Doob martingale $Y_k = E[Z_T|Z_0, \ldots, Z_k]$: 1) the effect variance $\sum c_i^2$ is small and 2) the expectation $Y_0 = E[Z_T]$ is nice.

**Lemma 17** (Phase $j > 1$). If $\|Z_0^u\| \leq \frac{1}{8}\|Z_0^s\|$ and $l_j \leq \|Z_0^u\| \leq l_{j+1}$, $\tau_j = O(n)$ such that $\|Z_k^s\| \leq \frac{1}{8}l_{j+1}$ and $\|Z_{\tau_j}^u\| > l_{j+1}$ with probability $1 - \exp(-\Omega(\sqrt{\log n}))$.

Note that in contrast to Lemmas 15 and 16 which show upper bounds for hitting times, this lemma characterizes the behavior of $Z$ at time $\tau_j$.

5 Phase portrait

In this section, we prove Theorem 11 which will follow immediately from theorem 18, by analyzing the fixed points of the function $F_{\text{ND}}$ defined in (8). We can classify the fixed points into three types: symmetric, anti-symmetric and eccentric. Lemma 19 characterizes the property of symmetric fixed points; Lemma 20, anti-symmetric fixed points; and Lemma 21, eccentric fixed points. The following section introduces the symmetry property of the flow on $F_{\text{ND}}$ and Theorem 18 is proved in the next one.

5.1 Setup and examples

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

$$0 = f_{\text{ND}}(px_1 + qx_2) - x_1,$$

$$0 = f_{\text{ND}}(px_2 + qx_1) - x_2.$$ (13)

Denote the solutions of equation (13) as

$$\gamma_1 = \left\{ (x_1^{(1)}, x_2^{(1)}) \in [0, 1]^2 : x_1^{(1)} = f_{\text{ND}} \left( px_1^{(1)} + qx_2^{(1)} \right) \right\},$$

$$\gamma_2 = \left\{ (x_1^{(2)}, x_2^{(2)}) \in [0, 1]^2 : x_2^{(2)} = f_{\text{ND}} \left( px_2^{(2)} + qx_1^{(2)} \right) \right\}.$$ (14)

Note that the system of Equation 13 is symmetric with respect to two axes $x_1 = x_2$ and $x_1 + x_2 = 1$, so we define four disjoint regions of $[0, 1]^2$:

$$R_1 = \{ (x_1, x_2) \in [0, 1]^2 : x_1 < x_2 \text{ and } x_1 + x_2 < 1 \},$$

$$R_2 = \{ (x_1, x_2) \in [0, 1]^2 : x_1 < x_2 \text{ and } x_1 + x_2 > 1 \},$$

$$R_3 = \{ (x_1, x_2) \in [0, 1]^2 : x_1 > x_2 \text{ and } x_1 + x_2 < 1 \},$$

$$R_4 = \{ (x_1, x_2) \in [0, 1]^2 : x_1 > x_2 \text{ and } x_1 + x_2 > 1 \}.$$ 

With this symmetric property, we classify the fixed points of (13) 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 2 shows some examples of a dynamic with different \(p, q\).

Figure 2: In Theorem 18 there are three critical values \(\delta_{\text{symm}}, \delta_{\text{ecce}}\) and \(\delta_{\text{anti}}\). 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 20. In case (b), the \(p - q\) is bigger such that there are two extra saddle anti-symmetric fixed points. For some specific update function \(f_{\text{ND}}\) 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 21. Finally in case (d), the \(p - q\) is big enough such that the antisymmetric fixed points become attracting which is characterized in Lemma 20.

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

\[
\tilde{F}_{\text{ND}}(x) = \begin{cases} 
m_1 x & \text{if } x < 0 \\
_{\text{ND}}(x) & \text{if } x \in [0, 1] \\
m_1(x - 1) + 1 & \text{if } x > 1
\end{cases}
\]

We can have \(\tilde{F}_{\text{ND}}\) by using \(\tilde{F}_{\text{ND}}\) in (8) instead of \(f_{\text{ND}}\).

5.2 Proof of Theorem 11

The following theorem is a detailed characterization of the flow \(x^{\text{ND}}\) with \(F_{\text{ND}}\), and Theorem 11 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 18** (Phase portrait). Fix the flow \(x^{\text{ND}}\) with \(p, q\) and \(F_{\text{ND}}\) defined in (8), depending on the property of \(f_{\text{ND}}\) there are two situations

1. If there exists \(\delta_e\) such that equation (13) 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 19 and \(\delta_{\text{anti}}\) is defined in Lemma 20 and \(\delta_{\text{ecce}}\) defined in Lemma 21 such that there are three cases:

---

4To make \(f_{\text{ND}} \in C^2(\mathbb{R}, \mathbb{R})\), 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 \(C^2\) function moreover it can be arbitrary close to the above definition if we take \(\epsilon\) small enough.
(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 19 and $\delta_{\text{anti}}$ is defined in Lemma 20, 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 18.

Lemma 19 (symmetric fixed points). Given $F_{\text{ND}}$ with $p, q$ and $f_{\text{ND}}$, 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 (13) only has the above three fixed points.

Lemma 20 (anti-symmetric fixed points). Given $F_{\text{ND}}$ with $p, q$ and $f_{\text{ND}}$ and $\delta_{\text{symm}}$ in Lemma 19, there exists $\delta_{\text{anti}} > \delta_{\text{symm}}$ such that there are two cases for the anti-symmetric fixed points in Equation (13) 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 20, 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 $f_{\text{ND}}$ such that the system has extra stable eccentric fixed points before the anti-symmetric fixed points become stable, e.g. Figure 2. Though we can use simulation to estimate the phase space, the following lemma shows: Given $f_{\text{ND}}$, suppose there exists $\delta_{e} < \delta_{\text{anti}}$ such that the system with $\delta_{e} = p_{e} - q_{e}$ in Equation (13) has an eccentric fixed point. Then there exists $\delta_{\text{ecce}} < \delta_{\text{anti}}$ such that for all $\delta_{e}'$ such that $\delta_{\text{ecce}} < \delta_{e}' < \delta_{\text{anti}}$ the system (13) has attracting eccentric fixed points fixed points. By symmetry, we only state the result in $R_{2}$.

Lemma 21 (eccentric fixed points). Given $F_{\text{ND}}$ with $p, q, f_{\text{ND}}, \delta_{\text{symm}}$ and $\delta_{\text{anti}}$ in Lemma 19, 20, if there exists $\delta_{e} < \delta_{\text{anti}}$ such that equation (13) 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 (13) 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 $\delta_{e}$ such that the there exists a eccentric fixed point and anti-symmetric saddle points.
Now we are ready to prove Theorem 18.

**Proof of Theorem 18.** The main statement of theorem is proved by Lemma 20 and 21. Now we prove the case 1 and 2 are indeed gradient-like. Because it's only a two dimensional system, by Proposition 3, 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 20, 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 20 and 21, the system only have 5 fixed points. Secondly, the saddle point have stable manifold 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.

6 Stochastic gradient descent and Theorem 10

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 method by stalling their process.

Our analysis in Section 4 can be applied to these problems. Formally, given an objective function $F : \mathbb{R}^d \to \mathbb{R}$, an popular heuristic to minimize $F$ is by gradient descent method:

$$x_{t+1} = x_t - \eta \nabla F(x_t),$$  \hspace{1cm} (15)

The gradient descent is well-studied when the objective function is convex: for any constant $\epsilon$, $|F(x_t) - \min_{x \in \mathbb{R}^d} F(x)| \leq \epsilon$ in time $O(1/\eta)$. In this section, we want to study the convergence property when $F$ is non-convex. In particular, we are interested in the time complexity with respect to the step size $\eta$.

6.1 Bounded stochastic gradient descent algorithm

We now state one general framework to converge to local minima.

**Algorithm 1:** Bounded Stochastic Gradient Descent Algorithm

<table>
<thead>
<tr>
<th>Result: Finding a local minimal value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> An objective function $F : \mathbb{R}^d \to \mathbb{R}$, the step length $\eta$, the running time $T$, and the initial point $x_1$</td>
</tr>
<tr>
<td><strong>Output:</strong> A point $x \in \mathbb{R}^d$</td>
</tr>
</tbody>
</table>

1 for $t = 1, 2, \ldots, T$ do
2 Sample a perturbation $U_{t+1}$ with properties defined in Theorem 22
3 $x_{t+1} = x_t - \eta (\nabla F(x_t) + U_{t+1})$
4 end

Using the same argument for Theorem 10 and Proposition 28 we have:

**Theorem 22** (Bounded Stochastic Gradient Descent Algorithm). Given a constant $d$, an objective function $F \in C^3(\mathbb{R}^d, \mathbb{R}^d)$, a compact set $B \subset \mathbb{R}^d$ which is positive invariant for Equation (15), constants $D, d_1, d_2 > 0$, such that
1. The objective function $F$ has

(a) a continuous third derivative.

(b) a constant number of fixed points in $B$, \( \{ \beta \in B : \nabla F(\beta) = 0 \} \). Moreover, these fixed points are non-degenerate: $\nabla^2 F(\beta)$ is invertible.

2. The perturbation of the process in Algorithm 1 satisfies

(a) $X_t \in B$ with probability 1.

(b) $E[U_{t+1} | x_t] = 0$ for all $x_t \in B$,

(c) $\|U_{t+1}\| \leq D$,

(d) For all $x_t \in \cup_{\beta : \nabla F(\beta) = 0} B(\beta, \epsilon)$, $d_1 \mid d \prec \Cov[U_{k+1} | X_k] \prec d_2 \mid d$

where $B(\beta, r)$ is a ball in $\mathbb{R}^d$, \( \{ y : \| y - \beta \| \leq r \} \).

Then for any $\epsilon > 0$, there exist $\eta > 0$ and $T = O\left( \frac{\log 1/\eta}{\eta} \right)$ such that for all initial points $x_1 \in B$ $\|x_t - x^*\| \leq \epsilon$ for some local minimal $x^*$ and $t \leq T$ with high probability.

Note that though this theorem only shows the hitting time is $O(\frac{\log 1/\eta}{\eta})$ with high probability, with Lemma 46 it is not hard to show the process stays in an arbitrary neighborhood of a local minimal after hitting a neighborhood of the local minimal.

6.2 Related work

For the time complexity with respect to the step size $\eta$, this framework contains several previous results as special cases, and provides a tighter convergence time upper bound. For example, Ge et al. [25] propose the following algorithm:

**Algorithm 2: Noisy Gradient Descent [25]**

| 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 $x \in \mathbb{R}^d$ |

1. for $t = 1, 2, \ldots, T$ do
2. Sample a perturbation $U_{t+1} \sim S^{d-1}$ (Random point on unit sphere)
3. $x_{t+1} = x_t - \eta (\nabla F(x_t) + U_{t+1})$
4. end

They show the convergent time to constant neighborhood of some local minima is $O(1/\eta^2)$ which is weaker than Theorem [22] when the objective function satisfies our condition.

Similarly, Jin et al. [32] proposes a perturbed gradient descent algorithm:
Algorithm 3: Perturbed Gradient Descent [32]

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 $x \in \mathbb{R}^d$

1 for $t = 1, 2, \ldots$ do
2  if $\|\nabla F(x_t)\|$ is small then
3    $U_{t+1} \sim S^{d-1}$
4  else
5    $U_{t+1} = 0$
6  end
7  $x_{t+1} = x_t - \eta (\nabla F(x_t) + U_{t+1})$
8 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 22 when the objective function satisfies our condition.

Remark 23. Here we put some comparison between Theorem 22 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 algorithm. Instead of requiring the perturbation to be a uniform point in the unit sphere, our result only requires the noise is bounded and the covariance matrix is positive definite (Theorem 22).

3. In gradient flow, the stable and unstable manifold 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 extends to reinforced random walks with non-gradient flows.

On the other hand, our result doesn’t handle some aspects in Ge et al. [25], Jin et al. [32]:

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

2. We do not work out the dependency of running time on the dimension (and several parameters), although we believe our analysis is dimension free.

3. The number of fixed points in our work is constant, and they avoid this condition by assuming a uniform lower bound of positive eigenvalue of all saddle points which ensures a universal constant improvement after escape from any saddle points.

References


A Primer of dynamical system

A.1 Fundamental theorem of dynamical system

An opposite concept to “recurrence” is transit. How do 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 \( \Psi \) 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 \( x \in \mathcal{X} \) is a **fixed point** if \( O_x = \{x\} \) that is \( f(x) = 0 \), and we use \( \text{Fix}_f \) to denote the set of fixed points.

**Periodic point** A point \( x \in \mathcal{X} \) is a **periodic point** of \( f \) if \( \exists T \geq 0 \) such that \( \varphi(x, T; f) = x \), and we use \( \text{Per}_f \) to denote the set of periodic points.

**\( \omega \)-recurrent** For other non-periodic points \( x \in \) the long term behavior can be characterized as \( \omega \)-limit set of \( x \): \( \omega(x) = \{y : \exists t_1 \to +\infty, \lim_{t \to \infty} d(\varphi(x, t_1), y) = 0\} \), and we call \( x \) \( \omega \)-recurrent if \( x \in \omega(x) \) If we change \( +\infty \) to \( -\infty \) in above definition, it called \( \alpha \)-limit set \( \alpha(x) \) of \( x \). We call \( L_f = \bigcup_{x \in \mathcal{X}} \omega(x) \cup \bigcup_{x \in \mathcal{X}} \alpha(x) \) the limit set of \( f \).

**Chain recurrent** An \( \epsilon \)-chain of length \( T \) from a point \( x \) to \( y \) is a sequence of points \( (x_{\ell})_{0 \leq \ell \leq n} \) and a sequence of time \( (t_{\ell})_{1 \leq \ell \leq n} \) such that \( x_0 = x, x_n = y \), and \( d(\varphi(x_{i-1}, t_i), x_i) < \epsilon \) for \( 1 \leq \ell \leq n \) with \( t_\ell \geq 1 \) and \( \sum_{\ell=1}^T t_\ell = T \). We define a relation \( \sim_{CR} \) on \( CR_f \). Similar to \( \omega \)-limit we define \( \Omega^+(x) = \bigcap_{\epsilon > 0} \bigcap_{T > 0} \{y : \exists \alpha, T \text{ chain from } x \to y\} \), and a point \( x \) is said to be **chain recurrent** for the flow \( f \) if \( x \in \Omega^+(x) \). The set of chain recurrent points of \( f \) is called the chain recurrent set of \( f \) denoted as \( CR_f \). We say \( x \sim_{CR} y \) if and only if \( x \in \Omega^+(y) \) and \( y \in \Omega^+(x) \).

It is not hard to show

\[
\text{Fix}_f \subseteq \text{Per}_f \subseteq L_f \subseteq CR_f \subseteq \mathcal{X}
\]

Now we are ready to state the **fundamental theorem of dynamical system**, 

**Definition 24** (Complete Lyapunov function). Let \( \varphi(\cdot, \cdot; f) \) be a flow with \( f \) on a metric space \( \mathcal{X} \). A **complete Lyapunov function** for \( f \in C^0 \) is a continuous function \( \Psi : \mathcal{X} \to \mathbb{R} \) such that

1. For all \( s < t \) and \( x \in \mathcal{X} \setminus CR_f \), \( \Psi(\varphi(x, s; f)) > \Psi(\varphi(x, t; f)) \),

2. for all \( x, y \in CR_f \), \( x \sim_{CR} y \) if and only if \( \Psi(x) = \Psi(y) \).

3. \( \Psi(CR_f) \) is nowhere dense subset of \( \mathbb{R} \).

By constructing a complete Lyapunov function, Conley shows:

**Theorem 25** (Fundamental theorem of dynamical system [44]). Every flow on a compact metric space has a complete Lyapunov function \( V : \mathcal{X} \to \mathbb{R}_+ \).

One interpretation of this theorem is that the space of the dynamics can be decomposed into two parts: points exhibiting a particular type of recurrence, and points proceed in a gradient-like fashion.
A.2 Morse-Smale and gradient-like dynamics

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

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

$$W_{loc}^s(x, U, f) \triangleq \{ y \in U : \varphi(y, t; f) \in U, \forall t > 0 \text{ and } d(\varphi(y, t; f), x) \to 0 \text{ as } t \to \infty \}$$

$$W_{loc}^u(x, U, f) \triangleq \{ y \in U : \varphi(y, t; f) \in U, \forall t < 0 \text{ and } d(\varphi(y, t; f), x) \to 0 \text{ as } t \to -\infty \}$$

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

**Definition 26** (Morse-Smale flow). Let $\varphi(\cdot, \cdot; f)$ be a flow on $X = \mathbb{R}^d$. $\varphi$ 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 $i = 1, \ldots, l$
2. $\mathcal{C}R_f = \text{Per}_f$
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.

Note that the gradient flow is a special case of gradient-like flow

**Definition 27** (Gradient flow). A flow $\varphi(\cdot, \cdot)$ on $\mathbb{R}^d$ is called gradient flow if there is a real valued function $V : \mathbb{R}^d \to \mathbb{R}$ such that

$$\frac{d}{dt} \varphi(x, t) = -\nabla V(x).$$

**Proposition 28.** Let $V : \mathbb{R}^d \to \mathbb{R}$ be a $C^2$ function such that each critical point is nondegenerate, i.e., at each point $\beta$ where $\nabla V(x) = 0$, the matrix of second partial derivatives $\nabla^2 V(\beta)$ has nonzero determinant. Then the gradient flow with $V$ has all the fixed points are hyperbolic and the chain recurrent set for the flow equals the set of fixed points.

The above proposition shows the (non-degenerate) gradient flows are Morse-Smale system if and only if the stable and unstable manifolds are transverse.

Let $\{\beta_1, \ldots, \beta_m\} = \text{Fix}_f$ be the set of fixed point of $f$, and $W^s_i$ and $W^u_i$ be the stable and unstable manifold associated to $\beta_i$. The Morse-Smale system has the following property.

**Lemma 29.** Let $f$ be a Morse-Smale system on $X$. Let $\beta_i \succ \beta_j$ mean there is a trajectory not equal to $\beta_i$ or $\beta_j$ whose $\alpha$-limit set is $\beta_i$ and whose $\omega$-limit set is $\beta_j$. 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 $\dim W^u_i \geq W^u_j$
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. Furthermore, by the Fundamental theorem of dynamical system \(25\) we have\(^5\)

**Corollary 30** (Theorem 12 in Akin [2]). If \(f \in C^2\) is a Morse-Smale system then there exists a complete Lyapunov function \(V : X \to \mathbb{R}\) such that

1. \(V \in C^2\) is smooth.
2. \(\frac{d}{dt} V(\varphi(x_0,t))|_{t=0} < 0\) for all non fixed points of \(f\).

We use Lie derivative to simplify the notion, \(\mathcal{L}_f V(x_0) \triangleq \frac{d}{dt} V(\varphi(x_0,t))|_{t=0}\).

**B Basic math**

**B.1 Markov chain**

Let \(M = (X_t, P)\) be a discrete time-homogeneous Markov chain with a finite state space \(\Omega\) 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

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

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

By the Markov property, the expected hitting time \(\tau_Q(x) = \mathbb{E}_M[T_Q(x)]\) can be written as linear equations.

\[ \mathcal{L} \tau_Q(x) = -1 \text{ where } x \notin Q \]
\[ \tau_Q(x) = 0 \text{ where } x \in Q \]

**Corollary 31** (Maximum principle [20]). Given a Markov chain \(M\) with state space \(\Omega\) and a set of states \(Q \subseteq \Omega\), suppose \(s_Q : \Omega \to \mathbb{R}\) is a non-negative function satisfying

\[ \mathcal{L}s_Q(x) \leq -1 \text{ where } x \notin Q, \]
\[ s_Q(x) \geq 0 \text{ where } x \in Q. \]

(16)

Then \(s_Q(x) \geq \tau_Q(x)\) for all \(x \notin Q\).

\(^5\)For Morse Smale system, we have a stronger notion of potential function \(\xi\)-function [39]. However, it often requires the flow to be smooth.
When Markov chain has some symmetric property, we can project the original Markov chain to other simpler Markov chain. Formally, \( M \) is lumpable with respect to the partition \( \Omega_Y = (y_1, \ldots, y_k) \) where \( \Omega_X = \bigcup_{i \in [k]} y_i \) if and only if, for any subsets \( y_i \) and \( y_j \) in the partition, and for any states \( x, x' \) in subset \( y_i \),

\[
\sum_{x'' \in y_j} P(x, x'') = \sum_{x'' \in y_j} P(x', x'').
\]

Alternatively we say \( M \) is lumpable with respect to a function \( g : \Omega_X \rightarrow \Omega_Y \) if \( X \) is lumpable with respect to the partition \( \{ g^{-1}(y) : y \in \Omega_Y \} \), and define the lumped process of \( X \) under \( g \) as the Markov chain \( Y := (Y_t)_{t \in \mathbb{N}_0} \) defined by \( Y_t := g(X_t) \).

### B.2 Linear algebra

In this section, we state some basic results of linear algebra. Given symmetric matrices \( A, A' \in \mathbb{R}^{d \times d} \), \( A < A' \) denotes \( A' - A \) is positive definite.

**Definition 32** (Majorize [31]). Given two real-valued sequences \( x, y \in \mathbb{R}^d \), we say that \( x \) majorizes \( y \) if for all \( k \leq n \) and for all length \( k \)-sub-sequence \( i_1 < i_2 < \ldots < i_k \),

\[
\sum_{i_j} x_{i_j} \geq \sum_{i_j} y_{i_j}
\]

with equality for \( k = d \).

The following characterization of the majorization relationship tells us that the eigenvalues of the Hermitian part of a matrix \( A \) majorize the Hermitian parts of the eigenvalues of \( A \).

**Theorem 33.** Let \( x \in \mathbb{R}^d \) and \( z \in \mathbb{C}^d \). Then \( x \) majorizes \( \Re(z) \) if and only if there is an \( A \in \mathbb{R}^{d \times d} \) such that \( z \) is the vector of eigenvalues of \( A \) and \( x \) is the vector of eigenvalues of \( H(A) = \frac{1}{2}(A + A^* ) \)

**Corollary 34** (Quadratic form). Let \( A \in \mathbb{R}^{d \times d} \) with eigenvalues \( \rho(A) = \{ \lambda_1, \lambda_2, \ldots, \lambda_d \} \) with \( \lambda_{\min} \triangleq \min \Re(\lambda_i) \) and \( \lambda_{\max} \triangleq \max \Re(\lambda_i) \). For all \( v \in \mathbb{R}^d \),

\[
\lambda_{\min} \|v\|^2 \leq v^T Av \leq \lambda_{\max} \|v\|^2.
\]

Note that the process in [2] is exactly the explicit Euler method for [1]. The following lemma is useful to show these two processes are close to each other.

**Lemma 35** (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

\[
a_k \leq n b \left( \exp \left( \frac{k}{n} \right) - 1 \right).
\]

### 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 \( \sigma \)-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) \( \mathbb{E}[X_{k+1} | \mathcal{F}_k] = X_k \) for all \( k \), \( X \) is said 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 41** (Martingale Stopping theorem). 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 37** (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 \( |E[W_T | \mathcal{F}_{k-1}, W_k, \neg \mathcal{B}] - E[W_T | \mathcal{F}_{k-1}, W_k']| \leq c_k \). Then,

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

The following theorem shows this concentration property is dimension free.

**Theorem 38** (Vector-valued martingale \([33, 27]\)). Let \( g \) be a vector-valued function of \( n \) random variables \( X = (X_1, \ldots, X_n) \). If \( \sup_{x, x'} \|g(x) - g(x')\| \leq c_i \) where \( x \) and \( x' \) only differ by one variable, \( x = x_1, \ldots, x_i, \ldots, x_n, x'_i = x_1, \ldots, x'_{i}, \ldots, x_n \). Then,

\[
\Pr[\|g(X) - E[g(X)]\| \geq t] \leq 20 \exp\left(-\frac{t^2}{2 \sum c_i^2}\right).
\]

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

**Theorem 39** (Maximum tail \([23, 22]\)). 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 \( t \geq 0 \)

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

**Theorem 40** (Wormald’s method \([53]\)). Given differential equation \( (1) \) with \( f \in \mathcal{C}^2 \) initial condition \( x_0 \in \mathbb{R}^d \), and \( X_k \) be the reinforced random walk in \( (3) \) with \( X_0 = x_0 \),

\[
\left\|X_k - x\left(\frac{k}{n}\right)\right\|_\infty = o(1)
\]

for all \( k = O(n) \) with high probability.

\( T \) 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 41** (Martingale Stopping theorem). If \( (W_k)_{0 \leq k \leq n} \) is a martingale with respect to \( (\mathcal{F}_k)_{0 \leq k \leq n} \) and if \( T \) is a stopping time for \( (\mathcal{F}_k)_{0 \leq k \leq n} \) such that \( W_k \) is bounded, \( T \) is bounded, \( \mathbb{E}\{T\} < \infty \), and \( \mathbb{E}\{|W_{k+1} - W_k| | \mathcal{F}_k\} \) is uniformly bounded, then

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

27
C Missing proofs for Section 4

Proof of Lemma 12: Since (11) is a Morse-Smale system and \( V \) is a complete Lyapunov function, starting at \( x_0 \) we know there exists a fixed point \( \beta_i \in \text{Fix} \) such that \( \lim_{t \to \infty} d(\varphi(x_0, t), \beta_i) = 0 \) and \( V(\beta_i) \leq V(x_0) \). Therefore, given \( r > 0 \) a neighborhood of \( \beta_i, B(\beta_i, r) \), there is a constant \( t \) such that \( \varphi(x_0, t) \in B(\beta_i, r) \).

Moreover by Wormald’s method (Theorem 40), the random component \( X_k \) converges to \( B(\beta_i, 2r) \) in \( O(n) \) steps with high probability. Fixing \( N_i \) we can take \( r \) small enough and finish the proof.

C.1 Non attracting fixed points—Lemma 13

The rest of this section is devoted to prove the Lemma 13. Lemma 13 is very similar to the literature of leaving saddle points, and we discuss more details in section 6.

The proof of Lemma 13 has two parts: we first show the process is constant away from the fixed point \( \beta_i \) within time \( T_1 = O(n \log n) \) with high probability in Lemma 14 and we use the property of complete Lyapunov function, and show the value of \( V(X_{T_1}) \) is not much bigger than \( V(\beta_i) \). 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 \( V(\beta_i) \).

To define this two parts formally, We first define several neighborhood of \( \beta_i; N_i \subset B(\beta_i, r/2) \subset B(\beta_i, 3r/4) \subset B(\beta_i, r) \) where \( B(\beta_i, r) \) is the open ball with radius \( \beta_i \) and center at \( \beta_i \). Lemma 13 keeps track of the process when it enter the region \( N_i \) and stop after leaving \( B(\beta_i, r) \). Taking \( r \) small enough such that \( B(\beta_i, r) \) only has a single fixed point \( \beta_i \). Because the complete Lyapunov function \( V \in \mathbb{C}^2 \) and \( \mathcal{L}_f V(x) < 0 \) for all \( x \in B(\beta_i, r) \) which is a compact set, there exists \( \kappa > 0 \) such that

\[
\forall x \in B(\beta_i, r) \setminus B(\beta_i, r/2), \mathcal{L}_f V(x) < -\kappa. \tag{17}
\]

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

\[
\forall x \in N_i, \|V(x) - V(\beta_i)\| \leq \frac{\kappa r}{32Dr}. \tag{18}
\]

Proof of Lemma 13: Suppose the process starting at \( X_0 \in N_i \). Let \( V(k) \triangleq V(X_k) \), by Equation (18),

\[
V(0) \leq V(\beta_i) + \frac{\kappa r}{32Dr}.
\]

We show \( T_1 = O(n \log n/\rho) \) time the process starting at \( N_i \) leaves \( \beta_i; X_{T_1} \in B(\beta_i, 3r/4) \setminus B(\beta_i, r/2) \) with probability \( 1 - \rho \) in Lemma 14.

Because by direct computation the value of complete Lyapunov function \( V \) is a almost a supermartingale, \( \mathbb{E}[V(X_{k+1})] \leq V(X_k) + O\left(\frac{1}{n^2}\right) \), by Azuma’s inequality(Theorem 36), with high probability,

\[
V(T_1) \leq V(0) + \frac{\kappa r}{32Dr} \leq V(\beta_i) + \frac{\kappa r}{16Dr}.
\]

By Equation (17), \( \mathcal{L}_f V(x) \leq -\kappa \) for all \( x \in B(\beta_i, r) \setminus B(\beta_i, r/2) \), we run the process for additional
$T_2 = \frac{rn}{4D'}$ steps then

$$V(T_1 + T_2) = V(T_1) + \sum_{k=T_1}^{T_1+T_2} V(k+1) - V(k)$$

$$= V(T_1) + \sum_{k=T_1}^{T_1+T_2} \left( \frac{d}{dt} V(X_k) + O(\frac{1}{n^2}) \right) \frac{1}{n}$$

$$\leq V(T_1) + \sum_{k=T_1}^{T_1+T_2} \left( -\kappa + O(\frac{1}{n^2}) \right) \frac{1}{n}$$

$$\leq V(T_1) - \frac{\kappa r}{4D'} + O\left(\frac{1}{n^2}\right)$$

$$\leq V(\beta_t) - \frac{\kappa r}{8D'}$$

which shows the process leaves the neighborhood $N_i$ in $O(n \log n/\rho)$ time with probability $1 - \rho$. \qed

C.2 Proofs for Sect. 4.1

Proof of Lemma 15 This is proved by optional stopping time theorem. Given $Z_0 + \beta_k = X_0 \in N_i$, Suppose $T_0$ is the stopping time such that $\|Z_{T_0}\| \geq 8\|Z_{T_0}\|$ or $\|Z_{T_0}\| \leq t_1$. We consider the following random variables $W_{k}^s \triangleq (1 - \lambda_s/2n)^{-k} \|Z_{k}^s\|^2$. Suppose $W_{k}^s$ is a super martingale and $r$ small enough, and by optional stopping time theorem

$$\mathbb{E}[W_{T_0}^s] \leq W_0^s \leq r^2 \leq 1. \quad (19)$$

On the other hand, let $p = \Pr[T_0 \leq n \log n]$

$$\mathbb{E}[W_{T_0}^s] = \mathbb{E}\left[ \left( 1 - \frac{\lambda_s}{2n} \right)^{-T_0} \|Z_{T_0}^s\|^2 \right]$$

$$\geq \mathbb{E}\left[ \left( 1 - \frac{\lambda_s}{2n} \right)^{-T_0} \|Z_{T_0}^s\|^2 \right] = \left( 1 - \frac{\lambda_s}{2n} \right)^{-n \log n} \frac{l_1^2}{2}$$

$$\geq pl_1^2/2 + (1 - p) \left( 1 - \frac{\lambda_s}{2n} \right)^{-n \log n} \frac{l_1^2}{2}$$

$$\geq pl_1^2/2 + (1 - p)n^{\lambda_s/2}l_1^2/2.$$
Now, let's use induction show \( W_T^s \) is a supermartingale before stopping time \( T_0 \):

\[
\left( 1 - \frac{\lambda_s}{2n} \right)^{k+1} E[W_{k+1}^s | \mathcal{F}_k] = E \left[ \|Z_{k+1}^s\|^2 | \mathcal{F}_k \right] = E \left[ \left\| Z_k^s + \frac{1}{n} AZ_k^s + \frac{1}{n} (O(\|Z_k\|^2) + \text{noise}^s) \right\|^2 | \mathcal{F}_k \right] = \|Z_k^s\|^2 + E \left[ \left\langle Z_k^s, \frac{1}{n} AZ_k^s + \frac{1}{n} (O(\|Z_k\|^2) + \text{noise}^s) \right\rangle | \mathcal{F}_k \right] + O \left( \frac{1}{n^2} \right) \quad \text{(by (9))}
\]

\[
= \|Z_k^s\|^2 + \frac{1}{n} (Z_k^s)^\top AZ_k^s + \frac{1}{n} O(\|Z_k\|^3) + O \left( \frac{1}{n^2} \right) = \left( 1 - \frac{\lambda_s}{n} \right) \|Z_k^s\|^2 + \frac{1}{n} O(\|Z_k\|^3) + O \left( \frac{1}{n^2} \right) \quad \text{(by Corollary 34)}
\]

If we take \( N_i \) small enough and \( n \) large enough,

\[
\left( 1 - \frac{\lambda_s}{2n} \right)^{k+1} E[W_{k+1}^s | \mathcal{F}_k] \leq \left( 1 - \frac{\lambda_s}{2n} \right) \|Z_k^s\|^2 = \left( 1 - \frac{\lambda_s}{2n} \right)^{k+1} W_k^s.
\]

This completes the proof.

Proof of Lemma 16. Let \( T_1 \) be the stopping time that \( \|Z_{T_1}\| \geq Cl_1 \) for some constant \( C \). We first show the expectation of \( T_1 \) is much smaller than \( \tau_1 \). Then we show the stable component \( \|Z^s_k\| \) is small for all \( k \leq \tau_1 \). By union bound on these two event, we show with high probability \( \|Z_{T_1}\| \) is large and \( \|Z^s\| \) is small before \( \tau_1 \).

For the first part, because we are in a Euclidean space, the principle angle between \( E^u \) and \( E^s \) is bounded that is

\[
\theta_{us} = \min \{ \arccos(\langle v_s, v_u \rangle) : v_s \in E^s, \|v_s\| = 1, v_u \in E^u, \|v_u\| = 1 \} > 0. \quad (20)
\]

As a result for all \( Z = Z^u + Z^s \), we can lower bound

\[
\|Z\| \geq \|Z^u\| / \sin \theta_{us}, \quad (21)
\]

so it is sufficient to lower bound the magnitude of unstable component, \( \|Z^u\| \). Let \( a^u_{\text{noise}} \triangleq d_1 \operatorname{Tr}((P^u)^\top P^u) > 0 \) and \( W_k \triangleq \|Z^u_k\|^2 - a^u_{\text{noise}} n^2 k \). If \( W_k \) is a submartingale, by optional stopping theorem (Theorem 41) \( E[W_{T_1} | F_0] \geq E[W_0] \geq 0 \) and

\[
E[\|Z^u_{T_1}\|^2] \geq a^u_{\text{noise}} \frac{E[T_1]}{n^2}. \quad (22)
\]

Therefore by (21) and (22),

\[
E[T_1] \leq \frac{n^2}{a^u_{\text{noise}}} E[\|Z^u_{T_1}\|^2] \leq \frac{(n \sin \theta_{us})^2}{a^u_{\text{noise}}} E[\|Z^u_{T_1}\|^2] \leq \frac{(Cn \sin \theta_{us} l_1)^2}{a^u_{\text{noise}}} = O(n \log^{2/3} n).
\]

By Markov inequality there exists \( \tau_1 = O(n \log n) \) such that \( \|Z_k\| \) is greater than \( Cl_1 \) for some \( k \leq \tau_1 \) with probability \( 1 - 1/(\log n)^{1/3} = 1 - o(1) \).
Now, let’s show $W_k$ is a submartingale with respect to $\mathcal{F}_k$ before stopping time $T_1$. Let $Z_{k+1}^u = Z_k^u + D_k$ where $D_k = \frac{1}{n}(AZ_k^u + O(\|Z_k\|^2) + \text{noise}^u)$:

$$
\mathbb{E}[W_{k+1}|\mathcal{F}_k] = \mathbb{E} \left[ \|Z_{k+1}^u\|_2^2 - \frac{a_{\text{noise}}^u}{n^2} (k + 1)|\mathcal{F}_k\right] \\
= \mathbb{E} \left[ \langle Z_k^u + D_k, Z_k^u + D_k \rangle|\mathcal{F}_k\right] - \frac{a_{\text{noise}}^u}{n^2}(k + 1) \\
= W_k + 2\mathbb{E} \left[ \langle Z_k^u, D_k \rangle|\mathcal{F}_k\right] + \mathbb{E} \left[ \langle D_k, D_k \rangle|\mathcal{F}_k\right] - \frac{a_{\text{noise}}^u}{n^2}.
$$

To prove $\mathbb{E}[W_{k+1}|\mathcal{F}_k] \geq W_k$, it is sufficient to show the following two claims:

$$
2\mathbb{E} \left[ \langle Z_k^u, D_k \rangle|\mathcal{F}_k\right] - o(1/n^2) \geq 0 \tag{23} \\
\mathbb{E} \left[ \langle D_k, D_k \rangle|\mathcal{F}_k\right] \geq \frac{a_{\text{noise}}^u}{n^2} \tag{24}
$$

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

$$
2\mathbb{E} \left[ \langle Z_k^u, D_k \rangle|\mathcal{F}_k\right] = 2\langle Z_k^u, \mathbb{E} [D_k|\mathcal{F}_k]\rangle \\
= \frac{2}{n} \langle Z_k^u, A Z_k^u + O(\|Z_k\|^2)\rangle \\
= \frac{2}{n} \left( \langle Z_k^u \rangle^\top A Z_k^u + O(\|Z_k\|^3)\right) \\
\geq \frac{2}{n} (\lambda_u \|Z_k^u\|^2 + O(\|Z_k\|^3)) \quad \text{(by Corollary 34)} \\
> \frac{1}{n} O(\|Z_k\|^3) = o(1/n^2). \quad \text{(} \|Z_k\| = O(\|Z_k\|) = O(l_1) \text{)}
$$

For (23), we use the variance of noise is bounded below by some constant

$$
\mathbb{E} \left[ \langle D_k, D_k \rangle|\mathcal{F}_k\right] = \frac{1}{n^2} \|AZ_k^u + O(\|Z_k\|^2)\|^2 + \frac{1}{n^2} \mathbb{E} \left[ \langle AZ_k^u + O(\|Z_k\|^2), \text{noise}^u \rangle\right] + \frac{1}{n^2} \mathbb{E} \left[ \|\text{noise}^u\|^2\right] \\
\geq \frac{1}{n^2} \mathbb{E} \left[ \|\text{noise}^u\|^2\right] \\
\geq \frac{1}{n^2} d_1 \text{Tr}((P^u)^\top P^u) = \frac{1}{n^2} a_{\text{noise}}^u \quad \text{(by Lemma 42 and definition of } a_{\text{noise}}^u)\text{)}
$$

For the second part, $\|Z_k^u\| = o(l_1)$ for all $k \leq \tau_1$, we can use similar argument in Lemma 15 to show it’s true with high probability.

Finally because with high probability $T_1 < \tau_1$ such that $\|Z_{T_1}\| \geq Cl_1$ and $\|Z_k^u\| = o(l_1)$ for all $k \leq \tau_1$, we have $\|Z_{T_1}^u\| \geq 2l_1$ which completes the proof. \qedhere

Lemma 42 (projected noise). Given a $d$-dimensional random vector $X \in \mathbb{R}^d$, matrices $P, S \in \mathbb{R}^{d \times d}$, and $0 < d_1 < d_2$ where $\mathbb{E}[X] = 0$, $\text{Cov}[X] = S$, $P$ is not the zero matrix, and $S$ is positive definite matrix with $d_1 \mathbb{I}_d < S < d_2 \mathbb{I}_d$, then

$$
0 < d_1 \text{Tr}(P^\top P) < \mathbb{E} \left[ \|PX\|^2 \right] < d_2 \text{Tr}(P^\top P).
$$
Proof.

\[
\mathbb{E} \left[ \|PX\|^2 \right] = \mathbb{E} \left[ \text{Tr} \left( X^T P^T PX \right) \right] \\
= \mathbb{E} \left[ \text{Tr} \left( P^T PXX^T \right) \right] \\
= \text{Tr} \left( P^T P \mathbb{E} \left[ XX^T \right] \right) \\
= \text{Tr} \left( P^T PS \right) > 0 
\]  
(linearity of trace)

Because \( S \) is positive definite and \( P^T P \) is positive semi-definite and not the zero matrix.

Finally, since \( d_1 I \prec S \), \( S - d_1 I \) is positive definite, and

\[
\mathbb{E} \left[ \|PX\|^2 \right] - d_1 \text{Tr}(P^T P) = \text{Tr} \left( P^T PS \right) - d_1 \text{Tr}(P^T P) = \text{Tr} \left( P^T P (S - d_1 I) \right) > 0. 
\]

\[\Box\]

Proof of Lemma 17. Let \( \tau_j = Cn, T_j \) be the stopping time, \( T_j = \arg \min \{X_t \notin B(\beta_i, \sqrt{r})\} \) given \( X_0 = Z_0 + \beta_i \) defined in the statement of Lemma 17 and \( r \) small enough such that (9) holds. Here we abuse the notation and define \( Z_k \) as a new process by Equation (9) 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 \( \tau_j \),

\[
\Pr[T_j > \tau_j] = 1 - o(1/\log n); 
\]  
(25)

2. The expectation at time \( \tau_j \), \( \mathbb{E}[Z_{\tau_j}] \), is nice,

\[
l_{j+1} \geq 8\|\mathbb{E}[Z_{\tau_j}^s]\| \text{ and } \|\mathbb{E}[Z_{\tau_j}^u]\| > l_{j+1}, \text{ and} 
\]  
(26)

3. \( Z_{\tau_j} \) is concentrated

\[
\Pr \left[ l_{j+1} \geq 8\|Z_{\tau_j}^s\| \text{ and } \|Z_{\tau_j}^u\| > l_{j+1} \right] = 1 - o(1/\log n). 
\]  
(27)

Before proving these, let’s do some computation to gain some intuition. To compute the \( \mathbb{E}[Z_{\tau_j}] \) suppose \( T_j > \tau_j \) we can use the linear function \( Ax \) to approximate \( f(x) \) and tower property of expectation:

\[
\mathbb{E}[Z_{k+1}] = \mathbb{E}\left[ \mathbb{E}[Z_{k+1} | \mathcal{F}_k] \right] \\
= \mathbb{E} \left[ \mathbb{E} \left[ Z_k + \frac{1}{n} (AZ_k + O(\|Z_k\|^2) + \text{noise}) | \mathcal{F}_k \right] \right] \quad \text{(by Equation (9))} \\
= \left( 1 + \frac{1}{n} A \right) \mathbb{E} [Z_k] + \frac{1}{n} \mathbb{E} \left[ O(\|Z_k\|^2) \right]. 
\]

Apply this recursive relation repeatedly and we have,

\[
\mathbb{E}[Z_{\tau_j}] = \left( 1 + \frac{1}{n} A \right)^{\tau_j} \mathbb{E} [Z_0] + \frac{1}{n} \sum_{k<\tau_j} \left( 1 + \frac{1}{n} A \right)^{\tau_j-k} \mathbb{E} \left[ O(\|Z_k\|^2) \right]. 
\]  
(28)
Therefore, suppose the norm $\|Z_k\|^2$ for all $0 \leq k < \tau_j$ are small, the value $E[Z_{\tau_j}]$ can be approximated by the first term, $(1 + \frac{1}{n}A)^{\tau_j} E[Z_0]$. Formally, it is sufficient to show for all constant $\epsilon > 0$,

$$\|Z_k\|^2 \leq \epsilon \|Z_0\| \text{ for all } 0 \leq k < \tau_j. \quad (29)$$

**Equation (25):** We define $W_k \triangleq (1 + 2\lambda_{\max} - k) \|Z_k\|^2$ where $\lambda_{\max} = \max_{\lambda \in \rho(A)} \Re(\lambda)$ is the maximum real part of eigenvalues of $A$. By Corollary 33 and similar argument in Lemma 15, $W_k$ is a supermartingale such that $E[W_{k+1} | \mathcal{F}_k] \leq W_k$.

Let’s apply Theorem 39 on $(W_k)$). Because for all $k \leq Cn \|Z_{k+1}\|^2 - \|Z_k\|^2 = O\left(\frac{1}{n}\right)$ uniformly, $D = O(1/n)$, $c_k = |W_{k+1} - W_k| = O\left((1 + 2\lambda_{\max} - k)^2 \frac{1}{n}\right)$, and $\sum c_i^2 = O\left(\frac{1}{n}\right)$. Let $\delta = \frac{\log n}{\sqrt{n}}$. By Theorem 39,

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

Let $\mathcal{E}$ be the good event that $\max_{k \leq \tau_j} W_k < W_0 + \delta$. 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 $0 \leq k \leq \tau_j$

$$\|Z_k\|^2 \leq \left(1 + \frac{2\lambda_{\max}}{n}\right)^k (\|Z_0\|^2 + \delta) \leq 2r \exp(2\lambda_{\max}C) \cdot \|Z_0\|. \quad (30)$$

Given $\epsilon, C, \lambda_{\max} > 0$, we can take $r$ small enough such that $2r \exp(2\lambda_{\max}C) \leq \epsilon \leq 1$. Moreover, this also proves Equation (25), because $\|Z_k\| \leq \sqrt{\epsilon r} \leq \sqrt{r}$.

**Equation (26):** Now we are ready to prove the first part. By Equation (28) and (30), let $\mathcal{E}$ be the event defined in (29), we have

$$E[Z_{\tau_j}] = \left(1 + \frac{1}{n}A\right)^{\tau_j} E[Z_0]$$

$$= \frac{1}{n} \sum_{k < \tau_j} \left(1 + \frac{1}{n}A\right)^{\tau_j - k} E[O(\|Z_k\|^2)] \quad (by \ (28))$$

$$= \frac{1}{n} \sum_{k < \tau_j} \left(1 + \frac{1}{n}A\right)^{\tau_j - k} \left(E\left[O(\|Z_k\|^2) \mid \mathcal{E}\right] + E\left[O(\|Z_k\|^2) \mid \neg\mathcal{E}\right]\right)$$

$$= \frac{1}{n} \sum_{k < \tau_j} \left(1 + \frac{1}{n}A\right)^{\tau_j - k} \left(E\left[O(\|Z_k\|^2) \mid \mathcal{E}\right] + O(Pr[\neg\mathcal{E}])\right) \quad (by \ (30) \ and \ Z_{\tau_j} \in B)$$

$$\leq r O(\|Z_0\|) + \exp\left(-\Omega\left(\sqrt{\log n}\right)\right) = \tilde{c}_j. \quad (by \ \mathcal{E})$$

Therefore, we have if $\|Z_0\| \leq l_j$,

$$E[Z_{\tau_j}] = \left(1 + \frac{1}{n}A\right)^{\tau_j} Z_0 + rO(l_j), \quad (31)$$

and for the unstable component and stable component we have

$$\|E[Z_{\tau_j}^u]\| \geq \left(1 + \frac{\lambda_u}{n}\right)^{\tau_j} \|Z_0^u\| \geq \exp(\lambda_u C) l_j - rO(l_j). \quad (32)$$

$$\|E[Z_{\tau_j}^s]\| \leq \frac{\exp(-\lambda_s C)}{8} l_j + rO(l_j). \quad (33)$$
and the constant of $O(l_j)$ depends on $C$, dimension $d$, and the principle angle $\theta_{us}$ defined in (20). This proves Equation (26) by taking $C$ large enough and $r$ small enough.

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

$$Y_k(Z_0,\ldots,Z_k) = \mathbb{E}[Z_{\tau_j}|Z_0,\ldots,Z_k] \in \mathbb{R}^d.$$  

and prove Equation (27) by using concentration property of vector-valued martingale $Y_k$ (Theorem 38 and 37). With good event $\mathcal{E}$, we want to bound $\{c_k\}_{0 \leq k \leq \tau_j}$ the “variability” of each variable $Z_0,\ldots,Z_{\tau_j}$ on the martingale $Y_k$ condition on this good event defined in (34),

$$c_k = \sup \left\{ \| \mathbb{E}[Z_{\tau_j}|Z_0,\ldots,Z_{k-1},Z_k = z_k,\mathcal{E}] - \mathbb{E}[Z_{\tau_j}|Z_0,\ldots,Z_{k-1},Z_k = z_k,\mathcal{E}] \| \right\}.$$  

Equivalently, $c_k$ is the 2-norm error with initial difference $\|z_k - z_k'\| = O(1/n)$. Formally by (31) and $\mathcal{E}$, we have $c_k = O(1/n)$ for all $k \leq \tau_j$ and $\sum_{k=0}^{\tau_j} c_k^2 = O(1/n)$. By concentration property of vector-valued martingale $Y_k$ (Theorem 38), for any constant $D' > 0$

$$\Pr \left[ \|Z_{\tau_j} - \mathbb{E}[Z_{\tau_j}]\| \geq \frac{l_j}{16D'} \right] \leq O\left( \exp \left( -\Omega(nl_j^2) \right) \right) + \Pr[\mathcal{E}] = \exp \left( -\Omega\left( \sqrt{\log n} \right) \right)$$  

(35)

Therefore, by Equations (32), (33), and (35), with probability $1 - \exp \left( -\Omega\left( \sqrt{\log n} \right) \right) = 1 - o(1/\log n)$ we have,

$$\|Z_{\tau_j}^u\| \geq \|\mathbb{E}[Z_{\tau_j}^u]\| - \frac{l_j}{16D'} \geq \left( \exp \left( \lambda_u C - O(r) - \frac{1}{16D'} \right) \right) l_j \geq 2l_j = l_{j+1}.$$  

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

$$\|Z_{\tau_j}^u\| \leq \|\mathbb{E}[Z_{\tau_j}^u]\| + \frac{l_j}{16D'} \leq \left( \exp \left( -\lambda_u C + O(r) + \frac{1}{2D'} \right) \right) \frac{l_j}{8} \leq \frac{1}{8} l_j \leq \frac{1}{8} l_{j+1}.$$  

which proves Equation (27).

\[\square\]

D Missing proofs for Sect. 5

**Proof of Lemma 19**. We first show there is no fixed point outside $[0,1]^2$, that is the curve $\gamma_1$ and $\gamma_2$ do not have intersection outside.

Let $(x_1,x_2) \in \gamma_1 \cap \gamma_2$. When $m_0 = f'_{ND}(0) = 0$, if $px_1 + qx_2 \leq 0$ or $px_2 + qx_1 \leq 0$ by the definition of $f'_{ND}$ and $\gamma_1$, $(x_1,x_2) = (0,0)$. On the other hand, when $m_0 = f'_{ND}(0) > 0$, $f_{ND}$ is monotone, the above solution curve can be rewritten with respect to

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

(36)

$$\gamma_1 = \{(x_1, x_2) \in [0,1]^2 : x_2 = g(x_1)\}$$  

$$\gamma_2 = \{(x_1, x_2) \in [0,1]^2 : x_1 = g(x_2)\}$$  

(37)

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 $\delta_{sym} = 1/f'_{ND}(1/2)$, we want to show $(0,0), (1,1)$ and $(0.5,0.5)$ are the only intersections between $\gamma_1$ and $\gamma_2$ in $[0,1]^2$ which by symmetry is enough to show the curve $\gamma_1$ is in $R_1 \cup R_3 \cup$
We first show the number of anti-symmetric fixed points is two, than analyze and we show
\[
\{ (0, 0), (1, 1), (0.5, 0.5) \}. \text{ By Definition 6, } f_{ND}(0) = 0, f_{ND}(1/2) = 1/2, \text{ and } f_{ND} \text{ is strictly convex in } [0.0.5], g(0) = 0, g(0.5) = 0.5, \text{ and } g \text{ is strictly concave in } [0, 0.5], \text{ so for all } x_1 \in (0, 0.5),
\]
\[
g(x_1) = g \left( (1 - 2x_1) \cdot 0 + 2x_1 \cdot \frac{1}{2} \right) > \left( (1 - 2x_1) \cdot g(0) + 2x_1 \cdot g \left( \frac{1}{2} \right) \right) = x_1,
\]
and we show \( \gamma_1 \) 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 \( p - q < \delta_{\text{symm}} = 1/f_{ND}'(0.5). \) 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
\[
g(x_1) + x_1 < 1. \tag{39}
\]
Combining equations (38) and (39) 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 2 it is sufficient to study the linear approximation of the dynamics at these points:
\[
\nabla F_{ND}_{(x_1, x_2)} = \begin{bmatrix} -1 + p f_{ND}'(p x_1 + q x_2) & q f_{ND}'(p x_1 + q x_2) \\ q f_{ND}'(q x_1 + p x_2) & -1 + p f_{ND}'(q x_1 + p x_2) \end{bmatrix} \tag{40}
\]
When \((x_1, x_2) = (0, 0), \nabla F_{ND}_{(0,0)} = \begin{bmatrix} -1 + p f_{ND}'(0) & q f_{ND}'(0) \\ q f_{ND}'(0) & -1 + p f_{ND}'(0) \end{bmatrix} \) has trace \( 2(p f_{ND}'(0) - 1) \) and determinant \( ((p - q)f_{ND}'(0) - 1)(f_{ND}'(0) - 1) \). Thus \( \nabla F_{ND}_{(0,0)} \) has two negative real eigenvalues since \( f_{ND}'(0) < 1. \)

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

Proof of Lemma 20. 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 \( \gamma_1 \) overlaps with \( R_2. \) Therefore there exists a non-symmetric intersection between \( \gamma_1 \) and the line \( x_1 + x_2 = 1, (x_1', x_2') \) with \( x_1' \neq x_2' \) which is also in the intersection of \( \gamma_1 \) and \( \gamma_2 \) due to the symmetry.
\[
\begin{align*}
x_1'(a) &= f_{ND}(p x_1'(a) + q x_2'(a)) \\
x_2'(a) &= \frac{1}{q} f_{ND}(p x_2'(a) + q x_1'(a)) \\
1 &= x_1'(a) + x_2'(a) \quad \text{and} \quad x_1'(a) < x_2'(a)
\end{align*}
\tag{41}
\]
Because \( f \) is convex in \([0, 0.5]\), the system only has two anti-symmetric fixed points \( (x_1'(a), x_2'(a)) \) and \( (1 - x_1', 1 - x_2') \).

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,
\[
1 = f_{ND}(s(a)) + f_{ND}(t(a)) \tag{42}
\]
\[
p + q = f_{ND}(s(a)) - f_{ND}(t(a)) \tag{43}
\]
\[
1 = s(a) + t(a) \quad \text{and} \quad s(a) > t(a) \tag{44}
\]

Because \( 1 = x_1^{(a)} + x_2^{(a)} \) and the symmetry of \( f_{ND} \), we have \( \bar{f}_{ND}(s^{(a)}) = \bar{f}_{ND}(t^{(a)}) \) and call it \( m^{(a)}(\delta) \). By Equation (43) and the convexity of \( f_{ND} \), as \( \delta \) 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 (40), the matrix \( \nabla \bar{F}_{ND}|_{(x_1^{(a)},x_2^{(a)})} \) 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 [21]** Let \( (x_1^{(a)},x_2^{(a)}) \) be the anti-symmetric fixed point defined in (41). 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 \( C^2 \) function by implicit function theorem, we can parametrize curves (14) as \( (x_1^{(1)},x_2^{(1)}) \) and \( (x_1^{(2)},x_2^{(2)}) \) of \( \gamma_1 \), and \( \gamma_2 \) respectively. Given \( \delta_e < \delta_{\text{anti}} \), by Lemma 20 \( (x_1^{(a)},x_2^{(a)}) \) is a saddle point,

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

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

\[
\frac{dx_2^{(2)}}{dx_1^{(2)}}|_{(x_1^{(e)},x_2^{(e)})} < \frac{dx_2^{(1)}}{dx_1^{(1)}}|_{(x_1^{(e)},x_2^{(e)})} < m^{(a)}(\delta_e) < 1
\]

(45)

Let \( I \subseteq (\delta_e,\delta_{\text{anti}}) \) be the set of \( \delta \) such that the system (37) has eccentric fixed points. We want to show the system has an eccentric fixed point when \( \delta \) is between \( \delta_e \) and \( \delta_{\text{anti}} \). 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 (37), we know the set \( I \) is closed. To show \( I \) is open, without loss of generality, we show there is a neighborhood of \( \delta_e \) contained in \( I \). Given \( (x_1^{(e)},x_2^{(e)}) \) with \( \delta_e \), fixing \( x_1 = x_1^{(e)} \), let’s consider and the movement of \( x_2^{(1)}(\delta) \) and \( x_2^{(2)}(\delta) \) as \( \delta \) changes around \( \delta_e \) where \( x_2^{(1)}(\delta) \) (and \( x_2^{(2)}(\delta) \)) is the highest intersection between \( x_1 = x_1^{(e)} \) and \( \gamma_1 \) (\( \gamma_2 \) respectively).

\[
\frac{d}{d\delta} \left(x_2^{(1)} - x_2^{(2)}\right) > 0.
\]

(46)

Informally, by Equation (37), as \( \delta \) changes, the curve \( \gamma_1 \) is stretched vertically \( (x_2 \text{ direction}) \) and the movement is proportional to the change rate of \( \delta \). On the other hand, \( \gamma_2 \) is stretched horizontally \( (x_1 \text{ direction}) \), and by Equation (45) the slope is smaller than 1, so the vertically increment rate is smaller than the rate of \( \delta \). 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 (37),

\[
\frac{dx_2^{(1)}}{d\delta} = \frac{1}{2(1-\delta)} (x_2^{(1)} - x_1^{(1)}) = \frac{1}{2(1-\delta)} (x_2^{(e)} - x_1^{(e)}).
\]
Similarly,
\[
\left(1 + \frac{1}{1 - \delta} \left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_1))} - 1 \right) \right) \frac{dx^{(2)}_2}{d\delta} = \frac{1}{2(1 - \delta)}(x^{(e)}_2 - x^{(e)}_1)
\]

Therefore, to prove Equation (46), it is sufficient to show
\[
\left(1 + \frac{1}{1 - \delta} \left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2))} - 1 \right) \right) > 1.
\]

This can be proved by taking derivative at Equation (37) with respect to \(x^{(2)}_2\) and applying Equation (45),
\[
1 = \left(1 + \frac{1}{1 - \delta} \left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2))} - 1 \right) \right) \frac{dx^{(2)}_2}{dx^{(2)}_1} < \left(1 + \frac{1}{1 - \delta} \left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2))} - 1 \right) \right).
\]

Now, let’s prove the eccentric fixed point is stable. Note that by (45) and (46), for all \(\delta > \delta_e\),
\[
0 < \left. \frac{dx^{(2)}_2}{dx^{(2)}_1} \right|_{(x^{(e)}_1, x^{(e)}_2)} < \left. \frac{dx^{(1)}_2}{dx^{(1)}_1} \right|_{(x^{(e)}_1, x^{(e)}_2)} < 1.
\]

Rewrite the above inequality in terms of \(f_{\text{ND}}\) we have,
\[
1 > \frac{1}{1 - \delta} \left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_1))} - \delta \right) > \left[ \frac{1}{1 - \delta} \left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2))} - \delta \right) \right]^{-1} > 0.
\]

By Equation (40), the matrix \(\nabla F_{\text{ND}}(x^{(e)}_1, x^{(e)}_2)\) is
\[
\begin{bmatrix}
-1 + p f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_1)) & q f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_1)) \\
q f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2)) & -1 + p f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2))
\end{bmatrix}.
\]

The trace is negative, because \(f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_1)) < 1\) and \(f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2)) < 1/\delta\). The determinant is positive, because \(\left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_1))} - \delta \right) \cdot \left( \frac{1}{f'_{\text{ND}}(f^{-1}_{\text{ND}}(x^{(e)}_2))} - \delta \right) > (1 - \delta)^2\). Therefore, the \((x^{(e)}_1, x^{(e)}_2)\) is a stable fixed point.

\[\Box\]

E Node dynamics on bi-blockmodel

To prove the first part, our proof has two steps: given an arbitrary neighborhood of consensus states \(Q\), the Markov chain \(X^{\text{ND}}\) reaches \(Q\) in \(O(n \log n)\) with high probability, and it hits the consensus states in \(O(n \log n)\) with constant probability if \(X^{\text{ND}}_0 \in \overline{Q}\) when \(Q\) small enough. The first one is proved in Lemma 43 and the second part is proved in Lemma 45.

**Lemma 43** (Reaching neighborhood \(Q\)). In case 1 of Theorem 2, given arbitrary neighborhoods \(M_0, M_1 \subset [0, 1]^2\) such that \((0, 0) \in M_0\) and \((1, 1) \in M_1\), the hitting time of \(X^{\text{ND}}\) to set \(Q \triangleq M_0 \cup M_1\) is
\[
Pr[\forall \sigma_0 \in \{0, 1\}^n, T_Q(\sigma_0) = O(n \log n)] = 1 - o(1),
\]
where \(T_Q(\sigma_0)\) denotes the stopping time of \(S^{\text{ND}}\) such that \(\text{pos}(S^{\text{ND}}_{T_Q}) \in Q\) from the initial state \(\sigma_0\).
Proof of Theorem 9. For the first part, by Lemma 43 consensus states. Let $E = \{0, 0\}, (1, 1), Q$ in $O(n \log n)$ with high probability if the noise is well-behaved:

$$\exists d_1, d_2 > 0, \forall x \in \Omega_X \setminus Q, d_1 \|d - d_2\| \preceq \text{Cov}[U(x)] \prec d_2 \|d.$$  \hspace{1cm} (49)

which is proved in Lemma 44.

□

Lemma 44 (Well-behaved noise). Given $X^{ND}$ defined in (7), there exist $d_1, d_2 > 0$, for all $x \in \Omega_X \setminus Q$:

$$d_1 \|d - d_2\| \preceq \text{Cov}[U(x)] \prec d_2 \|d$$

where $U(x) \triangleq n(X' - E[X'])$ condition on $X = x$.

Lemma 45 (Reaching consensus). In the first case of Theorem 9, there exist $T = O(n \log n)$, neighborhoods $M_0, M_1$, and $Q$ in $[0, 1]^2$ where $(0, 0) \in M_0$, $(1, 1) \in M_1$, and $Q \triangleq M_0 \cup M_1$, such that for all post($\sigma_0) \in Q$

$$\Pr[T(\sigma_0) \leq T] \geq 1/6$$

where $T(\sigma_0)$ denotes the hitting time of $S^{ND}$ to consensus states $0^n$ or $1^n$ with initial state $\sigma_0$.

Lemma 46 (Potential wall). Given a time homogeneous Markov chain $S_k$ with state space $\Omega$, if there exist constants $d \in \mathbb{N}$, $D \in \mathbb{R}_+$ compact sets $Q \subset Q' \subset \mathbb{R}^d$, functions $\phi : \Omega \rightarrow \mathbb{R}^d$, and $f \in C^2(\mathbb{R}^d, \mathbb{R}^d)$ with an attracting fixed point $\beta_a \in Q$, if $\{X_k\}_{k \in \mathbb{N}} = \{\phi(S_k)\}_{k \in \mathbb{N}}$ is a function of Markov chain in $\mathbb{R}^d$, and

1. $X_{k+1} = X_k + \frac{1}{n} (f(X_k) + U(S_k))$, and set $U_{k+1} \triangleq U(S_k)$.
2. For all $X_k \in \Omega$, $E[U_{k+1} \mid S_k] = 0$.
3. For all $X_k \in \Omega$, $\|U_{k+1}\| \preceq D$

For all $s_0$ with $\phi(s_0) \in Q$ and $T \geq 1$

$$\Pr[\forall k < T, X_k \in Q'] \geq 1 - T \exp(-\Omega(n)).$$

With above four lemmas, we are ready to prove the Theorem 9.

Proof of Theorem 9. For the first part, by Lemma 44 $X^{ND}$ reaches a fixed neighborhood of consensus states $(0, 0), (1, 1), Q$ in $O(n \log n)$ with high probability. By Lemma 45, if $Q$ is small enough, the process hits consensus states in $T_c = O(n \log n)$ with probability at least 1/7. Therefore

$$\Pr[\forall \sigma_0 \in \{0, 1\}^n, T(\sigma_0) \leq T_c] \geq 1/7.$$  \hspace{1cm} (50)

Because the $X^{ND}$ is a Markov chain which is bounded in $B \subseteq [0, 1]^2$ and the upper bound in (50) is independent of initial states, we can show upper bound the expected hitting time to consensus states. Let $E_c$ be the event of hitting $Q$ in time $T_c = O(n \log n)$, and by Equation (50) we have $\Pr[E_c] \geq 1/7$. Because the process $X^{ND}$ is a Markov chain and bounded, we can partition the time into intervals with duration $T_c$, the process $X^{ND}$ reaches consensus with probability at most $\Pr[E_c]$ and this bounds are independent of different time intervals so the expected number of intervals for $E_c$ to happens is bounded by the expectation of a geometric random variable with success probability $\Pr[E_c]$ which is constant. Therefore, $\text{ME}(K(n, p, q), f_{ND}) = O(n \log n)$.
For the second part, by Theorem 11 there is an extra attracting fixed point $\beta_a$ of $\bar{F}_{ND}$. By Lemma 46, there exists neighborhoods of $\beta_a$, $Q$ and $Q'$ such that for any $\sigma_0$ with $\phi(\sigma_0) \in Q$ and $T \geq 1 \Pr [X_T \in Q'] \geq 1 - T \exp(-\Omega(n))$. Therefore, with initial state $\sigma_0$

$$\Pr [T(G, f_{ND}, \sigma_0) \geq k] \geq \Pr [X_k \in Q'] \geq 1 - k \exp(-\Omega(n))$$

Because the hitting time is a non-negative random variable

$$\mathbb{E}[T(G, f_{ND}, s_0)] = \sum_k \Pr [T(G, f_{ND}, s_0) \geq k] \geq \sum_k 1 - k \exp(-\Omega(n)) = \exp(\Omega(n)).$$


\[\square\]

**Proof of Lemma 46** Because $\beta_a$ is an attracting fixed point, all the eigenvalue of $A \triangleq \nabla f|_{\beta_a}$ has negative real part which is called stable matrix (or sometimes Hurwitz matrix), and by Lyapunov theorem there exists a positive definite matrix $P$ such that $PA + A^TP = -\mathbb{I}_d$. We define $V(x) \triangleq x^TPx$. Therefore, with Taylor expansion on $V$ and the property of $A$, we have

$$\mathbb{E}[V(X_{k+1}) | S_k] = \mathbb{E} \left[ V(X_k) + \frac{1}{n} \nabla V(X_k) \cdot (f(X_k) + U_{k+1}) + O \left( \frac{1}{n^2} \right) | S_k \right]$$

$$\leq V(X_k) + \frac{2}{n} x_k^TP \cdot (f(X_k)) + O \left( \frac{1}{n^2} \right)$$

$$\leq V(X_k) + \frac{2}{n} x_k^TPA x_k + \frac{L}{n} \|x_k\|^3 \quad (L \text{ bounded because } f \in C^2 \text{ in } B)$$

$$\leq V(X_k) + \frac{1}{n} x_k^T(PA + A^TP)x_k + \frac{L}{n} \|x_k\|^3$$

$$\leq V(X_k) - \frac{1}{n} \|x_k\|^2 + \frac{L}{n} \|x_k\|^3 \quad (PA + A^TP = -\mathbb{I}_d)$$

Therefore the value $V(X_k)$ is a super martingale and there exists $r > 0$ such that $\mathbb{E}[V(X_{k+1}) | S_k] - V(X_k) \leq -r$ for all $X_k \in Q'$ when $Q'$ are small enough. Furthermore, because $P$ is positive definite we can take $Q \subset Q'$ small enough such that the potential value has constant separation: $\max_{x \in Q} V(x) < \min_{x \notin Q'} V(x)$.

Suppose there exists $0 \leq l \leq T$ such that $X_l \notin Q'$. Because $X_0 = \phi(s_0) \in Q$, there exists an interval of time from $k$ to $l$ such that $X_k \in Q$, $X_l \notin Q'$ and $X_{\ell} \in Q' \setminus Q$ for all $k < \ell < l$, we define this event as $E_l$. Because each step the process $X_k$ can only increase by $1/n$ and the potential value in $Q$ and outside $Q'$ has constant separation, the time interval is $l - k \geq cn$ for some constant $c > 0$ however, such event $E_l$ happen with probability

$$\Pr [X_l \notin Q'] \leq \Pr [E_l] \leq \exp(-\Omega(n))$$

by Azuma’s inequality. The proof is finished by taking union bound on $l$.

\[\square\]

**Proof of Lemma 44** By the definition of $X^ND$, given $X = x = (x_1, x_2) \in \Omega_X \setminus Q$, define the difference to be $Y \triangleq n(X' - X)$ where $Y = (Y_1, Y_2) \in \{(0, 0), (1, 0), (-1, 0), (0, 1), (0, -1)\}$ only have these five possible outcomes, and we can compute these directly:
\[ p_1^+(x) \triangleq \Pr[Y = (1, 0) \mid X = x] = \frac{1 - x_1}{2} f_{ND}(px_1 + qx_2), \]
\[ p_1^-(x) \triangleq \Pr[Y = (-1, 0) \mid X = x] = \frac{x_1}{2} (1 - f_{ND}(px_1 + qx_2)), \]
\[ p_2^+(x) \triangleq \Pr[Y = (0, 1) \mid X = x] = \frac{1 - x_2}{2} (f_{ND}(qx_1 + px_2)), \]
\[ p_2^-(x) \triangleq \Pr[Y = (0, -1) \mid X = x] = \frac{x_2}{2} (1 - f_{ND}(qx_1 + px_2)). \]

We omit \( x \) when it is clear. Then by the definition of \( U(x) \) and \( Y \),
\[ \text{Cov}[U(x)] = \text{Cov}[n (X' - \mathbb{E}[X']) \mid X = x] \]
\[ = \text{Cov}[n (X' - x) \mid X = x] \]
\[ = \text{Cov}[Y \mid X = x] \]
\[ = \begin{bmatrix}
\text{Var}[Y_1] & \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1] \mathbb{E}[Y_2] \\
\mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1] \mathbb{E}[Y_2] & \text{Var}[Y_2]
\end{bmatrix} \]
\[ = \begin{bmatrix}
(p_1^+ + p_1^-) - (p_1^+ - p_1^-)^2 & -(p_1^+ - p_1^-)(p_2^+ - p_2^-) \\
-(p_1^+ - p_1^-)(p_2^+ - p_2^-) & p_2^+ + p_2^- - (p_2^+ - p_2^-)^2
\end{bmatrix}. \]

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 \( \text{Cov}[U(x)] \) can be simplified as,
\[ \text{Cov}[U(x)] = \begin{bmatrix}
S_1 - D_2^2 & -D_1 D_2 \\
-D_1 D_2 & S_2 - D_2^2
\end{bmatrix}. \tag{51} \]

Because \( \text{Cov}[U(x)] \) is symmetric, the eigenvalues are real. By Gershgorin circle theorem and (51), the eigenvalues are upper bounded by
\[ \max \{ S_1 - D_2^2 + |D_1 D_2|, S_2 - D_2^2 + |D_1 D_2| \} \leq 1, \]
and lower bounded by
\[ \min \{ S_1 - D_2^2 - |D_1 D_2|, S_2 - D_2^2 - |D_1 D_2| \}, \tag{52} \]
so to find \( d_1 \) it is sufficient to lower bound Equation (52).

Because \( Q = M_0 \cup M_1 \), there exists constant \( \alpha > 0 \) such that 1-norm balls \( \{ x \in [0, 1]^2 : \| x \|_1 \leq \alpha \} \) and \( \{ x \in [0, 1]^2 : \| x - (1, 1) \|_1 \leq \alpha \} \) are inside \( Q \). Thus, if \( (x_1, x_2) \in \Omega_X \setminus Q \), \( px_1 + qx_2 \), \( qx_1 + bx_2 \) are in \( [\alpha, p(1 - \alpha)] \), so
\[ 0 < f_{ND}(\alpha x) \leq f_{ND}(px_1 + qx_2) \text{ and } f_{ND}(qx_1 + px_2) \leq f_{ND}(p(1 - \alpha)) < 1 \tag{53} \]
As a result, \( p_1^+, p_1^- \), \( p_2^+ \) and \( p_2^- \) are smaller or equal to \( \frac{1}{2} f_{ND}(p(1 - \alpha)) \), and \( |D_1|, |D_2| \leq \frac{1}{2} f_{ND}(p(1 - \alpha)) \). Moreover,
\[ \begin{aligned}
(51) & \geq \min \{ S_1 - f_{ND}(p(1 - \alpha))|D_1|, S_2 - f_{ND}(p(1 - \alpha))|D_2| \} \\
& \geq (1 - f_{ND}(p(1 - \alpha)) \min \{ S_1, S_2 \}. 
\end{aligned} \]

Because \( S_1 = p_1^+ + p_1^- \) is a convex combination of \( f_{ND}(px_1 + qx_2)/2 \) and \( (1 - f_{ND}(px_1 + qx_2))/2 \), and \( S_2 = p_2^+ + p_2^- \) is a convex combination of \( f_{ND}(qx_1 + px_2)/2 \) and \( (1 - f_{ND}(qx_1 + px_2))/2 \), by (53), \( \min \{ S_1, S_2 \} \geq \frac{1}{2} \min \{ f_{ND}(\alpha x), 1 - f_{ND}(p(1 - \alpha)) \} \),
\[ (51) \geq (1 - f_{ND}(p(1 - \alpha))) \cdot \frac{1}{2} \min \{ f_{ND}(\alpha x), 1 - f_{ND}(p(1 - \alpha)) \} > 0 \]
Therefore, we can take $0 < d_1 < \frac{1}{2}(1 - f_{\text{ND}}(p(1-\alpha))) \cdot \min\{f_{\text{ND}}(q\alpha), 1 - f_{\text{ND}}(p(1-\alpha))\}$ and $d_2 = 2$ which completes the proof. \hfill \Box

**Proof of Lemma 47.** Let $\psi(k) = \sum_{1 \leq \ell \leq k} d_\ell$ and $\psi(0) = 0$. By direct computation, for all $0 < k < m$

$$
\mathcal{L} \psi(k) = p^+(k) (\psi(k+1) - \psi(k)) - p^-(k) (\psi(k) - \psi(k-1)) = p^+(k) d_{k+1} - p^-(k) d_k \quad \text{(definition of } \psi) \\
\leq -1 \quad \text{(definition of } d_k)
$$

Finally, $\mathcal{L} \psi(m) = -p^-(k) (\psi(k) - \psi(k-1)) - p^-(k) d_k \leq -1$. Therefore $\psi(m)$ is a upper bound for the maximum expected hitting time by Corollary 31. \hfill \Box

### E.1 From neighborhood of attracting fixed points to fixed points

In this section, we want to prove Lemma 45: once the process $X_{\text{ND}}$ hits the set $Q$ defined in Lemma 43 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 47, we give a simple upper bound for hitting time of birth-and-death chain. In Lemma 48, a uniform bound for (54) is given for our process.

**Lemma 47 (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

$$
\Pr[W' = W + 1 \mid W = \ell] = p^+(\ell) \\
\Pr[W' = W \mid W = \ell] = 1 - p^+(\ell) - p^-(\ell) \\
\Pr[W' = W - 1 \mid W = \ell] = p^-(\ell)
$$

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

$$
d_m \geq \frac{1}{p^-(m)} \text{ and } d_{\ell-1} \geq \frac{1}{p^-(\ell-1)} + \left(\frac{p^+(\ell+1)}{p^-(\ell-1)}\right) d_\ell \quad (54)
$$

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

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

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

**Lemma 48.** Let $h(\sigma) \triangleq \|\text{Pos}(\sigma)\|_1$. There exist positive constants $\alpha$, $\gamma$, and $\epsilon$, such that for all $S_{k}^{\text{ND}}$ with $h(S_{k}^{\text{ND}}) \leq \epsilon n$,

$$
\Pr \left[ h(S_{k+1}^{\text{ND}}) = h(S_k^{\text{ND}}) - 1 \mid S_{k}^{\text{ND}} = \sigma_0 \right] \geq \gamma h(\sigma_0)/n, \quad (55)
$$

and

$$
\frac{\Pr \left[ h(S_{k+1}^{\text{ND}}) = h(S_k^{\text{ND}}) + 1 \right]}{\Pr \left[ h(S_{k+1}^{\text{ND}}) = h(S_k^{\text{ND}}) - 1 \right]} \leq 1 - \alpha. \quad (56)
$$
Proof of Lemma 45. Without loss of generality, suppose \( \text{pos}(\sigma_0) \in M_0 \). Consider a function \( h : \Omega \rightarrow \mathbb{N}_0 \) where \( h(\sigma) \triangleq \|\text{Pos}(\sigma)\|_1 \). Let \( V_k = h(S_k) \) is a stochastic process on \( \mathbb{N}_0 \) and the process \( S_k \) reaches \( 0^n \) if and only \( h(S_k^{ND}) = 0 \). With \( M_0 \) and \( h \) we define \( m_0 = \max\{ h(\sigma) : \text{pos}(\sigma) \in M_0 \} = \Theta(n) \).

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

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

(57)

Then show

\[
\Pr[V_T = 0] \geq \Pr[V_T \geq 2m_0]
\]  

(58)

With the above two equations, we have

\[
\Pr[T \leq 3\tau'] \geq \Pr[T \leq 3\tau' \land V_T = 0]
\]

\[
\geq 1 - \Pr[V_T \neq 0] - \Pr[T \geq 3\tau']
\]

(union bound)

\[
\geq 1/2 - 1/3 = 1/6
\]

(by Markov inequality and (58))

Now let’s prove the Equation (57) and (58). For Equation (57) 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 \( 0 < \ell < 2m_0 \)

\[
\Pr[W' = W + 1 \mid W = \ell] = \max_{\sigma : h(\sigma) = \ell} \Pr[V' = V + 1 \mid V = h(\sigma)]
\]

\[
\Pr[W' = W - 1 \mid W = \ell] = \min_{\sigma : h(\sigma) = \ell} \Pr[V' = V - 1 \mid V = h(\sigma)]
\]

(59)

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 48 and 47, the expected hitting time of \( W_k \) to state 0 is upper bounded by \( \sum_{\ell \leq 2m_0} d_{\ell} \) where \( d_{\ell} \) is defined in Lemma 47. By Lemma 48, we can set \( d_{2m_0} = \frac{n}{\gamma_2m_0} = O(1) \), for all \( 1 \leq \ell < 2m_0 \), \( d_{\ell} = \frac{1}{\gamma_2} + (1 - \alpha)d_{\ell+1} \). 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 (58) is true, because \( V_k \) is a supermartingale, \( \mathbb{E}[V_{k+1} \mid S_k^{ND}] \leq V_k \) by Lemma 48.

Proof of Lemma 48. This Lemma shows if the fraction of opinion 1 in \( V_1 \) and \( V_2 \) is smaller than \( \alpha \), the number of 1 opinion decrease fast. Given configuration \( S_k \), 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 \( f_{ND}(1) = 1, f_{ND}(0.5) = 0.5 \), there exists \( m_1 \) such that \( f_{ND}'(1) < m_1 < 1 \) and for all \( 0 < 1 - x < \epsilon \)

\[
f_{ND}(x) \leq 1 + m_1(x - 1).
\]

(60)

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

\[
f_{ND}(x) \geq m_0 x.
\]

(61)
Let’s first prove (55). The event that \( h(S_{k+1}) = h(S_k) - 1 \) is equal at time \( k+1 \) a node with opinion 1 is chosen and updates its opinion to 0,

\[
\Pr[h(S_{k+1}) = h(S_k) - 1 | S_k] = \frac{a_k}{n} \Pr[v_1 \in V_1 \text{ updates to 0}] + \frac{b_k}{n} \Pr[v_2 \in V_2 \text{ updates to 0}]
\]

\[
= \frac{a_k}{n} \left( 1 - f_{\text{ND}} \left( \frac{2a_k}{n} + \frac{2b_k}{n} \right) \right) + \frac{b_k}{n} \left( 1 - f_{\text{ND}} \left( \frac{2a_k}{n} + \frac{2b_k}{n} \right) \right)
\]

\[
\geq \frac{a_k}{n} m_1 \left( 1 - \frac{2a_k}{n} - \frac{2b_k}{n} \right) + \frac{b_k}{n} m_1 \left( 1 - \frac{2a_k}{n} - \frac{2b_k}{n} \right) \quad \text{(by (60))}
\]

\[
\geq \frac{a_k + b_k}{n} m_0 (1 - 2\epsilon)
\]

\[
\geq \frac{m_1}{2n} (a_k + b_k) = \frac{m_1}{2} h(S_k)/n \quad \text{(if } \epsilon \text{ smaller than 1/4)}
\]

Therefore this proves (55) by taking \( 0 < \gamma < \frac{m_1}{2} \).

For the (56), with (55), it is sufficient to show there exists \( \delta \) such that \( \Pr[h(S_{k+1}) = h(S_k) - 1] - \Pr[h(S_{k+1}) = h(S_k) + 1] \) is greater than \( \delta h(S_k)/n \). This can be done by computation

\[
\Pr[h(S_{k+1}) = h(S_k) - 1] - \Pr[h(S_{k+1}) = h(S_k) + 1] = \mathbb{E}[h(S_{k+1})] - h(S_k)
\]

\[
= \mathbb{E}[a_{k+1} + b_{k+1}] - a_k - b_k
\]

\[
= f_{\text{ND}}(pa_k/n + qb_k/n) + f_{\text{ND}}(qa_k/n + pb_k/n)
\]

\[
\geq m_0(pa_k/n + qb_k/n) + m_0(qa_k/n + pb_k/n) \quad \text{(by (61))}
\]

\[
\geq m_0(a_k/n + b_k/n) = m_0 h(S_k)/n
\]

, and these complete the proof for (56).