Dynamic Factorization and Partition of Complex Networks

Lin F. Yang, Vladimir Braverman, Tuo Zhao, Mengdi Wang *

Abstract

Finding the reduced-dimensional structure is critical to understanding complex networks. Existing approaches such as spectral clustering are applicable only when the full network is explicitly observed. In this paper, we focus on the online factorization and partition of implicit large-scale networks based on observations from an associated random walk. We propose an efficient and scalable nonconvex stochastic gradient algorithm. It is able to process dependent data dynamically generated by the underlying network and learn a low-dimensional representation for each vertex. By applying a diffusion approximation analysis, we show that the nonconvex stochastic algorithm achieves nearly optimal sample complexity. Once given the learned low-dimensional representations, we further apply clustering techniques to recover the network partition. We show that, when the associated Markov process is lumpable, one can recover the partition exactly with high probability. The proposed approach is experimented on Manhattan taxi data.

1 Introduction

Network data arise in many applications and research areas, including but not limited to social science, economics, transportation, finance, power grid, artificial intelligence, etc. Examples include protein-protein interaction networks Junker and Schreiber (2011), phone communication networks Newman (2001), collaboration networks Blondel et al. (2008), correlation networks between stock prices and the gravitational interaction network of dark matter particles in cosmology (Peebles, 1980; Mantegna, 1999; Nerur et al., 2005). Due to the highly complex nature of these networks, many efforts have been devoted to investigate reduced-order representations from high-dimensional data (e.g. Chung (1997); Pizzuti (2008); Page et al. (1999); Chen and Yuan (2006)).

In this paper, we focus on learning from the dynamic “state-transition” data, which are snapshots of a random walk associated with the implicit network. For example, records of taxi trips can be used to reveal the traffic network of a metropolitan Liu et al. (2012); Benson et al. (2016).

*Lin Yang, Vladimir Braverman are affiliated with Department of Computer Science at Johns Hopkins University; Lin Yang is also affiliated with Department of Physics and Astronomy at Johns Hopkins University; Tuo Zhao is affiliated with School of Industrial and Systems Engineering at Georgia Institute of Technology; Mengdi Wang is affiliated with Department of Operations Research and Financial Engineering, Princeton University. Email: lyang@jhu.edu, mengdiw@princeton.edu; Mengdi Wang is the corresponding author.
Each trip can be viewed as a fragmented sample path realized from a city-wide Markov chain that characterizes the traffic dynamics. For another example, reinforcement learning applications such as autonomous driving and game AI are modeled as Markov decision processes Sutton and Barto (1998). Given trajectories of game snapshots or a game simulator, it is of vital interest to identify the low-dimensional representation of the “state” of game. Existing approaches for network partition do not address the Markov nature of state-transition data. They often require explicit knowledge and pre-computation of large matrices, which cannot scale to large-scale problems and is not even possible for online learning applications. Efficient algorithms are in demand.

Motivated by the need to analyze state-transition data, we propose an efficient and scalable approach for online factorization and partition of implicit complex networks. We start by employing a basic nonconvex stochastic gradient descent (SGD) algorithm and tailor it towards processing Markov transition data. Then we show that the SGD learns low-dimensional representations of the network in an online fashion, and by further applying clustering techniques, we can recover the underlying partition structure with high probability. Though the stochastic optimization problem is highly nonconvex, we show that the SGD converges to the global optima with high probability, even if the initial solution is chosen uniformly at random. Our proof uses the diffusion approximation of random process Harold et al. (1997). We characterize the discrete trajectory of the stochastic updates by their continuous-time limits, which is the solution to an ordinary differential equation (ODE). We further show that the process at sufficiently large time is well approximated by an Ornstein-Uhlenbeck process, whose fluctuation can be precisely characterized. Despite of the spherical geometry and many unstable equilibria of the optimization problem, we establish global convergence with a near-optimal sample complexity guarantee.

Our work is partly motivated by Weinan et al. (2008), which establishes the connection between networks and a class of lumpable Markov chains. It proposes an optimization framework to identify the partition structure when the transition matrix is known a priori. Our method is also related to the class of online eigenvalue decomposition methods for representation learning Allen-Zhu and Li (2016); Li et al. (2016); Xie et al. (2015); Arora et al. (2016); Jain et al. (2016). However, none of the existing methods and analysis are applicable to Markov transition data and dynamic network partition. We provide more discussions about related methods in Section 7.

The paper is organized as follows: In Section 2, we briefly review some preliminaries of networks and associated Markov chains; In Sections 3-4, we propose the dynamic network partition approach and establish its theoretical properties; In Section 6, we provide thorough numerical experiments; In Section 7, we discuss the related work; In 5, we provide the proof sketch of our main results; A full proof is deferred to the appendix.

**Notation:** We denote \([n] = \{1, 2, \ldots, n\}\). Given two matrices \(U \in \mathbb{R}^{m \times r_1}, V \in \mathbb{R}^{m \times r_2}\) with orthonormal columns, where \(1 \leq r_1 \leq r_2 \leq m\), we denote the principle angle between two matrices by

\[
\Theta(U, V) = \text{diag}\left[\cos^{-1}\left(\sigma_1(U^T V)\right), \cos^{-1}\left(\sigma_2(U^T V)\right), \ldots, \cos^{-1}\left(\sigma_{r_1}(U^T V)\right)\right],
\]

where \(\sigma_i(A)\) is the \(i\)-th largest singular value of matrix \(A\). We also use \(\cos(\cdot)\) and \(\sin(\cdot)\) to act on matrices and denote entry-wise functions. For a matrix \(V\), we denote by \(V_{\cdot j}\) its \(j\)-th column vector and by \(V_{i \cdot}\) its \(i\)-th row vector. We denote by \(V_{1:r, r}\) the sub-matrix of the first \(r\) columns. We denote by \(\|\cdot\|_F\) the Frobenius norm of a matrix, and denote by \(\|\cdot\|_2\) the Euclidian norm of a vector or the
spectral norm of a matrix. We denote by $e_i \in \mathbb{R}^s$ the $i$-th standard unit vector for any $s \geq i$: $(e_i)_i = 1$ and $(e_i)_j = 0$ for $j \neq i$. We also denote by $0_{m \times n} \in \mathbb{R}^{m \times n}$ the matrix with all 0 entries.

2 Preliminaries

Let us review the basics of networks and the associated Markov chains.

2.1 Networks and Associated Markov Chains

Let $G = (S, E)$ be a network with $m$ vertices (a weighted directed graph), where $S = \{s_1, s_2, \ldots, s_m\}$ denotes the vertex set, $E = \{w_{i,j} \geq 0 : i, j \in [m]\}$ denotes the edge set, and $w_{i,j}$ denotes the weight of the edge $(s_i, s_j)$.

Consider the random walk that is naturally associated with the network $G$: We denote by $P = (p_{i,j}) \in \mathbb{R}^{m \times m}$ its probability transition matrix, where each state of the Markov chain $P$ corresponds to a vertex in $G$. The random walk $P$ generated by the network $G$ satisfies

$$P[s^{(t)} = s_j | s^{(t-1)} = s_i] = p_{i,j}, \text{ where } \forall i, j : p_{i,j} = \frac{w_{i,j}}{w_i} \text{ and } w_i = \sum_{j \in [m]} w_{i,j}. $$

The stationary distribution of the Markov chain is $\mu_i = \frac{w_i}{\sum_{j \in [m]} w_j}$. When the network $G$ is undirected, the corresponding Markov chain is reversible and satisfies the following detailed balance condition

$$\forall i \neq j, \mu_i p_{i,j} = \mu_j p_{j,i} \text{ and } \sum_{i \in [m]} u_i p_{i,j} = \mu_j, \quad (1) $$

i.e., $DP = PD$, where $D = \text{diag}(\mu_1, \mu_2, \ldots, \mu_m)$. In this paper, we focus on fully connected and undirected networks, where $\mu_i > 0$ for all $i \in [m]$. We note that our analysis are generalizable to directed networks.

2.2 Our Problem of Interest

Given a sample trajectory $s^{(0)}, s^{(1)}, \ldots, s^{(t)}, \ldots$ of state transitions of the unknown Markov chain $P$, our objective is to develop an online learning method to extract reduced-order information about the Markov chain and recover the network partition.

We are interested in complex networks that can be approximated using reduced-order representations. In particular, we focus on Markov chains that are nearly low-rank, which is defined as follows:

**Definition 1** (Nearly Low-Rank Markov Chains). A Markov chain with transition matrix $P$ is nearly low-rank if there exist matrices $F_1, F_2 \in \mathbb{R}^{m \times m}$, where $\text{rank}(F_1) = r$ and $\|F_2\|_2 < \sigma_r(F_1)$ such that

$$DP = F_1 + F_2 \quad \text{and} \quad F_1^T F_2 = 0_{m \times m}, \quad (2) $$

and $F_1 = \mathcal{U}^T \Sigma \mathcal{V}$, where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ is a diagonal matrix with $1 \geq \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, and $\mathcal{U}, \mathcal{V} \in \mathbb{R}^{n \times r}$ are orthonormal matrices.
Each row of $M = D^{-1}V \in \mathbb{R}^{m \times r}$ can be viewed as an $r$-dimensional representation of a vertex of $G$. The matrix $M$ gives an approximate decomposition of the Markov chain, which has a similar spirit as spectral clustering Chung (1997). We also consider an important special case of “lumpable” Markov chains, which are defined below.

**Definition 2** (Lumpable Markov chains Weinan et al. (2008)). A Markov chain on states $S$ with transition matrix $P$ is lumpable with respect to partition $S = S_1 \cup S_2 \ldots \cup S_r$ if the transitions between these sets satisfy the strong Markov property, i.e., for any $s_k, s_h \in S_i$,

$$
\sum_{s_j \in S_j} p_{k, j} = \sum_{s_j \in S_j} p_{h, j}.
$$

We can view $S_1, \ldots, S_r$ as “meta states” of the Markov chain. The meta states suffice to characterize the dynamics of a complex Markov chain. For lumpable Markov chains, one can obtain the optimal partition of the corresponding network by clustering based on the $r$-dimensional representations Weinan et al. (2008).

## 3 Method

Recall that we are interested in learning from Markov transition data. In particular, consider the scenario where we only observe state-to-state transitions of a Markov process over $S$:

$$s^{(1)}, s^{(2)}, s^{(3)}, \ldots, s^{(n-1)}, s^{(n)}, \ldots,$$

without knowing the transition matrix $P$ in advance. For notational convenience, we simplify the notation of the states to $S = \{1, 2, \ldots, m\}$.

### 3.1 Online Factorization of Markov Chains

To handle the dependency of the Markov process, we need to downsample the data. Specifically, we divide the sequence of $n$ samples into $b$ blocks with block size $\tau$ for some $\tau \geq 2$:

$$s^{(1)}, s^{(2)}, \ldots, s^{(\tau)}, s^{(\tau+1)}, s^{(\tau+2)}, \ldots, s^{(2\tau)}, \ldots, s^{(b-1)\tau+1}, s^{(b-1)\tau+2}, \ldots, s^{(b\tau)}.$$

For the $k$-th block, we select the last two samples and define $Z^{(k)} \in \mathbb{R}^{m \times m}$

$$Z^{(k)}_{s^{(k\tau-1)}, s^{(k\tau)}} = 1 \quad \text{and} \quad Z^{(k)}_{j\ell} = 0 \text{ for all } (j, l) \neq (s^{(k\tau-1)}, s^{(k\tau)}). \quad (3)$$

Here we choose a large enough $\tau$ such that $\forall k \geq 1$, $E[Z^{(k)}|s^{(0)}] \approx DP = F_1 + F_2$, where $F_1 = U^T \Sigma V$ and $F_2$ are given in Definition 1. We will specify the choice of $\tau$ in Section 4.

Let us formulate the Stochastic Transition Matrix Decomposition Problem as

$$(U^*, V^*) = \arg\max_{U, V} \text{tr}[U^T \Sigma V] \quad \text{subject to} \quad U^T U = V^T V = I_r, \quad (4)$$

4
where the expectation

\[ \mathbb{E} Z := \lim_{n \to \infty} n^{-1} \sum_{j=1}^{n} Z(j) \]

is taken over the invariant distribution of the Markov chain. Note that \( U^* \) and \( V^* \) are global optima to (4), and they satisfy \( U^* = \overline{U} O \) and \( V^* = \overline{V} O \) for some orthonormal matrix \( O \in \mathbb{R}^{r \times r} \). By using a self-adjoint dilation, we denote

\[
A^{(k)} = \begin{bmatrix} 0_{m \times m} & Z^{(k)} \\ Z^{(k)\top} & 0_{m \times m} \end{bmatrix} \in \mathbb{R}^{2m \times 2m}, \quad \mathbb{E} A = \begin{bmatrix} 0_{m \times m} & \mathbb{E} Z \\ \mathbb{E} Z^\top & 0_{m \times m} \end{bmatrix} \in \mathbb{R}^{2m \times 2m}
\]

recast (4) into a symmetric decomposition problem as follows

\[
W^* = \arg\max_{W \in \mathbb{R}^{2mxr}} \text{tr}[W^\top \mathbb{E} A W] \quad \text{subject to} \quad W^\top W = I_r,
\]

(5)

where \( W = \frac{1}{\sqrt{2}} [U^\top, V^\top]^\top \in \mathbb{R}^{2mxr} \).

To solve (5), we adopt the generalized Hebbian algorithm (GHA) which was originally developed for training neural nets and principal component analysis. GHA, also referred as Sanger’s rule Sanger (1989), is essentially a stochastic primal-dual algorithm. Specifically, let

\[
L(W, L) = \text{tr}[W^\top \mathbb{E} A W] - \text{tr}[L(W^\top W - I_r)]
\]

be the Lagrangian function of (5), where \( L \in \mathbb{R}^{r \times r} \) is the Lagrangian multiplier matrix. By checking the KKT condition, we obtain

\[
\mathbb{E} A W^* + W^* L^* = 0 \quad \text{and} \quad W^* W - I_r = 0,
\]

(6)

where \( L^* \) is the optimal Lagrangian multiplier. The above KKT condition further implies

\[
L^* = -W^* \mathbb{E} A W^*.
\]

(7)

GHA is essentially inspired by (6) and (7). Specifically, at the \( k \)-th iteration, GHA takes

\[
\text{Dual Update} : \quad L^{(k)} = \frac{W^{(k)\top} A^{(k)} W^{(k)}}{\text{Markov sample of } W^{(k)\top} \mathbb{E} A W^{(k)}}
\]

(8)

\[
\text{Primal Update} : \quad W^{(k+1)} = W^{(k)} + \eta \frac{(A^{(k)} W^{(k)} - W^{(k)} L^{(k)})}{\text{Markov sample of } \nabla_W L(W^{(k)}, L^{(k)})}
\]

(9)

where \( \eta > 0 \) is the learning rate. Combing (8) with (9), we get a dual-free update of GHA as follows,

\[
W^{(k+1)} = W^{(k)} + \eta (A^{(k)} W^{(k)} - W^{(k)} W^{(k)\top} A^{(k)} W^{(k)}).
\]

We formally state the algorithm as follows:

In Algorithm 1, the initial solution \( W^{(0)} \) is drawn uniformly from the set of all orthonormal matrices. It does not require any warm-up initialization or prior information. Algorithm 1 makes update online and uses \( O(m r) \) space, while a batch method needs \( O(m^2) \) space to store the explicit transition matrix.
### Algorithm 1 SGD for Online Factorization of Markov Chains

**Output:** A stream of Markov transition data \( s^{(1)}, s^{(2)}, s^{(3)}, \ldots, s^{(n-1)}, s^{(n)} \).

**Initialize:** Sample matrix \( G \in \mathbb{R}^{2m \times r} \) with i.i.d. entries from \( \mathcal{N}(0, 1) \); \( W^{(0)} \leftarrow QR(G) \), \( k \leftarrow 0 \).

**Repeat:**

- For every \( \tau \) state transitions, obtain \( A^{(k)} \) using the two most recent states by Eq. (3);
- \( W^{(k+1)} \leftarrow W^{(k)} + \eta \left[ A^{(k)} W^{(k)} - W^{(k)} W^{(k)} \tau A^{(k)} W^{(k)} \right] \);
- \( k \leftarrow k + 1 \);

**Until** stopping condition is satisfied

**Output** \( [\hat{U}; \hat{V}] \leftarrow \sqrt{2} W^{(k)} \)

### 3.2 Recovering The Network Partition

Recall that the \( m \times r \) matrix \( M = D^{-1} V \) gives reduced-order representation for each vertex of the network. As long as we can estimate \( D, V \), we would be able to partition the network by applying a clustering algorithm such as the \( k \)-means. Let us describe the overall procedure:

1. Run Algorithm 1 on the Markov transition data and obtain \([\hat{U}; \hat{V}]\).
2. Let \( \hat{\mu} \) be the empirical estimate of the stationary distribution, i.e., \( \hat{\mu}_i = \sum_{j=1}^{n} I(s_j = i) / n \). Let \( \hat{D} = \text{diag}(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m) \). Now each row of \( \hat{M} = \hat{D}^{-1} \hat{V} \) gives an approximate \( r \)-dimensional representation for the corresponding state/vertex.
3. Find a set of centers \( C = [c_1, c_2, \ldots, c_r] \subset \mathbb{R}^r \) by solving the following problem:

   \[\begin{align*}
   \hat{C} &= \arg\min_{C} \sum_{i=1}^{m} \min_{c \in C} d^2(\hat{M}_{is}, c),
   \end{align*}\]  

   where \( d(\hat{M}_{is}, c_j) = \|\hat{M}_{is} - c_j\|_2 \).
4. Output the partition that assigns each state to its closest center.

### 4 Theory

For notational convenience, we denote \( \mu(\Omega) = \sum_{i \in \Omega} \mu_i \) for \( \Omega \subset S \). We introduce the merging conductance (Mihail, 1989) of a Markov chain by

\[\Phi = \min_{\Omega \subset [m]} \frac{\sum_{j \in \Omega} \sum_{i \in \Omega} \mu_i \mu_j P_{ij}}{\sum_{j \in \Omega} \mu_j} \text{ subject to } \mu(\Omega) \leq 1/2.\]

The parameter \( \Phi \) is a generalization of the Cheeger’s constant, which characterizes the bottleneck of a network. For recurrent Markov chains that are rapidly mixing, \( \Phi \) can be treated as a constant. Besides, we denote \( \mu_{\max} = \max_{i \in [m]} \mu_i \) and \( \mu_{\min} = \min_{i \in [m]} \mu_i \). Our first theorem establishes the finite-sample error bound for estimating the Markov chain factorization using Algorithm 1.
Theorem 3 (Sample Complexity for Learning from Markov Transition Data). Suppose that the Markov chain with transition matrix $P$ is nearly low-rank (as in Definition 1). Given a sufficiently small pre-specified error $\epsilon > 0$, we choose

$$\eta \approx \frac{e \left( \sigma_r(F_1) - \|F_2\|_2 \right)}{r} \quad \text{and} \quad \tau \approx \frac{2}{\Phi^2} \log \left( \sqrt{\frac{\mu_{\text{max}}}{\mu_{\text{min}}}} \frac{1}{\eta} \right).$$

Then Algorithm 1 takes

$$N = O \left( \max \left[ \frac{r \tau}{e \left( \sigma_r(F_1) - \|F_2\|_2 \right)^2} \log \frac{m r}{e \left( \sigma_r(F_1) - \|F_2\|_2 \right) \cdot e \mu_{\text{min}} \cdot \log m} \right] \right)$$

sample state transitions such that $\hat{U}$, $\hat{V}$, and $\hat{\mu}$ satisfy

$$\| \sin \Theta(\hat{U}, U) \|_F^2 + \| \sin \Theta(\hat{V}, V) \|_F^2 \leq \epsilon \quad \text{and} \quad \max_{i \in [m]} |\hat{\mu}_i - \mu_i| \leq \sqrt{\epsilon} \mu_i$$

with probability at least $3/4$.

The proof of Theorem 3 is sketched in Section 5 and given in Section A. The $3/4$ success probability can be boosted to $1 - \delta$ for arbitrary $0 < \delta < 1$ by repeating Algorithm 1 for $O\left( \log \frac{1}{\delta} \right)$ times (for more details, please refer to Appendix C). Theorem 3 has two implications:

1. For approximately lumpable and recurrent Markov chains, we have $\tau \approx \log \left( \frac{1}{\epsilon} \right)$, which is fairly small in practice;

2. To obtain an $\epsilon$-optimal solution, we need $N \approx \frac{1}{\epsilon} \log^2 \left( \frac{1}{\epsilon} \right)$ sample state transitions. This is consistent with the optimal sample complexity for independent principal component analysis Allen-Zhu and Li (2016), so this result is near-optimal.

To our best knowledge, this is the first result that works for Markov transition data.

Next we show that we can recover the underlying partition structure of lumpable Markov chains.

Theorem 4 (Recovery of Partition Structure for Lumpable Markov Chains). Under the assumptions of Theorem 3, then there exists an absolute constant $C$, such that with probability at least $3/4$, for any $s_i, s_j \in S$, we have

$$\| \hat{M}_{s_i} - M_{s_i} \|_2^2 - \| \hat{M}_{s_j} - M_{s_j} \|_2^2 \leq \frac{C \epsilon}{\mu_{\text{min}}}$$

where $\epsilon$ is the pre-specified error defined in Theorem 3. Moreover, the procedure of Section 3.2 exactly recovers the network partition with probability $3/4$, as long as the Markov chain is lumpable with metastates $S_1, \ldots, S_r$ and satisfies

$$\forall l, s_i \in S_l, s_j \in S^c_l : \| M_{s_i} - M_{s_j} \|_2^2 > \frac{2C \epsilon}{\mu_{\text{min}}^2}.$$

The proof is given in Section A.1. Theorem 4 implies that our proposed partition approach can exactly recover the partition of a lumpable Markov chain under mild conditions. It is possible to extend our analysis to approximately lumpable Markov chains, which is left for future research.
5 Proofs of Main Results

The proof of Theorem 3 takes two major steps: (1) We show that the dynamics of our algorithm can be approximated by an ordinary differential equation (ODE); (2) To analyze the sample complexity, we show that after proper rescaling of time, the algorithm can be characterized by a Stochastic Difference Equation (SDE). The SDE allows us to analyze the error fluctuation when the iterates are within a small neighborhood of the global optimum.

5.1 ODE Characterization of Algorithm 1

Let \( \bar{R} \in \mathbb{R}^{2m \times 2m} \) be the eigenvectors of \( \mathbb{E}A \) in (5). We consider a transformation by \( \bar{R} \):

\[
W^{(k)} = \bar{R}W^{(k)} \quad \text{and} \quad \bar{A}^{(k)} = \bar{R}^T A^{(k)} \bar{R}.
\]

Let \( \Lambda = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{2m}) = \mathbb{E}\bar{A} \) with that \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{2m} \). To demonstrate an ODE characterization for the trajectory of the algorithm, we introduce a continuous time \( t \). Recall where \( \eta \) is the learning rate. We denote \( \bar{W}(t) = \bar{W}^{(t/\eta)} \). For notation simplicity, we may drop \( (t/\eta) \) if it is clear from the context. For \( r+1 \leq i \leq 2m \), we define the subspace angle as

\[
\gamma_i^{(\eta)}(t) = \left\| \cos \Theta(\bar{R}_{si}, W) \right\|_F = \left\| e_i^T \bar{R}^T W \right\|_F = \left\| e_i^T \bar{W} \right\|_F,
\]

where \( e_i \in \mathbb{R}^{2m} \) is the \( i \)-th standard unit vector. We use \( \eta \) as a superscript to emphasize the dependence on \( \eta \). To show a global convergence of \( \gamma_i^{(\eta)}(t) \), we characterize its upper bound in the following lemma.

Lemma 5 (Principle Angle Upper Bound). Let \( E = (e_1, e_2, \ldots, e_r) \in \mathbb{R}^{2m \times r} \). Suppose that \( W \) has orthonormal columns and \( E^TW \) is full rank. For any \( X \in \mathbb{R}^{2m \times s} \) with \( s \geq 1 \), we have

\[
\left\| X^TW \right\|_F \leq \left\| X^TW \cdot (E^TW)^{-1} \right\|_F.
\]

Accordingly, we define an uniform upper bound of \( \gamma_i^{(\eta)} \) as \( \gamma_i^{(\eta)} = \left\| e_i^T \bar{W} \cdot (E^T \bar{W})^{-1} \right\|_F \). The next theorem establishes the continuous time limit for \( \gamma_i^{(\eta)} \).

Lemma 6 (ODE Convergence). Given \( \bar{W}^{(0)} \) such that \( E^T \bar{W}^{(0)} \) is invertible, under the limit of \( \eta \to 0 \), at each time \( t > 0 \), for all \( r < i \leq 2m \), \( \gamma_i^{(\eta)}(t) \) converges to the solution of the following ODE, \( \frac{d\gamma_i^{(\eta)}(t)}{dt} = b_i \gamma_i^{(\eta)}(t) \) with probability 1, where \( b_i \leq 2(\sigma_i - \sigma_r) \).

Lemma 6 essentially implies the global convergence of the algorithm. Specifically, the solution to the above ODE is

\[
\gamma_i(t) = \gamma_i(0) e^{b_i t/2} \leq \gamma_i(0) e^{(\sigma_i - \sigma_r)t}, \quad \forall \ r < i \leq 2m,
\]

which implies \( \gamma_i^{(\eta)}(t) \to 0 \) for any \( r < i \leq 2m \) as \( \eta \to 0 \) and \( t \to \infty \). Since

\[
\left\| \sin \Theta(E, \bar{W}(t)) \right\|_F^2 = \sum_{i>r} \gamma_i^{(\eta)}(t),
\]
we obtain
\[ \left\| \sin \Theta \left( \overline{U}(t), \overline{U} \right) \right\|_F^2 + \left\| \sin \Theta \left( \overline{V}(t), \overline{V} \right) \right\|_F^2 \leq 2 \left\| \sin \Theta \left( E, \overline{W}(t) \right) \right\|_F^2 \leq 2 \sum_{i>r} \gamma_i(0)e^{(\sigma_{i+1} - \sigma_i)t} \to 0. \] (12)

5.2 SDE Characterization of Algorithm 1

Our ODE approximation of the algorithm shows that after sufficiently many iterations, i.e., for any large enough \( t \), the algorithm solution can be arbitrarily close to the true subspace, \( \text{span}(\overline{R}_{s1}, \overline{R}_{s2}, \ldots, \overline{R}_{sr}) \).

To obtain the “rate of convergence”, however, we need to study the variance of the trajectory at time \( t \). Note that such a variance is of order \( \mathcal{O}(\eta) \), and vanishing under the limit of \( \eta \to 0 \). To characterize the variance, we need to rescale the updates by a factor of \( \eta^{-1/2} \), i.e., after rescaling, the variance is of order \( \mathcal{O}(1) \). Specifically, the rescaled update is defined as

\[ \zeta_i^{(\eta)}(t) = \eta^{-1/2} \overline{W}^{([t/\eta])^T} e_i \in \mathbb{R}^r. \]

Note that given \( W^{(k)} \) such that \( \text{span}(W^{(k)}) = \text{span}(\overline{R}_{s1}, \overline{R}_{s2}, \ldots, \overline{R}_{sr}) \), we have

\[ \mathbb{E}(W^{(k+1)}|W^{(k)}) = W^{(k)} \).

Namely, any matrix in \( \text{span}(\overline{R}_{s1}, \overline{R}_{s2}, \ldots, \overline{R}_{sr}) \) is a fixed point for Equation (16), in expectation. We consider a regime, where the algorithm has already run for sufficient many iterations such that

\[ \left\| \sin \Theta \left( R_{s1:r}, W^{(N_1)} \right) \right\|_F^2 = \left\| \sin \Theta \left( E, \overline{W}^{(N_1)} \right) \right\|_F^2 \leq \eta^{1/c}, \]

for some constant \( c > 1 \). By restarting the counter, we denote \( \overline{W}^{(0)} := \overline{R}^T W^{(N_1)} \). Now we define \( \overline{W}^{(0)^T} \Lambda \overline{W}^{(0)} = \Gamma^T \overline{\Lambda} \Gamma \), where \( \Gamma \in \mathbb{R}^{r \times r} \) is an orthonormal matrix and \( \overline{\Lambda} = \text{diag}(\sigma'_1, \sigma'_2, \ldots, \sigma'_r) \), with \( \sigma'_1 \geq \sigma'_2 \geq \ldots \geq \sigma'_r \geq 0 \).

Denote \( \zeta_{i,j}^{(\eta)}(t) = \eta^{-1/2} \left( e_j^T \overline{W}^{([t/\eta])^T} e_i \right) \), for \( i = r + 1, r + 2, \ldots, 2m \) and \( j = 1, 2, \ldots, r \), where \( e_j^r \in \mathbb{R}^r \) denotes the \( j \)-th standard unit vector in \( \mathbb{R}^r \). We establish the following lemma.

**Lemma 7 (SDE Convergence).** Given

\[ \left\| \sin \Theta \left( E, W^{([t/\eta])} \right) \right\|_F^2 \leq \mathcal{O}(\eta^{1/c}) \]

for all \( t \geq 0 \), then for any \( i > r \) and \( j \in [r] \), the trajectory of \( \zeta_{i,j}^{(\eta)}(t) \) converges to the solution of the following SDE in probability, as \( \eta \to 0 \),

\[ d\zeta_{i,j} = K_{i,j} \zeta_{i,j} dt + G_{i,j} dB_{i,j} \]

(13)

where \( B_{i,j} \) is the standard Brownian motion and constants

\[ K_{i,j} \leq (\sigma_i - \sigma_r), \quad \sum_{i>r} G_{i,j}^2 \leq B \quad \text{for any} \quad j \in [r], \]

with some absolute constant \( B \).
Notice that (13) is a Fokker-Plank equation, which admits the following solution,
\[
ζ_{i,j}(t) = ζ_{i,j}(0) \exp[K_{i,j}t] + G_{i,j} \int_0^t \exp[K_{i,j}(s-t)] dB_{i,j}(s).
\] (14)

Therefore, we show that each \(ζ^{(η)}_{i,j}(t)\) weakly converges to an Ornstein-Uhlenbeck (OU) process, which is widely studied in existing literature (Meucci, 2009). Since the drifting term is driven by \(K_{i,j} < 0\), the OU process eventually becomes a pure random walk, i.e., the first term of R.H.S. in (14) goes to 0. Recall that \(ζ^{(η)}_{i,j}(t)\) characterizes the sin angle of the subspaces, i.e.,
\[
\left\|\sin Θ(E, W^{(\lfloor t/η \rfloor)})\right\|_F^2 = η \sum_{i>r} \sum_{j=1}^{r} G_{i,j}^2 \int_0^t \exp[2K_{i,j}t] dt = O\left(\frac{ηr}{∥F_1∥ - ∥F_2∥}\right).
\] (15)

Thus the fluctuation of \(ζ^{(η)}_{i,j}(t)\) is essentially the error fluctuation of the algorithm after sufficiently many iterations. By (14), we obtain
\[
E \left\|\sin Θ(E, W^{(\lfloor t/η \rfloor)})\right\|_F^2 = η \sum_{i>r} \sum_{j=1}^{r} G_{i,j}^2 \int_0^t \exp[2K_{i,j}t] dt = O\left(\frac{ηr}{∥F_1∥ - ∥F_2∥}\right).
\]

Given the error parameter \(ε > 0\), we need η to satisfy
\[
O\left(\frac{ηr}{∥F_1∥ - ∥F_2∥}\right) ≤ ε.
\]

Combining with a Markov inequality and Equation (12), we obtain the following lemma,

**Lemma 8 (Sample Complexity).** Given a sufficiently small \(ε > 0\), with at most
\[
N = O\left(\frac{rB}{ε (σ_r(F_1) - ∥F_2∥)^2} \log \frac{∑_{i>r} γ_i^2(0) \cdot B}{ε (σ_r(F_1) - ∥F_2∥)^2}\right)
\]
iterations, we have
\[
P\left[\left\|\sin Θ(\bar{U}, \bar{V})\right\|_F^2 + \left\|\sin Θ(\bar{V}, \bar{U})\right\|_F^2 > ε\right] ≤ \frac{1}{10}.
\]

**Proof of Theorem 3:** With standard characterizations of the random matrices (e.g. Tao (2012)), we obtain the value of \(∑_{i>r} γ_i^2(0) = \text{poly}(m)\) with probability close to 1 when \(m\) is large. Combining with the above Lemmas, we prove the first part of (11). The second part of (11) follows from standard results of Markov chain (e.g. Levin et al. (2009)). This concludes the proof of Theorem 3.

### 6 Experiments

We experiment with the proposed method on two data sets.
Figure 1: (a) An illustration of lumpable Markov chain: the full network (Top) and 3 meta-states in the simplified network (Bottom); (b) The convergence in subspace angle of 100 simulations: fixed stepsize (Top) and diminishing step size (Bottom); (c) Time complexity, nonconvex SGD v.s. convex MSG: fixed stepsize (Top) and diminishing step size (Bottom).

6.1 Simulated Data

Consider a 12-vertex lumpable network that has 3 meta-states (see Figure 1(a)). We let its probability transition matrix be $P_{ex}$ (which is specified in (21) of Appendix B). We generate the Markov transition data by simulating the random walk according to $P_{ex}$. We test our algorithms on the simulated transition data using $10^4$ samples for 100 independent trials. In every single trial, the meta states are correctly recovered. Figure 1(b) shows that the convergence rate of subspace angle are consistent with Theorem 3. For comparison, we test a convex relaxation algorithm (MSG Arora et al. (2016)) for online solution of Problem (4) (there is no theoretic guarantee for applying MSG to depend data). Figure 1(c) shows that our algorithm significantly outperforms MSG in wall-clock time complexity using both fixed and diminishing stepsizes.

6.2 Manhattan Taxi Data

We experiment using a real dataset that contains $1.1 \times 10^7$ trip records of NYC Yellow cabs from January 2016 (TLC, 2017). Each entry records the coordinates of the pick-up and drop-off locations, distance and length of trip, and taxifares. We discretize the map into a fine grid (with cell size roughly 10m) and model each taxi trip as a single state transition of a Markov chain. For example, a taxi picks up a customer at cell $s_1$ and drops off the customer at cell $s_2$, then picks up a customer at $s_3$... We can view $s_1, s_2, s_3, ...$ as the path of states visited by an implicit city-wide random walk. In order to guarantee recurrence of the random walk, we preprocess data by dropping cells that are rarely visited. We end up with 2017 locations and a total of $10^7$ effective trips. We apply Algorithm 1 and the partition procedure of Section 3.2 to the taxi trip data and illustrate the results in Figure 2(a)–2(c). Our method reveals a very informative partition of the Manhattan city according to traffic dynamics. We compare our online algorithm with a batch partition procedure and observe that they generate highly similar results (Figure 2(d)) when $r = 4$. A practically impressive observation is: our algorithm uses less than 1 Mbytes memory for $r = 4, 10, 15$. 
7 Conclusions and Discussions

We have developed an online learning method for analyzing dynamic transition data, which outputs the low-dimensional representation of an implicit network and reviews its underlying partition structure. Our method has superior space and computation complexity. We show that it achieves near-optimal global convergence and sample complexity by using an ODE-SDE argument.

Our algorithm and analysis is related to online singular value decomposition methods that have been considered by Jain et al. (2016); Li et al. (2016); Allen-Zhu and Li (2016); Xie et al. (2015); Arora et al. (2016); Chen et al. (2017). We summarize the improvements of our results:

1. All existing analyses require independent samples, while ours applies to dependent Markov samples;

2. Jain et al. (2016); Li et al. (2016); Allen-Zhu and Li (2016); Chen et al. (2017) analyzed Oja's algorithm for PCA problem, which conducts QR factorization in each iteration. Our algorithm does not require such decomposition - each iteration uses only vector-to-vector inner products;

3. Xie et al. (2015) analyzed a similar algorithm as ours but their results require a sufficiently near-optimal initial solution, which is not available in our problems;
Arora et al. (2016) investigated a method based on convex relaxation of SVD but they achieve
a sub-optimal sample complexity - $\tilde{O}(1/\epsilon^2)$. Moreover, their algorithm needs to compute an
expensive Fantope projection with a computational complexity $O(m^3)$ per iteration.

In summary, Markov transition data carries rich information about the underlying structure
of complex networks and stochastic systems. We hope this work will motivate more research in
this area and faster algorithms for applications in networks and reinforcement learning.

References

Allen-Zhu, Z. and Li, Y. (2016). First efficient convergence for streaming k-pca: a global, gap-free,
and near-optimal rate. arXiv preprint arXiv:1607.07837.

Arora, R., Mianjy, P. and Marinov, T. (2016). Stochastic optimization for multiview representa-
tion learning using partial least squares. In Proceedings of The 33rd International Conference on
Machine Learning.

Benson, A. R., Gleich, D. F. and Lim, L.-H. (2016). The spacey random walk: A stochastic process
for higher-order data. arXiv preprint arXiv:1602.02102.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of commu-
unities in large networks. Journal of statistical mechanics: theory and experiment 2008 P10008.

Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interac-
tion network. Bioinformatics 22 2283–2290.

Chen, Z., Yang, L. F., Li, C. J. and Zhao, T. (2017). Online multiview representation learning:
Dropping convexity for better efficiency. arXiv preprint arXiv:1702.08134.

Chung, F. R. (1997). Spectral graph theory, vol. 92. American Mathematical Soc.

Cohen, M. B., Lee, Y. T., Miller, G., Pachocki, J. and Sidford, A. (2016). Geometric median
in nearly linear time. In Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of
Computing. ACM.

Ethier, S. N. and Kurtz, T. G. (2009). Markov processes: characterization and convergence, vol. 282.
John Wiley & Sons.

Harold, J., Kushner, G. and Yin, G. (1997). Stochastic approximation and recursive algorithm
and applications. Application of Mathematics 35.

Jain, P., Jin, C., Kakade, S. M., Netrapalli, P. and Sidford, A. (2016). Streaming pca: Matching
matrix bernstein and near-optimal finite sample guarantees for ojas algorithm. In 29th Annual
Conference on Learning Theory.

Junker, B. H. and Schreiber, F. (2011). Analysis of biological networks, vol. 2. John Wiley & Sons.
Levin, D. A., Peres, Y. and Wilmer, E. L. (2009). *Markov chains and mixing times*. American Mathematical Soc.

Li, C. J., Wang, M., Liu, H. and Zhang, T. (2016). Near-optimal stochastic approximation for online principal component estimation. *arXiv preprint arXiv:1603.05305*.

Liu, Y., Kang, C., Gao, S., Xiao, Y. and Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of geographical systems* 14 463–483.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* 11 193–197.

Meucci, A. (2009). Review of statistical arbitrage, cointegration, and multivariate ornstein-uhlenbeck.

Mihail, M. (1989). Conductance and convergence of markov chains-a combinatorial treatment of expanders. In *Foundations of computer science, 1989., 30th annual symposium on*. IEEE.

Minsker, S. et al. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli* 21 2308–2335.

Nerur, S., Sikora, R., Mangalaraj, G. and Balijepally, V. (2005). Assessing the relative influence of journals in a citation network. *Communications of the ACM* 48 71–74.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98 404–409.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab.

Peebles, P. J. E. (1980). *The large-scale structure of the universe*. Princeton university press.

Pizzuti, C. (2008). Ga-net: A genetic algorithm for community detection in social networks. *Parallel problem solving from nature–PPSN X* 1081–1090.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks* 2 459–473.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge.

Tao, T. (2012). *Topics in random matrix theory*, vol. 132. American Mathematical Society Providence, RI.

TLC, N. (2017). Nyc taxi and limousine commission (tlc) trip record data. URL *http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml*

Weinan, E., Li, T. and Vanden-Eijnden, E. (2008). Optimal partition and effective dynamics of complex networks. *Proceedings of the National Academy of Sciences* 105 7907–7912.
We first assume that the sample $A^{(k)}$s are i.i.d. to each other. In particular, we assume $s^{(i)}$ is generated from the stationary distribution $\mu$ and $s^{(i+1)}$ is generated from distribution $P_{\theta^{(i)}}$. Later on, we will remove this requirement to prove Theorem 3. We denote $D$ as the distribution for the i.i.d samples described above. It can be verified that $D$ satisfies the following two properties.

**Assumption 1** (Subgaussian moments). Suppose $A \sim D$. Then for each $r \in \mathbb{N}_+$,

$$\mathbb{E}A = \Sigma \quad \text{and} \quad \mathbb{E}\|A\|_2^r \leq C_1,$$

for some constants $C_1$ dependent only on $m$.

For our distribution $D$, we observe that $\Sigma = \begin{bmatrix} 0_{m \times m} & DP \\ P^T D^T & 0_{m \times m} \end{bmatrix} \in \mathbb{R}^{2m \times 2m}$.

**Assumption 2** (Rotational Invariant Variance). Suppose $A \sim D$. Then for any set of orthonormal vectors $v_1, v_2, \ldots, v_m$,

$$\sum_{j=2}^m \mathbb{E}(v_i^T A v_j)^2 \leq B,$$

for some constant $B$.

We also introduce the following notations,

$$E_r = (e_1, e_2, \ldots, e_r) \in \mathbb{R}^{2m \times r} \quad \text{and} \quad \ov{E} = (e_{r+1}, e_{r+2}, \ldots, e_{2m}) \in \mathbb{R}^{2m \times (2m-r)}$$

where $e_i$ is the $i$-th standard unit vector in $\mathbb{R}^m$. Let

$$\Sigma = \ov{R} \Lambda \ov{R}^T$$

be the eigenvalue decomposition of $\Sigma$, i.e., $\ov{R} = (v_1, v_2, \ldots, v_{2m})$ are the eigenvectors of $\Sigma$ and $\Lambda = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m, -\sigma_m, -\sigma_{m-1}, \ldots, -\sigma_1)$. Initialize with a random matrix $W^{(0)} \in \mathbb{R}^{2m \times r}$ with orthonormal columns. Let $\eta > 0$ be the choice of the learning rate. The update for step $k$ is

$$W^{(k+1)} \leftarrow W^{(k)} + \eta \left( A^{(k)} W^{(k)} - W^{(k)} \cdot W^{(k)}^T A^{(k)} W^{(k)} \right).$$

Let $\ov{W} = \ov{R}^T W$ and $\ov{A} = \ov{R}^T A \ov{R}$, we obtain,

$$\ov{W}^{(k+1)} \leftarrow \ov{W}^{(k)} + \eta \left( \ov{A}^{(k)} \ov{W}^{(k)} - \ov{W}^{(k)} \cdot \ov{W}^{(k)}^T \ov{A}^{(k)} \ov{W}^{(k)} \right). \quad (16)$$
Proof of Lemma 5. Since $E$ and $W$ have orthonormal rows, we have that $\|E^TW\|_2 \leq 1$. Then we immediately have
\[
\|X^TW\|_F \leq \|X^T \cdot (E^TW)^{-1} (E^TW)\|_F \leq \|X^T \cdot (E^TW)^{-1}\|_F \|E^TW\|_2,
\]
as desired. \qed

Proof of Lemma 6 (On Independent Sample). First we note that if $E^T W^{(k)}$ is invertible, then, for sufficiently small $\eta$, $E^T W^{(k+1)}$ is invertible with probability 1. Further we will show that $\tilde{\gamma}_i^{(n)}(t)$ is monotonically decreasing, which implies $\tilde{\gamma}_i^{(n)}(t)$ is finite for all $t > 0$. Thus $E^T W^{(k)}$ is invertible for all $k > 0$.

It is easy to verify that $(W^{(k)}, \tilde{\gamma}_i^{(n)}(\eta k))_{k=1}^n$ form a sequence satisfying strong Markov property. By Corollary 4.2 in §7.4 of Ethier and Kurtz (2009), if
\[
b_i = \lim_{\eta \to 0} \mathbb{E} \left[ \frac{\Delta \tilde{\gamma}_i^{(n)}(\eta)}{\eta} \right] \mathbb{E} \left[ W \right] \leq \infty \quad \text{and} \quad \sigma_i^2 = \lim_{\eta \to 0} \mathbb{E} \left[ \frac{\Delta \tilde{\gamma}_i^{(n)}(\eta)}{\eta} \right] \mathbb{E} \left[ W \right] = 0,
\]
where
\[
\Delta \tilde{\gamma}_i^{(n)}(\eta) = \tilde{\gamma}_i^{(n)}(t + \eta) - \tilde{\gamma}_i^{(n)}(t) = \|e_i^T (W + \Delta W)(E^T W + E^T \Delta W)^{-1}\|^2_2 - \|e_i^T (W)(E^T W)^{-1}\|^2_2.
\]
then the sequence of updates converges to the solution of the following ODE in probability
\[
d\tilde{\gamma}_i^2(t) = b_i dt.
\]
By a simple calculation,
\[
\Delta \tilde{\gamma}_i^{(n)}(\eta) = 2e_i^T \Delta W (E^T W)^{-1} (W^T E)^{-1} W^T e_i \\
- 2e_i^T W (E^T W)^{-1} (E^T \Delta W) (E^T W)^{-1} (W^T E)^{-1} W e_i + O(\eta^2 \|A\|_2^2);
\]
\[
\left[\Delta \tilde{\gamma}_i^{(n)}(\eta)^2\right] = O(\eta^2 \|A\|_2^2).
\]
Plugging (16) in to $\Delta \tilde{W}$, we have,
\[
\frac{1}{\eta} \mathbb{E} \left[ \frac{\Delta \tilde{\gamma}_i^{(n)}(\eta)}{\eta} \right] \mathbb{E} \left[ W \right] = 2\sigma_i e_i^T \tilde{W} \left( E^T W \right)^{-1} \left( W^T E \right)^{-1} W^T e_i \\
- 2e_i^T \tilde{W} \left( E^T W \right)^{-1} \left( E \Lambda W \right) \left( E W \right)^{-1} \left( W^T E \right)^{-1} W e_i + O(\eta);
\]
\[
= 2\sigma_i e_i^T \tilde{W} \left( E^T W \right)^{-1} \left( W^T E \right)^{-1} W^T e_i \\
- 2e_i^T \tilde{W} \left( E^T W \right)^{-1} \left( \Lambda e_i^T W \right) \left( E^T W \right)^{-1} \left( W^T E \right)^{-1} W e_i + O(\eta);
\]
\[
\leq 2(\sigma_i - \sigma_r) e_i^T \tilde{W} \left( E^T W \right)^{-1} \left( W^T E \right)^{-1} W^T e_i + O(\eta)
\]
\[
= 2(\sigma_i - \sigma_r) \gamma_i^2 + O(\eta);
\]
(17)
and

$$E\left(\left|\Delta \left(\tilde{\gamma}_i^{(\eta)^2}\right)^2\right|W\right) = O(\eta^2).$$

(18)

Thus we have

$$b_i \leq 2(\sigma_i - \sigma_r)\gamma_i^2 \quad \text{and} \quad \sigma_i = 0,$$

as desired.

\[\square\]

**Lemma 9.** Suppose $W \in \mathbb{R}^{2m \times r}$ is a matrix with orthonormal columns. If $\|E^TW\|_F^2 \leq O(\delta)$, then

$$\sigma_{\min}(W^{(0)^T} \Lambda W^{(0)}) \geq \sigma_r - O(\delta).$$

**Proof of Lemma 9.** Since $\|E^TW\|_F^2 \geq r - O(\delta)$, for any $1 \leq i \leq r$, we have

$$\|We_i\|_2^2 \geq 1 - O(\delta),$$

where $e_i' \in \mathbb{R}^r$ is the $i$-th $r$-dimensional standard unit vector. Therefore,

$$e_i'^TW^T \Lambda We_i = e_i'^TW^T (E \Lambda_r E^T + \bar{E} \Lambda_r \bar{E}^T) We_i'$$

$$\geq \sigma_i e_i'^TW^T E E^T We_i' - O(\delta)$$

$$\geq \sigma_r - O(\delta),$$

Here $\Lambda_r = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ and $\bar{\Lambda}_r = \text{diag}(\sigma_{r+1}, \sigma_{r+2}, \ldots, \sigma_{2m})$. Hence,

$$\sigma_{\min}(W^T \Lambda W) \geq \sigma_r - O(\delta).$$

\[\square\]

**Proof of Lemma 7 (On independent sample).** Let $\tilde{\zeta}_{i,j} = \lim_{\eta \to 0} \eta^{-1/2} \left( e_j'^T \Gamma W^{(l/\eta)^T} e_i \right)$. Then

$$d\tilde{\zeta}_{i,j} = \lim_{\eta \to 0} \Delta \tilde{\zeta}_{i,j}^{(\eta)}.$$ 

Hence

$$\Delta \tilde{\zeta}_{i,j}^{(\eta)} = \left( e_j'^T \Gamma \Delta W e_i \right) \eta^{1/2}$$

$$= \eta^{1/2} \left[ e_j'^T \Gamma \left( W W^T - W^T \bar{A} W \cdot W^T \right) e_i \right].$$

Therefore, we have

$$\frac{1}{\eta^{1/2}} E \left( \Delta \tilde{\zeta}_{i,j}^{(\eta)} | W \right) = \eta^{1/2} E \left[ e_j'^T \left( \Gamma \bar{W} W e_i - \Gamma W^T \bar{A} \bar{W} W^T e_i \right) \right]$$

$$= \eta^{1/2} \left[ \sigma_i e_j'^T \Gamma W e_i - e_j'^T \Gamma \bar{W} \bar{A} \bar{W} W^T e_i \right]$$

$$= \eta^{1/2} \left[ \sigma_i e_j'^T \Gamma W e_i - \sigma_j' e_j'^T \Gamma \bar{W} W^T e_i \right]$$

$$\leq \eta (\sigma_i - \sigma_r) \tilde{\zeta}_{i,j}^{(\eta)} + O(\eta \delta).$$
We now turn to compute the infinitesimal variance $G_{i,j}$.

$$
\left( \Delta \kappa_{i,j}^{(\eta)} \right)^2 = \frac{1}{\eta} \left( e_j^T \Gamma \Delta W^T e_i \right)^2.
$$

And

$$
\mathbb{E}\left[ (e_j^T \Gamma \Delta W^T e_i)^2 \right] = \eta^2 \mathbb{E}\left[ (e_j^T \Gamma \Delta W^T (I - \bar{W}' \bar{W}) e_i)^2 \right].
$$

We further observe that $\Gamma \Delta W^T e_j'$ is perpendicular to $(I - \bar{W}' \bar{W}) e_i$ for any $r < i \leq 2m$, and that

$$
\left\| (I - \bar{W}' \bar{W}) e_i \right\| \leq 1.
$$

Thus, for any $j = 1, 2, \ldots, r$,

$$
\sum_{i=r+1}^{2m} \mathbb{E}\left[ (e_j^T \Gamma \Delta W^T (I - \bar{W}' \bar{W}) e_i)^2 \right] \leq B.
$$

By Corollary 4.2 in §7.4 of Ethier and Kurtz (2009), we obtain the desired SDE with $K_{i,j} \leq (\sigma_i - \sigma_r)$ and $\sum_{i>r}^{2m} G_{i,j}^2 \leq B$.

Proof of Lemma 8 (On Independent Sample). First of all, after $N_1$ steps, the algorithm has a solution $W^{(N_1)}$ with that

$$
\mathbb{E}\left[ \left\| \bar{R}' W^{(N_1)} \right\|_F^2 \right] = O(\eta^{1/c_0}).
$$

By Markov’s inequality, with probability at least $1 - c_1$ for arbitrarily $c_0 \geq 1$ and $0 < c_1 \leq 1$,

$$
\left\| \bar{R}' W^{(N_1)} \right\|_F^2 = O(\eta^{1/c_0}).
$$

Starting from $N_1$, the trajectory of $\zeta_{i,j}$ can be represented by the SDE (13). The SDE is a standard Ornstein-Uhlenbeck process, whose solution is

$$
\zeta_{i,j}(t) = \zeta_{i,j}(0) e^{K_{i,j} t} + G_{i,j} \int_0^t e^{K_{i,j} (t-s)} dB_{i,j}(s).
$$

We have

$$
\mathbb{E}[\zeta_{i,j}(t)] = \zeta_{i,j}(0) e^{K_{i,j} t}
$$

and

$$
\mathbb{E}[\zeta_{i,j}^2(t)] = \frac{G_{i,j}^2}{K_{i,j}} + \left[ \zeta_{i,j}(0)^2 - \frac{G_{i,j}^2}{K_{i,j}} \right] e^{K_{i,j} t} = O\left( \frac{G_{i,j}^2}{K_{i,j}} \right)
$$

for $t = O\left( \frac{1}{K_{i,j}} \right)$. Consider another $N_2 = O\left( \frac{\eta^{-1}}{K_{i,j}} \right)$ steps. Let $E$ be the following event,

$$
E = \left\{ \sum_{j=1}^k \sum_{i>k} \zeta_{i,j}^2(N_2 \eta) \leq \frac{c}{\eta} \right\}.
$$
Notice that
\[
\mathbb{E} \left[ \sum_{j=1}^{r} \sum_{i>r}^d \xi_{i,j}^2 (N_2 \eta) \right] = \mathcal{O} \left( \sum_{i,j} \frac{G_{i,j}^2}{K_{i,j}} \right) = \mathcal{O} \left( \frac{kB}{\sigma_r - \sigma_{r+1}} \right).
\]

By Markov's inequality, we have
\[
\Pr[\mathcal{E}] \leq \Pr \left[ \sum_{j=1}^{r} \sum_{i>r}^d \xi_{i,j}^2 (N_2 \eta) > \varepsilon \right] \leq \frac{rB\eta}{\varepsilon (\sigma_r - \sigma_{r+1})}.
\]

By setting
\[
\eta = \mathcal{O} \left( \frac{\varepsilon (\sigma_r - \sigma_{r+1})}{rB} \right)
\]
we have
\[
\Pr[\mathcal{E}] \leq \frac{1}{5}.
\]

Therefore, after total number of steps
\[
N = N_1 + N_2 = \mathcal{O} \left( \frac{kB}{\varepsilon (\sigma_r - \sigma_{r+1})^2} \log \frac{\sum_{i>r} \gamma_i^2 (0) \cdot B}{\varepsilon (\sigma_r - \sigma_{r+1})} \right)
\]  \hspace{1cm} (19)

we obtain the desired bound.

Before we prove the theorem for Markov samples, we first show a theorem for fully independent samples.

**Theorem 10** (Convergence Results for Independent Samples). Let $DP = F_1 + F_2$ be the probability transition matrix satisfying Assumption 1. Let $F_1 = V^\top \Sigma V$ with $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$. Let $0 < \varepsilon \ll 1$ be a sufficiently small parameter, and

\[
\eta \sim \Theta \left( \frac{\varepsilon (\sigma_r (F_1) - \|F_2\|)}{r} \right).
\]

Then Algorithm 1 takes
\[
N = \mathcal{O} \left\{ \max \left[ \frac{r}{\varepsilon (\sigma_{\min} (F_1) - \sigma_{\max} (F_2))^2} \log \frac{d r}{\varepsilon (\sigma_{\min} (F_1) - \sigma_{\max} (F_2)) \cdot \log d} \right] \right\}
\]
samples such that with probability at least $3/4$, $\hat{U}, \hat{V}, \hat{\mu}$ satisfy
\[
\| \sin \Theta(U, U) \|_F^2 + \| \sin \Theta(V, V) \|_F^2 \leq \varepsilon \quad \text{and} \quad \forall i \in [|S|] : |\hat{\mu}_i - \mu_i| \leq \sqrt{\varepsilon} \mu_i.
\]

**Proof.** We start with the first part of the theorem. By Lemma 8, we only need to calculate $\sum_{i>r} \gamma_i^2 (0)$ for a random initial. By standard random matrices theory (e.g. Tao (2012)), for randomly chosen $U(0)$ and $V(0)$, with constant probability arbitrarily close to 1,
\[
\sum_{i>k} \gamma_i^2 (0) = \text{poly}(mr).
\]
Plugging in to the equation (19), we obtain, for

$$N = O\left[\frac{r}{\epsilon (\sigma_r(F_1) - \|F_2\|_2)^2} \log \frac{mr}{\epsilon (\sigma_r(F_1) - \|F_2\|_2)}\right]$$

the output \(\hat{U}, \hat{V}\) satisfies

$$\|U^T \hat{U}\|_F^2 + \|V^T \hat{V}\|_F^2 \geq 2r - \epsilon.$$ 

By definition of \(\sin^2 \Theta\), we obtained the desired argument.

For the second part, we first consider a fixed \(i \in [m]\). Using Chernoff bound, for \(N \mu_i = \Omega\left(\frac{\log m}{\epsilon}\right)\)

$$\Pr[|\hat{\mu}_i - \mu_i| > \sqrt{\epsilon \mu_i}] \leq \frac{1}{\text{poly}(m)}.$$ 

Using union bound for all \(i \in [m]\), we obtain that

$$\Pr[\exists i: |\hat{\mu}_i - \mu_i| > \sqrt{\epsilon \mu_i}] \leq \frac{1}{\text{poly}(m)}$$

as desired. \(\square\)

We now proceed to present the results for dependent data. Before the proof of Theorem 3 we first present the mixing lemma of Markov chain.

**Lemma 11** (Markov Chain Mixing Time, Mihail (1989)). Suppose \(P \in \mathbb{R}^{m \times m}\) is the transition matrix of a Markov chain. Let \(\mu\) be the stationary distribution of the chain. Denote

$$\Phi = \min_{\Omega \subset [m]} \frac{\sum_{j \in \Omega, \ell \in \Omega} \sum_{i \in [m]} \frac{\mu_j p_{i,j}}{\mu_i} \frac{\mu_i p_{\ell,i}}{\mu_j}}{\sum_{j \in \Omega} \mu_j} \text{ subject to } \mu(\Omega) \leq 1/2,$$

as the merging conductance of the Markov chain. Then the mixing time, defined as

$$\tau(\epsilon) = \min_t \left\{ t: \forall t' > t, \max_{i,j} |(P^{t'})_{i,j} - \mu_j| \leq \epsilon \right\},$$

satisfies,

$$\tau(\epsilon) \leq \frac{2}{\Phi^2} \log \left( \frac{\mu_{\max}}{\mu_{\min}} \frac{1}{\epsilon} \right).$$

We now proceed with the proof of Theorem 3.

**Full Proof of Theorem 3.** Let \(s^{(0)}, s^{(1)}, s^{(2)}, \ldots, s^{(k)}, \ldots, s^{(n)}\) be samples from the Markov chain. We next show that the statements in Lemma 6 and Lemma 7 are true for Markov Samples: \(s^{(\tau-1)}, s^{(\tau)}, s^{(2\tau-1)}, s^{(2\tau)}, \ldots, s^{(k\tau-1)}, s^{(k\tau)}\), where \(\tau = \tau(\eta)\). We start with Lemma 6. We first observe that the updates of the Algorithm combining with Markov samples satisfy strong Markov property. Denote

$$A^{(i)} = \begin{bmatrix} 0 & Z^{(k)} \\ Z^{(k)\top} & 0 \end{bmatrix} \in \mathbb{R}^{2m \times 2m}.$$
Assumption 1 and Assumption 2 can be verified, conditioning on \( s^{(0)} \). By definition of \( \tau = \tau(\eta) \), we also observe that,

\[
\mathbb{E}(A^{(k)}|s^{(0)}) = \begin{bmatrix}
0 & DP + \tilde{E}P \\
P^\top D + P^\top \tilde{E}
\end{bmatrix} \in \mathbb{R}^{2m \times 2m}.
\]

where \( \tilde{E} \) is a diagonal matrix with

\[
\forall i \in \{m\} : |\tilde{E}_{i,i}| \leq \eta.
\]

Therefore, the spectrum of \( \mathbb{E}(A^{(k)}) \) differs with independent case by at most an additive \( \mathcal{O}(\eta) \) term. Thus Equation (17) holds, hence the rest of the proof of Lemma 6 is exactly the same as proof of the independent case.

Lemma 7 follows similarly by observing that the spectrum of \( \mathbb{E}(A^{(i)}) \) differs with independent case by at most an additive \( \mathcal{O}(\eta) \) term.

The rest of proof of this theorem is the same as that of Theorem 10. \( \square \)

A.1 Proof of Theorem 4

Proof. By Theorem 3, with probability at least 3/4, we have

\[
\left\| \tilde{V}^\top \tilde{V} \right\|_F^2 = \sum_{i,j=1}^r \left( \sum_{i=1}^m v_{i,j} \tilde{v}_{i,j} \right)^2 = \sum_{i,j=1}^r \left( \sum_{i=1}^r v_{i,j} \tilde{v}_{i,j} \right)^2 \geq r - \epsilon \quad \text{and} \quad \tilde{\mu}_i \in (1 \pm \epsilon') \mu_i.
\]

We condition on this event in the following proof. Let

\[
\tilde{V}^\top \tilde{V} = \tilde{U}^\top T \tilde{U},
\]

where \( T = \text{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_r) \) and \( \tilde{U}, \tilde{U}' \) are orthonormal matrices. Since for each \( i, |\tilde{\sigma}_i| \leq 1 \), we have

\[
r - \epsilon \leq \sum_{i=1}^r \tilde{\sigma}_i^2 \leq r.
\]

Fixing \( \sum_{i=1}^r \tilde{\sigma}_i^2 \), the minimum value of \( \sum_{i=1}^r \tilde{\sigma}_i \) is obtained when \( \tilde{\sigma}_1 = \tilde{\sigma}_2 = \ldots = \tilde{\sigma}_{r-1} = 1 \) and \( \tilde{\sigma}_r = 1 - \epsilon \).

Let \( \tilde{V} = V \tilde{U}^\top \) and \( \tilde{V} = \tilde{V} \tilde{U}^\top \). We have

\[
\sum_{i=1}^m \tilde{V}_{i}^\top \tilde{V}_{i} = \sum_{i=1}^m \sum_{j=1}^r \tilde{V}_{i,j} \tilde{V}_{i,j} = \sum_{j=1}^r \tilde{\sigma}_j \in [r - \epsilon, r]. \tag{20}
\]

For any \( s_i, s_j \in S \), the distance

\[
d(s_i, s_j) = \left\| \frac{V_{i}}{\mu_i} - \frac{V_{j}}{\mu_j} \right\| \quad \text{and} \quad d'(s_i, s_j) = \left\| \frac{\tilde{V}_{i}}{\mu_i} - \frac{\tilde{V}_{j}}{\mu_j} \right\|.
\]

Therefore, by triangle inequality,

\[
d'(s_i, s_j) = \left\| \frac{\tilde{V}_{i}}{\mu_i} - \frac{\tilde{V}_{i}}{\mu_i} + \frac{\tilde{V}_{i}}{\mu_i} - \frac{\tilde{V}_{j}}{\mu_j} + \frac{\tilde{V}_{j}}{\mu_j} - \frac{\tilde{V}_{j}}{\mu_j} \right\| \leq \left\| \frac{\tilde{V}_{i}}{\mu_i} - \frac{\tilde{V}_{j}}{\mu_j} \right\| + \left\| \frac{\tilde{V}_{j}}{\mu_j} - \frac{\tilde{V}_{j}}{\mu_j} \right\|.
\]

21
Now consider the second term,

\[
\left\| \frac{\hat{V}_{i} - \hat{V}_{j}}{\hat{\mu}_{i}} - \frac{\hat{V}_{i} - \hat{V}_{j}}{\hat{\mu}_{j}} \right\| \leq \left\| \frac{\hat{V}_{i}}{\hat{\mu}_{i}} - \frac{\hat{V}_{i}}{\hat{\mu}_{j}} \right\| + \left\| \frac{\hat{V}_{j}}{\hat{\mu}_{j}} + \hat{V}_{j} \right\|
\]

Since we have \( \hat{\mu}_{i} = (1 \pm \epsilon') \mu_{i} \), where \( \epsilon' = \sqrt{\epsilon} \),

\[
\left\| \frac{\hat{V}_{i} - \hat{V}_{j}}{\hat{\mu}_{i}} - \frac{\hat{V}_{i} - \hat{V}_{j}}{\hat{\mu}_{j}} \right\|^{2} \leq \frac{16}{\mu_{i}^{2}} \left\| \hat{V}_{i} - \hat{V}_{i} \right\|^{2} + \frac{16}{\mu_{j}^{2}} \left\| \hat{V}_{j} - \hat{V}_{j} \right\|^{2} + \frac{16\epsilon'^{2}}{\mu_{i}^{2}} \left\| \hat{V}_{i} \right\|^{2} + \frac{16\epsilon'^{2}}{\mu_{j}^{2}} \left\| \hat{V}_{j} \right\|^{2}.
\]

Observe that, for any \( i \in [n] \),

\[
\left\| \hat{V}_{i} - \hat{V}_{i} \right\|^{2} = \left\| \hat{V}_{i} \right\|^{2} + \left\| \hat{V}_{i} \right\|^{2} - 2 \hat{V}_{i}^{T} \hat{V}_{i}.
\]

Therefore

\[
\sum_{i=1}^{m} \left\| \hat{V}_{i} - \hat{V}_{i} \right\|^{2} = 2r - 2 \sum_{i=1}^{m} \hat{V}_{i}^{T} \hat{V}_{i} = 2r - 2 \sum_{j=1}^{r} \tilde{\sigma}_{j} \leq 2\epsilon.
\]

Thus, for each \( i \in [n] \)

\[
\left\| \hat{V}_{i} - \hat{V}_{i} \right\|^{2} \leq 2\epsilon.
\]

Furthermore, since

\[
\epsilon'^{2} \leq \epsilon
\]

we have

\[
|d''(s_{i}, s_{j}) - d''(s_{i}, s_{j})| \leq \frac{96\epsilon}{\mu_{\min}^{2}}.
\]

\[\square\]

### B Example for Lumapable Markov Chain

The full Markov chain has the following transition matrix,

\[
P_{ex} = \begin{bmatrix}
0 & 84 & 463 & 101 & 463 & 84 & 463 & 101 & 463 & 84 & 101 & 84 \\
323 & 0 & 2000 & 1250 & 625 & 1000 & 625 & 1000 & 1250 & 625 & 1250 & 625 \\
463 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
147 & 103 & 147 & 0 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
463 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
147 & 103 & 147 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
323 & 273 & 323 & 323 & 0 & 2000 & 250 & 250 & 250 & 250 & 250 & 250 \\
2000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
463 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
147 & 103 & 147 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
323 & 273 & 323 & 323 & 17 & 273 & 17 & 273 & 17 & 273 & 17 & 273 & 17 \\
2000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
463 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
147 & 103 & 147 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 \\
323 & 273 & 323 & 323 & 17 & 273 & 17 & 273 & 17 & 273 & 17 & 273 & 17 \\
2000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000 & 1000
\end{bmatrix}.
\]

(21)
The transition matrix on the three meta states is,

\[ P_r = \begin{bmatrix}
0 & 0.5880 & 0.4120 \\
0.3233 & 0.1389 & 0.5378 \\
0.2720 & 0.6461 & 0.0819
\end{bmatrix}. \]

The stationary distribution of \( P_{\text{ex}} \) is

\[ \mu_{\text{ex}} = \{0.105, 0.0874, 0.105, 0.0577, 0.105, 0.0874, 0.105, 0.0577, 0.0874, 0.0577, 0.0874\}, \]

and the merging conductance is roughly 0.06. Using \( 10^4 \) sample state transitions, we obtain a fairly close estimate of stationary distribution given by

\[ \tilde{\mu}_{\text{ex}} = \{0.106, 0.088, 0.107, 0.057, 0.102, 0.087, 0.105, 0.059, 0.057, 0.087, 0.058, 0.087\}. \]

## C Use Geometric Median to Boost Probability of Success

In this section we show that the geometric median indeed boosts the probability of success. In particular, let \( V \in \mathbb{R}^{2m \times r} \) be a fixed matrix with orthonormal columns. Let \( k = O(\log \frac{1}{\delta}) \) and \( V_1, V_2, \ldots, V_k \) be independent random column-orthonormal matrices from \( \mathbb{R}^{2m \times r} \) with

\[ \forall i \in [k] : \mathbb{P} \left[ \left\| V_i^T \mathcal{V} \right\|_F^2 < r - \epsilon \right] \leq \frac{1}{4}. \]

We denote \( H \) be the geometric median of \( V_1 V_1^T, V_2 V_2^T, \ldots, V_k V_k^T \) in Euclidean space, i.e.

\[ H := \arg\min_{Z \in \mathbb{R}^{2m \times 2m}} \sum_{j=1}^{k} \left\| V_j V_j^T - Z \right\|_F. \]

Let

\[ \bar{V} = \arg\min_{V_i \in [k]} \left\| V_i V_i^T - H \right\|_F. \]

The following proposition guarantees \( \bar{V} \) to be a good estimation of \( \mathcal{V} \).

**Proposition 12.**

\[ \mathbb{P} \left[ \left\| \bar{V}^T \mathcal{V} \right\|_F^2 < r - C\epsilon \right] \leq \delta, \]

for some absolute constant \( C \).

**Proof.** We first show that the quantity \( \left\| V_i^T \mathcal{V} \right\|_F^2 \) is related to Euclidean distance in \( \mathbb{R}^{2m \times 2m} \):

\[ \left\| V_i^T \mathcal{V} \right\|_F^2 = \text{tr}(V_i^T \mathcal{V} V_i^T) = \text{tr}(V_i V_i^T \mathcal{V} V_i^T) = r - \frac{\left\| V_i V_i^T - \mathcal{V} \right\|_F^2}{2}. \]

Thus we obtain,

\[ \forall i \in [k] : \mathbb{P} \left[ \left\| V_i V_i^T - \mathcal{V} \right\|_F^2 \geq 2\epsilon \right] \leq \frac{1}{4}. \]
Then by Theorem 3.1 of Minsker et al. (2015),

\[ \mathbb{P}\left[ \left\| H - \tilde{V} \tilde{V}^\top - \tilde{V} \tilde{V}^\top \right\|_F^2 \geq C_1 \epsilon \right] \leq \frac{\delta}{2} \]

for some absolute constant \( C_1 \) and appropriately chosen \( k \). Since the \( V_i \)'s are independent, we obtain

\[ \mathbb{P}\left[ \forall i \in [k]: \left\| V_i V_i^\top - \tilde{V} \tilde{V}^\top \right\|_F^2 \geq 2 \epsilon \right] \leq \frac{1}{4k} \leq \frac{\delta}{2}, \]

for appropriately chosen \( k \). Let \( \mathcal{E} \) be the following event,

\[ \left\| H - \tilde{V} \tilde{V}^\top \right\|_F^2 < C_1 \epsilon \quad \text{and} \quad \exists i^* : \left\| V_{i^*} V_{i^*}^\top - \tilde{V} \tilde{V}^\top \right\|_F^2 < 2 \epsilon. \]

By union bound, \( \mathbb{P}[\mathcal{E}] \geq 1 - \delta \). For the rest of the proof, we condition on \( \mathcal{E} \). Thus

\[
\left\| \tilde{V} \tilde{V}^\top - \tilde{V} \tilde{V}^\top \right\|_F = \left\| \tilde{V} \tilde{V}^\top - H + H - \tilde{V} \tilde{V}^\top \right\|_F \\
\leq \left\| \tilde{V} \tilde{V}^\top - H \right\|_F + \left\| \tilde{V} \tilde{V}^\top - H \right\|_F \\
\leq \left\| V_{i^*} V_{i^*}^\top - H \right\|_F + \sqrt{C_1 \epsilon} \\
\leq \sqrt{C_1 \epsilon} + \left\| V_{i^*} V_{i^*}^\top - \tilde{V} \tilde{V}^\top \right\|_F + \left\| \tilde{V} \tilde{V}^\top - H \right\|_F \\
\leq \left( 2 \sqrt{C_1} + \sqrt{2} \right) \sqrt{\epsilon}.
\]

Thus

\[
\left\| \tilde{V} \tilde{V}^\top - \tilde{V} \tilde{V}^\top \right\|_F^2 \leq \left( 2 \sqrt{C_1} + \sqrt{2} \right)^2 \epsilon \Rightarrow \left\| V_{i^*} V_{i^*}^\top \right\|_F^2 \geq r - \frac{\left( 2 \sqrt{C_1} + \sqrt{2} \right)^2}{2} \epsilon.
\]

The proof is completed by setting \( C = \frac{(2 \sqrt{C_1} + \sqrt{2})^2}{2} \).

Moreover, \( H \) can be obtained efficiently by running the algorithm presented in Cohen et al. (2016).