State Aggregation Learning from Markov Transition Data

Yaqi Duan
Princeton University
yaqid@princeton.edu

Zheng Tracy Ke
Harvard University
zke@fas.harvard.edu

Mengdi Wang
Princeton University
mengdiw@princeton.edu

Abstract

State aggregation is a model reduction method rooted in control theory and reinforcement learning. It reduces the complexity of engineering systems by mapping the system’s states into a small number of meta-states. In this paper, we study the unsupervised estimation of unknown state aggregation structures based on Markov trajectories. We formulate the state aggregation of Markov processes into a nonnegative factorization model, where left and right factor matrices correspond to aggregation and disaggregation distributions respectively. By leveraging techniques developed in the context of topic modeling, we propose an efficient polynomial-time algorithm for computing the estimated state aggregation model. Under some “anchor state” assumption, we show that one can reliably recover the state aggregation structure from sample transitions with high probability. Sharp divergence error bounds are proved for the estimated aggregation and disaggregation distributions, and experiments with Manhattan traffic data are provided.

1 Introduction

State aggregation is a long-existing approach for model reduction of complicated systems. It is widely used as a heuristic to reduce the complexity of control systems and reinforcement learning. The earliest idea of state aggregation is to aggregate “similar” states into a small number of subsets through a partition map. However, the meta-states and the partition map are often handpicked by practitioners based on domain-specific knowledge [20, 4]. Alternatively, the partition can be chosen via discretization in accordance with some priorly known similarity metric or feature functions [22]. More generally speaking, state aggregation also allows one to map states into meta-states through some probabilistic relation. Such models have been used for modeling large Markov decision processes, where each observed state is represented as a mixture over latent states [21, 4]. This model is known as soft state aggregation, where states and meta-states are mapped to each other via aggregation distributions and disaggregation distributions. Experiences and theoretical evidences suggest that the knowledge of a state aggregation model significantly reduces the complexity of solving the Markov decision problem [21].

Unfortunately, existing approaches that use the idea of state aggregation for optimal control and reinforcement learning typically require prior knowledges about the exact system dynamics and/or the state aggregation structures. Neither is available in emerging reinforcement learning applications that involve complicated black-box systems, like autonomous driving, computer games and cell evolution. In these cases, little prior structure is known about the underlying process, while a substantial amount of transition data are available. There lacks a principled approach that infers the state aggregation structure of a black-box system from unsupervised data.

In this paper, we study the estimation of a state aggregation model from sample transition data. We hope to develop an unsupervised state aggregation approach that requires minimal prior knowledge. Suppose that we observe a Markov trajectory \( \{X_0, X_1, \ldots, X_n\} \) generated by a state-transition system. In the case of hard state aggregation, the goal would be to find a partition mapping \( E \) such
that $\mathbb{P}(X_{t+1} \mid X_t) = \mathbb{P}(X_{t+1} \mid E(X_t))$. More generally, we consider the soft aggregation structure such that every state can be represented using a membership model over a number of meta-states. To this end, we hope to develop an efficient statistical procedure to estimate and compute the state aggregation model based on trajectories of the Markov chain. The soft state aggregation structure of a Markov chain can be formulated into a nonnegative factorization of the transition probability operator. Experience suggests that spectral decomposition is effective for analysis and state compression of Markov trajectories [23]. However, plain spectral decomposition does not recover the latent states, aggregation or disaggregation distributions. It remains open how to identify the state aggregation model effectively from data.

Inspired by a technique used in topic modeling [11], we develop an efficient method to solve this open problem. We introduce a notion of “anchor state”, which is an analog of anchor words, to uniquely identify meta-states of the Markov chain. Based on existence of anchor states, our method takes as input sample state transitions and performs a multi-step SVD approach, where the key is to properly normalize the data and rotate the singular vectors into the nonnegative cone by a simplex-hunting algorithm. The proposed method outputs the estimated state aggregation model, including the meta-states, aggregation/disaggregation distributions, and sets of anchor states that represent each meta-state. Numerical experiments are performed on simulated random walks and a Manhattan taxi-trip data set, in which we identified leading modes of the taxi traffic as well as anchor regions in the city. On the theoretical side, we prove high-probability error bounds for the total variation divergence between the estimated distributions and the ground truth. To the authors’ best knowledge, this is the first statistically-proven approach that uncovers unknown state aggregation structure from time series data.

2 Related Literatures

State aggregation is one of the earliest model reduction methods that were used to approximate known control systems with smaller ones. It also applies to reinforcement learning where the underlying transition probabilities are not known a priori. State aggregation reduces the complexity of the state space and therefore reduces computational costs for approximating the optimal value function or policy; see e.g., [16, 4, 21, 22, 19]. Similar to but more general than the state aggregation approach, a line of works, known as representation learning, construct basis functions for representing high-dimensional value functions, for which many methods have been developed; e.g. [10, 14, 17, 18]. The aforementioned methods typically require prior knowledge about structures of the problem, lacking statistical guarantees.

Two recent papers studied the statistical model reduction of Markov time series. The paper [23] proposed a class of spectral state compression methods of Markov process based on singular value thresholding of empirical transition matrices. It gives statistical error bounds for estimating the transition matrix and recovering the principal subspace associated with the transition matrix. Another related paper [13] studied the estimation of Markov transition matrix using maximum likelihood estimation with a low-rank constraint. Both [23] and [13] focus on estimation of a low-rank Markov model. However, their results do not apply to recovery of state aggregation model which corresponds to a particular nonnegative factorization model. In particular, the spectral methods cannot recover the aggregation and disaggregation distributions.

Estimating the state aggregation structure is essentially a problem of nonnegative matrix factorization (NMF) [12] under noise corruption. NMF is widely applied and many methods exist (see [6] for a comprehensive overview). However, many classical methods for NMF only perform well with independent and homoscedastic noise. The empirical transition matrix generated by a Markov chain does not satisfy these conditions. Fortunately, we discover that state aggregation is closely related to topic modeling. They share similar nonnegative factorization structure and similar noise structures. In particular, empirical transition distribution conditioned at a state is analogous to the bag-of-words representation of a “document”, each outgoing state can be viewed as a “dictionary word”, and a meta-state is analogous to a “topic”. This inspires us to borrow ideas from topic modeling to develop tractable methods for estimating state aggregation.

In particular, we find that the state aggregation model is strongly related to topic models with an “anchor word” assumption [2, 11]. The anchor word assumption has a natural analog in state aggregation models, which we call anchor state assumption. An anchor state is visited by only one
meta-state with a nonzero probability. The “anchor state” assumption is meaningful in applications: In Manhattan taxi-trip data, we have successfully identified a number of “anchor states”, each representing a location in the city that reflects the leading traffic modes hidden in the taxi-trip data.

The proposed method is adapted from the topic modeling method in [11]. It is able to identify the aggregation/disaggregation distributions by leveraging the technique in [11], where, in the context of a topic model with anchor words, they developed a principled way to find vertices corresponding to the singular vectors. Our method is customized to analyzing Markov transition data for the state aggregation problem. It uses a normalization technique, which is called Jin’s SCORE [8] and proved very useful in network data and text data analysis [9, 11]. The normalization scheme is particularly crafted to address the dependence issue in Markov data. Our method can compute estimates of both aggregation and disaggregation distributions, while the topic modeling method in [11] can recover only the right factors. To the best knowledge of the authors, this is the first work that develops a tractable method for estimating the meta-states, aggregation and disaggregation distributions with statistical guarantees.

3 The State Aggregation Model

In this section, we introduce the model of soft state aggregation and the notion of anchor states. We also discuss its relations with topic model and other latent-variable models for Markov chains.

Assumption 1 (Soft State Aggregation). Let \( r \) be a positive integer. The Markov chain \( X_0, X_1, \ldots, X_n \) admits a soft state aggregation, i.e., there exist random variables \( Z_t \in [r] \) such that

\[
P(X_{t+1} | X_t) = \sum_{k=1}^{r} P(Z_t = k | X_t) P(X_{t+1} | Z_t = k),
\]

for all \( t \) with probability 1, where \( P(Z_t | X_t) \) and \( P(X_{t+1} | Z_t) \) are independent of \( t \) and referred to as the aggregation distributions and disaggregation distributions, respectively.

The soft state aggregation model has been discussed in various literatures like [21, 23]. See Section 6.3.7 for a textbook review. Admitting a soft state aggregation is equivalent to a nonnegative factorization of the transition matrix. A Markov chain satisfy the soft state aggregation with \( r \) meta-states if and only if its transition matrix \( P \in \mathbb{R}^{p \times p} \) can be written in the form of

\[
P = UV^T,
\]

where \( U \in \mathbb{R}^{p \times r}, V \in \mathbb{R}^{p \times r} \) satisfy that \( U1 = 1 \) and \( V^T1 = 1 \). This decomposition means that one can map the states into meta-states while preserving the system’s dynamics. It also suggests that the transition dynamics is the sum of \( r \) leading modes, each mode corresponding to a rank-1 matrix. The smallest integer \( r \) such that the nonnegative factorization (1) holds is called the nonnegative rank of \( P \). Throughout the paper, we assume that the nonnegative rank of \( P \), which equals to the smallest number \( r \) such that Assumption 1 holds, is a known constant and does not scale with the number of states \( p \). Note that the nonnegative factorization (1) is a stronger model than spectral decomposition. As a result, the spectral decomposition method proposed in [23] does not apply to estimation of the state aggregation model.

We hope to estimate the state aggregation structure suggested in Assumption 1 and identify the meta-states \( \{Z_t\} \). However, the latent process \( \{Z_t\} \) is not necessarily identifiable. In fact, there exist infinitely many ways to specify meta-states and aggregation distributions to represent the same Markov process. In order to find a meaningful state aggregation structure, we require that there exist at least one “anchor state” that specifies each meta-state.

Assumption 2 (Anchor State Condition). For every meta-state \( k \in [r] \), there exists an anchor state \( x \in [p] \) and a nonnegative decomposition \( P = UV^T \) such that \( U1 = 1, V^T1 = 1, V_k(x) > 0 \) and \( V_s(x) = 0 \) for all \( s \neq k \), where \( V_1, \ldots, V_r \) denote columns of \( V \).

The anchor state condition is a natural analog of the anchor word condition for topic modeling [2] and the separability condition for nonnegative matrix factorization [5]. Under additional technical conditions, the anchor state condition would ensure unique identifiability of the meta-states and the nonnegative matrix factorization [5]. Notably, the state aggregation model with anchor state condition is related to several models, which we detail below:

3
1. **Relation to Topic Model.** The topic model is a statistical model for the generating process of text documents. Given a dictionary of $p$ distinct words, there exist $r$ unknown topics, each corresponding to a probability mass function over the dictionary, denoted by $A_k \in \mathbb{R}^p$ where $k = 1, \ldots, r$. Each document $i \in [n]$ admits a mixed membership $w_i \in \mathbb{R}^r$ over the $r$ topics, and words in document $i$ are iid drawn from the dictionary using the probability mass function $\sum_{k=1}^r w_i(k)A_k$. Let $D \in \mathbb{R}^{n \times p}$ be the matrix, each row of which contains the expected word distribution in a document. Then, $D$ admits a nonnegative factorization:

$$D = WA^\top,$$

where $A = [A_1, \ldots, A_r]$ is called the topic matrix and $W = [w_1, \ldots, w_n]^\top$ is called the weight matrix. The state aggregation model has a natural analog to topic models: the empirical transitions conditioned on each state can be thought as a document, and $U$ and $V$ are analogous to the weight matrix and topic matrix, respectively. The state aggregation model differs from topic models in the sense that both documents and words correspond to the same set of states and the factor matrices $U, V$ may admit special structures particular to the Markov system.

2. **Relation to Hard Aggregation.** A Markov chain admits a hard aggregation if there exists a partition of the state space $[p] = S_1 \cup \cdots \cup S_r$ such that $\mathbb{P}(X_{t+1} \mid X_t \in S_k) = \mathbb{P}(X_{t+1} \mid X_t = x)$ for all $x \in S_k$. This suggests that the transition matrix $P$ has a nonnegative decomposition $P = UV^\top$ where $U$ is a blockwise-constant matrix and each column of it is an indicator function of one of the subsets (Prop. 3 of [23]). Admitting a hard state aggregation model does not necessarily imply the existence of anchor state for every meta-state.

3. **Relation to Lumpable Markov Models.** A Markov process is called lumpable if the state space can be partitioned into blocks while still preserving the strong Markov property. However, a lumpable Markov chain is not necessarily low-rank (Prop. 4 of [23]), as there may be local dynamics within the blocks leading to a full-rank transition matrix. A particularly interesting case is when a Markov chain is both aggregatable and lumpable, which happens if and only if

$$P = ZPD^{-1}Z^T,$$

where columns of $Z \in \mathbb{R}^{p \times r}$ are indicator functions of the blocks, $\tilde{P} \in \mathbb{R}^{r \times r}$ is a stochastic matrix governing the block-to-block transitions, and $D \in \mathbb{R}^{r \times r}$ is a diagonal matrix with block sizes along its diagonal. Lumpability implies some form of symmetry between left and right singular spaces of $P$. When an aggregatable Markov chain is also lumpable, the transition matrix can be decomposed into the form $P = UV^\top$ by letting $U = Z\tilde{P}$ and $V = D^{-1}Z$. In this case, we can verify that the anchor state condition holds, therefore it is a special case that falls under our assumptions.

### 4 State Aggregation Learning

We develop a spectral state aggregation method for analyzing Markov transition data, by adapting a technique in [11] which was developed for topic modeling. Suppose that we are given sample pairs of state transitions $\{(X_1, X'_1), \ldots, (X_n, X'_n)\}$, where $\mathbb{P}(X'_i \mid X_i)$ follows the transition kernel $P$. We formulate a matrix of empirical state-transition counts $N \in \mathbb{R}^{p \times p}$, where $N_{ij} = \sum_{t=1}^n 1_{X_t=i,X'_t=j}$. We will use the count matrix $N$ as the input of our estimation procedure, which is detailedly described in Algorithm 1.

In what follows we explain the intuition of Algorithm 1 and assume for simplicity that $\{X_1, \ldots, X_n\}$ are independently generated from some distribution, so that rows of $N$ become independent. Given the count matrix $N$, the first step is to conduct a column-wise normalization of $N$:

$$N \rightarrow \tilde{N} = N[\text{diag}(N^\top 1)]^{-1/2}. \quad (2)$$

This normalization helps reduce noise heteroscedasticity and dependence. Roughly speaking, each row of $N$ has a multinomial distribution, as a result,

$$\mathbb{E}[(N - EN)^\top (N - EN)] \approx \text{diag}(EN^\top 1).$$

Hence, the normalization step (2) makes the “covariance” matrix of $N$ approximately an identity matrix.
The state aggregation model implies
\[ \tilde{N} \approx [\text{diag}(N1)]P[\text{diag}(N^T 1)]^{-1/2} = U^*(V^*)^T, \]
where \( U^* = [\text{diag}(N1)]U, \ V^* = [\text{diag}(N^T 1)]^{-1/2}V. \) It inspires the following steps of estimating \( V: \)
1. Estimate the column space of \( V^* \) by the leading right singular vectors of \( \tilde{N}. \)
2. Recover \( V^* \) from its column space using the non-negative constraint.
3. Estimate \( V \) from the estimate of \( V^*. \)

The first and third steps are trivial, so what remains is how to recover \( V^* \) from its column space.

Borrowing the idea of [11], we apply a normalization scheme, known as SCORE [8], on singular vectors of \( \tilde{N}. \) Let \( \hat{h}_1, \ldots, \hat{h}_r \) be the right singular vectors of \( \tilde{N}. \) The SCORE normalization divides each of \( \hat{h}_2, \ldots, \hat{h}_r \) by the leading singular vector \( \hat{h}_1 \) in an entry-wise fashion and obtains a \( p \times (r-1) \) matrix
\[ \hat{D} = [\text{diag}(\hat{h}_1)]^{-1}[\hat{h}_2, \ldots, \hat{h}_r]. \] (3)

Under mild conditions, \( \hat{h}_1 \) is a positive vector so \( \hat{D} \) is well-defined. It turns out that rows of \( \hat{D} \) exhibits a simplex geometry:

- There exists a low-dimensional simplex with \( r \) vertices such that all rows of \( \hat{D} \) are approximately contained in this simplex.
- When each meta-state has an anchor state, each vertex of the above simplex has a row of \( \hat{D} \) located around it. Therefore, we can use the point cloud of rows of \( \hat{D} \) to estimate these vertices.
- With knowledge of simplex vertices, one can directly transform the singular vectors into the nonnegative cone and get an estimate of \( V^*. \)

Algorithm [11] implemented these ideas and customize them to work for the state aggregation learning problem. We refer the readers to the appendix for a detailed explanation of the proposed method.

In Step 5, the estimated \( \hat{U} \) is not guaranteed to be a nonnegative matrix. There is an alternative way to Step 5: For each \( j \in [p], \) we project the \( j \)-th row of \( \hat{P} \) to the convex hull of columns of \( \hat{V}, \) where the projected vector is uniquely expressed as a convex combination of columns of \( \hat{V} \) with a combination coefficient vector denoted as \( \hat{u}_j. \) We use \( \hat{u}_j \) to estimate the \( j \)-th row of \( U. \) Numerical experiments suggest that both subroutines for estimating \( U \) work equally well. When \( r, L \) take known constant values, the overall method has a runtime polynomial in \( p. \)

## 5 Main Statistical Results

In this section we establish the main results on statistical upper bounds for estimating the aggregation and disaggregation matrices \( U, V. \)

First, we present a statistical error bound for recovering the disaggregation distributions (columns of \( V): \)

**Theorem 1 (Statistical error bound for \( V)).** Consider a \( p \)-state Markov chain with transition matrix \( P \) that satisfies Assumptions 1 and 2 with a fixed \( r. \) Let sample transitions \( \{(X_1, X_1'), \ldots, (X_n, X_n')\} \) be generated independently from the stationary distribution \( \pi. \) Fix positive constants \( \gamma_1-\gamma_4. \) Suppose \( \frac{\pi_{\text{max}}}{\pi_{\text{min}}} \leq \gamma_1 \) and \( P^rP \) is an irreducible matrix with \( \lambda_{\text{min}}(P^rP) \geq \gamma_2. \) Suppose the first right singular vector of \( [\text{diag}(\pi)]P[\text{diag}(\pi)]^{-1/2} \) satisfies that the ratio between its maximum and minimum entries is bounded by \( \gamma_3. \) Suppose \( V \) satisfies \( \min_{j \in [p]}|V1)_j| \geq \gamma_4/p. \) When \( r \leq C \) and \( n \geq p[\log(n)]^3, \) with probability \( 1-o(n^{-1}), \) the estimate \( \hat{V} \) given by Algorithm [11] satisfies
\[ \max_{1 \leq k \leq r} \|\hat{V}_k - V_k\|_1 \leq C \left( 1 + \frac{p^3\sqrt{p}}{n\sqrt{n}} \right) \sqrt{\frac{p[\log(n)]^3}{n}}. \]

Next, we present a statistical error bound for recovering the disaggregation distributions (rows of \( U): \)
Algorithm 1 Learning the Soft State Aggregation Model.

**Input**: empirical state-transition counts $\mathbf{N}$, number of meta-states $r$, a tuning integer $L \geq r$

**Output**: the estimate $\hat{\mathbf{V}}$ and $\hat{\mathbf{U}}$

1. Use (2)-(3) to obtain the matrix $\hat{\mathbf{D}}$ from $\mathbf{N}$.
2. Vertex hunting. Apply a $k$-means clustering to rows of $\hat{\mathbf{D}}$ with the number of clusters equal to $L$. Let $c_1, \ldots, c_L$ be the resulting $k$-means cluster centers. Search among all simplexes $\mathcal{S}$ whose $r$ vertices are selected from $c_1, \ldots, c_L$ and find the simplex such that $\max_{1 \leq j \leq L} \|c_j - P_{\mathcal{S}} c_j\|^2$ is minimized, where $P_{\mathcal{S}}$ denotes the projection to $\mathcal{S}$. Let $\hat{b}_1, \ldots, \hat{b}_r$ be the vertices of this simplex.
3. For $1 \leq j \leq p$, compute
   $$\hat{w}_j = \arg \min_{q \in \mathcal{S}_{r-1}} \left\| \hat{d}_j - \sum_{k=1}^{r} q_k \hat{b}_k \right\|^2,$$
   where $\mathcal{S}_{r-1}$ is the standard $(r-1)$-simplex. Let $\hat{\mathbf{W}}$ be the $p \times r$ matrix whose $j$-th row is $\hat{w}_j^T$.
4. Renormalize each column of the matrix $[\text{diag}(\hat{\mathbf{h}}_1)][\text{diag}(\mathbf{N}^T \mathbf{1})]^{1/2} \hat{\mathbf{W}}$
   by its $\ell^1$-norm. The resulting matrix is $\hat{\mathbf{V}}$.
5. Estimate the aggregation matrix $\mathbf{U}$ by
   $$\hat{\mathbf{U}} = \hat{\mathbf{P}} \hat{\mathbf{V}} (\hat{\mathbf{V}}^T \hat{\mathbf{V}})^{-1},$$
   where $\hat{\mathbf{P}} = [\text{diag}(\mathbf{N} \mathbf{1})]^{-1} \mathbf{N}$ is the empirical transition probability matrix.

**Theorem 2** (Statistical error bounds for $\mathbf{U}$). Under the conditions of Theorem 1, if additionally $\lambda_{\text{max}}(\mathbf{V}^T \mathbf{V}) \leq C/p$, then with probability $1 - o(1)$, the estimate $\hat{\mathbf{U}}$ given by Algorithm 1 satisfies
   $$\max_{1 \leq j \leq p} \|\hat{u}_j - u_j\|_1 \leq C \left(1 + \frac{p^3 \sqrt{p}}{n \sqrt{n}}\right) \sqrt{\frac{p \log(n)}{n}}.$$

Theorems 1 and 2 provide a total-variance estimation error bound that applies uniformly to every individual aggregation distribution and every individual disaggregation distribution.

In what follows, we explain the technical conditions required by Theorems 1,2 and their theoretical implications.

**Remark** (About regularity conditions). Our theoretical bound requires a few regularity conditions, but they are rather mild:

- The requirement that $\mathbf{P}^T \mathbf{P}$ is irreducible means it cannot be converted to a block-wise diagonal matrix via simultaneous row/column permutation. The case that $\mathbf{P}^T \mathbf{P}$ has a block-wise diagonal structure rarely happens. Note that hard state aggregation only implies $\mathbf{U}$ has a block structure but not $\mathbf{P}^T \mathbf{P}$. For lumpable Markov models, as long as the latent Markov process is irreducible, this condition is also satisfied.
- The condition of $\lambda_{\text{min}}(\mathbf{P}^T \mathbf{P}) \geq \gamma_2$ is mild because $\mathbf{P}$ is a low-rank matrix and its nonzero eigenvalues are typically at the constant order.
- The condition on the first right singular vector of $[\text{diag}(\mathbf{\pi})] \mathbf{P} [\text{diag}(\mathbf{\pi})]^{-1/2}$ is a form of incoherence and is commonly used in the literature of matrix completion and spectral analysis. We note that all coordinates of the first right singular are strictly positive, as a result of the irreducibility of $\mathbf{P}^T \mathbf{P}$.
- The condition $\min_{j \in [p]} (\mathbf{V}^1)_{jj} \geq \gamma_4/p$ will be satisfied if every state can be reached sufficiently often (at least $\Omega(1/p)$ probability) in one step conditioned on some meta-state.
Informally speaking, they require some form of irreducibility and reachability as well as that aggregation/disaggregation distributions be sufficiently independent so one can separate them.

Remark (Comparison with existing bounds). Existing works on estimating low-rank Markov models relied on spectral decomposition. They only gave error bounds for either estimating $P$ or estimating the principal subspace of $V$ [23, 13]. However, [23, 13] did not account for the nonnegative aspect of the factorization, thus they cannot recover any state aggregation structure. To our best knowledge, Theorems 1 and 2 give the first direct error bounds for estimating individual aggregation and disaggregation distributions.

Remark (Optimality of the error bounds). Consider the special case where $r = 1$, so the Markov chain becomes repeatedly and independently sampling from a fixed $p$-state distribution, which equals to the single column of $V$. In this case, the leading term of the error bound given by Theorem 1 is $O(\sqrt{p/n})$. It matches the minimax total variance divergence (up to a logarithmic factor) for estimating a $p$-state discrete distribution [7]. Using a similar argument, we can see that the leading error bound $O(\sqrt{p/n})$ given by Theorem 2 is also nonimprovable for estimating aggregation distributions, because there are $p$ aggregation distributions (with a support size $r$ which is treated as a constant) and each distribution is sampled for roughly $\Omega(n/p)$ times. Therefore our error bounds are optimal in their dependence on $n$ and $p$, as long as $n$ is sufficiently large.

Remark. The current results left open a few issues. First, we assumed for simplicity that the number of meta-states $r$ is a constant value, therefore our error bounds depend only on $n$ and $p$. It remains to analyze how the recovery error scales as $r$ varies. Second, we have assumed for simplicity that $X_t$ are generated independently from the stationary distribution. It remains to analyze how would the error change if the sample transitions are drawn from a long state trajectory. In this case, one needs to analyze the mixing properties of the Markov chain. These issues are left for future research.

6 Numerical Experiments

Finally we experiment the proposed state aggregation procedure on both simulated and real data.

6.1 Simulation Results

First we test our new approach on simulated sample transitions. For a $p$-state Markov chain with $r$ meta states, we first randomly create two matrices $U, V \in \mathbb{R}_{+}^{p \times r}$ such that each meta state has the same number of anchor states. After assembling a transition matrix $P = UV^T$, we generate random walk data $X_1, \ldots, X_n$ as input of Algorithm 1.
We carry out experiments with \( p = 1000, \ r = 6 \) and the number of anchor states equal to 25, 50, 75, 100 for each meta state. The results with varying sample size \( n \) are plotted in Figure 1. We observe that, when there are more anchor states, Algorithm 1 performs slightly better at recovering columns of \( \mathbf{V} \). In the log-log plot, the error in \( \hat{\mathbf{V}} \) scales linearly with \( n \). By fitting a curve, we see that the error curve (when there are 100 anchor states) approximately satisfies
\[
\ln \left( \max_{1 \leq k \leq r} \| \hat{\mathbf{V}}_k - \mathbf{V}_k \|_1 \right) \approx 5.0 - 0.51 \ln n.
\]
This validates our error upper bound in Theorem 1.

We next compare Algorithm 1 with the spectral state compression method proposed in [23], in terms of recovering the transition matrix \( \mathbf{P} \). Let \( \hat{\mathbf{U}} \hat{\mathbf{V}}^T \) be our estimator, and let \( \tilde{\mathbf{P}} \) be obtained using the competing method (computed by singular value thresholding). We measure the recovery error by using an averaged-per-row total variation divergence. Numerical results are plotted in Figure 2. We see that our proposed method consistently outperforms the competing estimator.

### 6.2 Experiments with NYC Taxi Data

We analyze a data set of \( 1.1 \times 10^7 \) NYC yellow cab trips that were collected in January 2016 [1]. We treat each taxi trip as a sample transition generated by a cite-wide Markov process over the New York city, where the transition is from a pick-up location to some drop-off location. We discretize the map into \( p = 1922 \) cells so that the Markov process becomes a finite-state one.

We apply the proposed state aggregation method to the taxi-trip data, and we test various sizes of the meta-states, i.e., \( r = 4, 6, 8, 10 \). For each vertex produced in the algorithm, there exists a cluster of anchor states around it that represent a same meta-state. In order to connect these anchor states with a region in the city, we select 4 points that are closest to the vertex, and locate the corresponding cells on the map. The convex hull of the cells forms an “anchor region”, which is plotted in Figure 3.

Let \( \hat{\mathbf{U}}, \hat{\mathbf{V}} \) be the estimated aggregation and disaggregation matrices. We use heat maps to visualize the columns of \( \hat{\mathbf{U}}, \hat{\mathbf{V}} \). Take \( r = 6 \) for example. We pick two meta-states, with anchor states in the downtown and midtown areas, respectively. We plot in Figure 4 the corresponding columns of \( \hat{\mathbf{U}} \) and \( \hat{\mathbf{V}} \). Recall that the columns of \( \hat{\mathbf{V}} \) are the disaggregation distributions, and columns of \( \hat{\mathbf{U}} \) can be thought of as some likelihood function for transitioning to corresponding meta-states. The heat maps can be viewed as some leading “modes” in the traffic-dynamics, corresponding to anchor regions in the downtown and midtown areas, respectively.

Furthermore, we analyze the estimated aggregation and disaggregation distributions to find a partition of the Manhattan city. Recall that in Algorithm 1, each state is projected onto a simplex, which can be represented as a convex combination of simplex’s vertices. For each state, we assign the state to a cluster that corresponds to largest value in the weights of convex combination. In this way, we can cluster the 1922 locations into a small number of regions. The partition results are shown in Figure 5, where anchor regions are marked in each cluster.

### 7 Summary

This paper proposes a spectral-based method for uncovering the state aggregation structure from Markov transition data. We establish sharp statistical error bound for each aggregation/disaggregation distribution. While our current results apply only to finite-state Markov chains, we hope they provide an initial attempt towards developing more powerful tools for statistical system identification. We also hope this paper would point out an intriguing connection between state aggregation and topic modeling, which might inspire more work in this area and motivate the developments of novel statistical methods for modeling complex systems.
Figure 3: Anchor regions identified using the state aggregation method from Manhattan taxi-trip data.

Figure 4: Estimated aggregation likelihoods (columns of $\hat{U}$) and disaggregation distributions (columns of $\hat{V}$) illustrated as heatmaps (where $r = 6$).

Figure 5: Partition of New York City by rounding the estimated disaggregation distributions to the closest vertices in the low-dimensional simplex.
References

[1] NYC Taxi and Limousine Commission (TLC) trip record data. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml Accessed June 11, 2018.

[2] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models—going beyond SVD. In Foundations of Computer Science (FOCS), pages 1–10, 2012.

[3] Dimitri P Bertsekas. Dynamic programming and optimal control. Athena scientific Belmont, MA, 2007.

[4] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming. Athena Scientific, Belmont, MA, 1996.

[5] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In Advances in neural information processing systems, pages 1141–1148, 2004.

[6] Nicolas Gillis. The why and how of nonnegative matrix factorization. Regularization, Optimization, Kernels, and Support Vector Machines, 12(257), 2014.

[7] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ1 loss. IEEE Transactions on Information Theory, 61(11):6343–6354, 2015.

[8] Jiashun Jin. Fast community detection by SCORE. Annals of Statistics, 43(1):57–89, 2015.

[9] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. arXiv:1708.07852, 2017.

[10] Jeff Johns and Sridhar Mahadevan. Constructing basis functions from directed graphs for value function approximation. In Proceedings of the 24th international conference on Machine learning, pages 385–392. ACM, 2007.

[11] Zheng Tracy Ke and Minzhe Wang. A new SVD approach to optimal topic estimation. arXiv:1704.07016, 2017.

[12] Daniel Lee and Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999.

[13] Xudong Li, Mengdi Wang, and Anru Zhang. Estimation of Markov chain via rank-constrained likelihood. Proceedings of the 35th international conference on Machine learning, 2018.

[14] Sridhar Mahadevan. Proto-value functions: Developmental reinforcement learning. In Proceedings of the 22nd international conference on Machine learning, pages 553–560. ACM, 2005.

[15] Sridhar Mahadevan. Learning representation and control in Markov decision processes: New frontiers. Foundations and Trends® in Machine Learning, 1(4):403–565, 2009.

[16] Andrew W Moore. Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces. In Machine Learning Proceedings 1991, pages 333–337. Elsevier, 1991.

[17] Ronald Parr, Christopher Painter-Wakefield, Lihong Li, and Michael Littman. Analyzing feature generation for value-function approximation. In Proceedings of the 24th international conference on Machine learning, pages 737–744. ACM, 2007.

[18] Marek Petrik. An analysis of laplacian methods for value function approximation in MDPs. In IJCAI, pages 2574–2579, 2007.

[19] Zhiyuan Ren and Bruce H Krogh. State aggregation in Markov decision processes. In Proceedings of the 41st IEEE Conference on Decision and Control, volume 4, pages 3819–3824. IEEE, 2002.

[20] David F Rogers, Robert D Plante, Richard T Wong, and James R Evans. Aggregation and disaggregation techniques and methodology in optimization. Operations Research, 39(4):553–582, 1991.

[21] Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Reinforcement learning with soft state aggregation. In Advances in neural information processing systems, pages 361–368, 1995.

[22] John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. Machine Learning, 22(1-3):59–94, 1996.
[23] Anru Zhang and Mengdi Wang. Spectral state compression of Markov processes. 
\textit{arXiv:1802.02920}, 2018.
Appendices

A Rationale of Algorithm 1

In this section, we explain the rationale of Algorithm 1, especially for:

- Why the SCORE normalization [8] produces a simplex geometry.
- How the simplex geometry is used for estimating $V$.

Write for short $n = N_1$ and $m = N_1^\top$. The normalized data matrix

$$\tilde{N} \approx \text{diag}(n)P[\text{diag}(m)]^{-1/2} \equiv \tilde{N}^\ast.$$  \hspace{1cm} (4)

The matrix $\tilde{N}^\ast$ can be viewed as the “signal” part of $N$. Let $h_1, \ldots, h_r$ be the right singular vectors of $\text{diag}(n)P[\text{diag}(m)]^{-1/2}$. They can be viewed as the population counterpart of $\hat{h}_1, \ldots, \hat{h}_r$. We define a population counterpart of the matrix $\hat{D}$ produced by SCORE:

$$D = [\text{diag}(h_1)]^{-1/2} [h_2, \ldots, h_r].$$  \hspace{1cm} (5)

From now on, we pretend that the matrix $\tilde{N}^\ast$ is directly given and study the geometric structures associated with the singular vectors and the SCORE matrix $D$.

A.1 The simplex geometry and explanation of steps of Algorithm 1

When $P = UV^\top$, the matrix $\tilde{N}^\ast$, defined in (4), also admits a low-rank decomposition:

$$\tilde{N}^\ast = U^* (V^*)^\top,$$

where

$$U^* = [\text{diag}(n)]U, \quad V^* = [\text{diag}(m)]^{-1/2}V.$$  \hspace{1cm}

By linear algebra, the span of the right singular vectors $h_1, \ldots, h_r$ is the same as the column space of $V^*$. It implies there exists a linear transformation $L \in \mathbb{R}^{r \times r}$ such that

$$H \equiv [h_1, \ldots, h_r] = V^* L.$$  \hspace{1cm} (6)

Since $V^*$ is a nonnegative matrix, each row of $H$ is an affine combination of rows of $L$. Furthermore, if $j$ is an anchor state, then the $j$-th row of $V^*$ has exactly one nonzero entry, and so the $j$-th row of $H$ is proportional to one row of $L$. This gives rise to the following simplicial-cone geometry:

**Proposition 1** (Simplicial cone geometry). Suppose $P = UV^\top$, each meta state has an anchor state, and $\text{rank}(U) = r$. Let $H = [h_1, \ldots, h_r]$ contain the right singular vectors of $\tilde{N}^\ast = \text{diag}(n)P[\text{diag}(m)]^{-1/2}$. There exists a simplicial cone in $\mathbb{R}^r$, which has $r$ extreme rays, such that all rows of $H$ are contained in this simplicial cone. Furthermore, for all anchor states $j$ of a meta state, the $j$-th row of $H$ lies exactly on one extreme ray of this simplicial cone.

**Remark.** Similar simplicial cone geometry has been discovered in the literature of nonnegative matrix factorization [5]. The simplicial cone there is associated with rows of the matrix that admits a nonnegative factorization, but the simplicial cone here is associated with singular vectors of the matrix. Since SVD is a linear projection, it is not surprising that the simplicial cone structure is retained in singular vectors.

However, in the real case, we have to apply SVD to the noisy matrix, then the simplicial cone is corrupted by noise and hardly visible. We hope to find a proper normalization of $H$, so that the normalized rows are all contained in a simplex, where all points on the extreme ray of the previous simplicial cone (these points do not overlap) fall onto one vertex of the current simplex (these points now overlap). Such a simplex geometry is much more robust to noise corruption and is easier to estimate.

How to normalize $H$ to obtain a simplex geometry is tricky. If all entries of $H$ are nonnegative, we can normalize each row of $H$ by the $\ell^1$-norm of that row, and rows of the resulting matrix are...
1.1. We now prove Propositions 2-3. Recall that by (6), for \( k \), \( l \), \( V \) the next proposition shows that we can recover \( w \) an anchor state \( h \), the coordinates of \( j \) from the standard simplex such that \( d \) be the \( j \)-th row of the simplex, for \( j \in [p] \). By the nature of a simplex, each point in it can be uniquely expressed as a convex combination of the vertices. This means, for each \( j \in [p] \), there exists a weight vector \( w_j \) from the standard simplex such that

\[
d_j = \sum_{k=1}^{r} w_j(k)b_k.
\]

The next proposition shows that we can recover \( V \) from \( w_1, \ldots, w_p \).

Proposition 3 (Relation of simplex and matrix \( V \)). In the setting of Proposition 2 each row of \( D \) is a convex combination of the vertices of \( S_0^r \), i.e., for each \( j \in [p] \), there exists \( w_j \) in the standard simplex such that \( d_j = \sum_{k=1}^{r} w_j(k)b_k \). Furthermore, consider the matrix \( W \in \mathbb{R}^{p \times r} \), whose \( j \)-th row equals to \( w_j^\top \). Then, \( W \) and \( V \) are connected by

\[
[\text{diag}(h_1)][\text{diag}(m)]^{1/2}W = V[\text{diag}(l_1)],
\]

where \( l_1 \) is the first column of \( L \) as defined in (5).

By Proposition 3 each column of the matrix

\[
[\text{diag}(h_1)][\text{diag}(m)]^{1/2}W
\]

is proportional to the corresponding column of \( V \). Since each column of \( V \) has a unit \( \ell^1 \)-norm, if we normalize each column of the above matrix by its \( \ell^1 \)-norm, we can exactly recover \( V \).

Once we have obtained \( V \), we can immediately recover \( U \) from \((P, V) \) by the relation:

\[
U = U(V^\top V)(V^\top V)^{-1} = PV(V^\top V)^{-1}.
\]

The above gives the following theorem:

Proposition 4 (Exact recovery of \( U \) and \( V \)). In the setting of Proposition 2 if we apply Algorithm 1 to the matrix \( [\text{diag}(n)]P[\text{diag}(m)]^{-1/2} \), it exactly outputs \( U \) and \( V \).

A.2 Proof of propositions

Proposition 1 follows from (6) and definition of simplicial cone. Proposition 2 is proved in Section 1.1. We now prove Propositions 2-3. Recall that by (6), for \( k \in [r] \),

\[
h_k = V^*l_k,
\]

where \( l_k \) is the \( k \)-th column of \( L \). When \( (\bar{N}^*)^\top \bar{N}^* \) is an irreducible matrix, by Perron’s theorem, all the coordinates of \( h_1 \) are strictly positive. This guarantees that \( D \) is well-defined. Additionally, for an anchor state \( j \) of the \( k \)-th meta state, \( h_1(j) = V_{jk}^*l_1(k) \), where \( V_{jk}^* > 0 \). Therefore, \( l_1(k) > 0 \) for \( k \in [r] \). We define a matrix

\[
B = [\text{diag}(l_1)]^{-1}l_2, \ldots, l_r
\]
By definition,  
\[ [1, \mathbf{D}] = [\text{diag}(\mathbf{h}_1)]^{-1} \mathbf{H}, \]  
and  
\[ [1, \mathbf{B}] = [\text{diag}(\mathbf{l}_1)]^{-1} \mathbf{L}. \]  
Combining them with (6) gives  
\[ [1, \mathbf{D}] = [\text{diag}(\mathbf{h}_1)]^{-1} \mathbf{V}^*[\text{diag}(\mathbf{l}_1)][1, \mathbf{B}]. \]  
Let  
\[ \mathbf{W} = [\text{diag}(\mathbf{h}_1)]^{-1} \mathbf{V}^*[\text{diag}(\mathbf{l}_1)]. \]  
The (9) implies  
\[ 1 = \mathbf{W} \mathbf{1}, \quad \mathbf{D} = \mathbf{W} \mathbf{B}. \]  
Since \( \mathbf{W} \) is a nonnegative matrix, the first equation implies that each row of \( \mathbf{W} \) is from the standard simplex, and the second equation implies that each row of \( \mathbf{D} \) is a linear combination of the \( r \) rows of \( \mathbf{B} \), where the combination coefficients come from the corresponding row of \( \mathbf{W} \). This has proved the simplex geometry stated in Proposition 2.

Note that the \( j \)-th row of \( \mathbf{D} \) is located on one vertex of the simplex if and only if the \( j \)-th row of \( \mathbf{W} \) is located on one vertex of the standard simplex. From the way we define \( \mathbf{W} \), its \( j \)-th row equal to  
\[ w_j^T = \frac{1}{h_1(j)\sqrt{m_j}} [V_{j1}l_1(1), V_{j2}l_1(2), \ldots, V_{jr}l_1(r)]. \]  
Since \( \mathbf{h}_1, \mathbf{l}_1 \) and \( \mathbf{m} \) are all positive vectors, \( w_j \) is located on one vertex of the standard simplex if and only if exactly one of \( V_{j1}, \ldots, V_{jr} \) is nonzero, where the latter is true if and only if \( j \) is an anchor state. This has proved Proposition 3.

Furthermore, from the way \( \mathbf{W} \) is defined above, using the fact that \( \mathbf{V}^* = [\text{diag}(\mathbf{m})]^{-1/2} \mathbf{V} \), we immediately find that  
\[ \mathbf{W} = [\text{diag}(\mathbf{h}_1)]^{-1}[\text{diag}(\mathbf{m})]^{-1/2} \mathbf{V} [\text{diag}(\mathbf{l}_1)], \]  
which is equivalent to  
\[ [\text{diag}(\mathbf{h}_1)][\text{diag}(\mathbf{m})]^{1/2} \mathbf{W} = \mathbf{V} [\text{diag}(\mathbf{l}_1)]. \]  
This has proved Proposition 3.

### B Proof of Theorems 1-2

Theorem 1 is adapted from Theorem 2.2 of [11], by connecting our setting with a topic modeling setting. Let \( \mathbf{n} = \mathbf{N} \mathbf{1} \). Since the sample transitions are generated independently from the stationary distribution \( \pi \), we have

- \( \mathbb{E} n_i = n \pi_i \). Since \( \pi_{\max}/\pi_{\min} \leq \gamma_1 \), \( \mathbb{E} n_i = \Omega(n/p) \). Combining it with large-deviation inequalities for Bernoulli variables (e.g., the Bernstein inequality), with probability \( 1 - o(n^{-1}) \), \( n_i \geq c_1 n/p \) simultaneously for all \( 1 \leq i \leq p \), for some constant \( c_1 > 0 \).
- Conditioning on \( \mathbf{n} \), the rows of \( \mathbf{N} \) are independent of each other, where the \( i \)-th row has a distribution of Multinomial \((n_i, \mathbf{p}_i)\), where \( \mathbf{p}_i \) is the \( i \)-th row of \( \mathbf{P} \).

Therefore, conditioning on \( \mathbf{n} \), each row of \( \mathbf{N} \) follows the same statistical model as the bag-of-words representation of a document of length of \( n_i \), generated from the topic matrix \( \mathbf{V} \) with a weight \( \mathbf{u}_i \) over the topics. Therefore, Theorem 1 is an application of Theorem 2.2 of [11] to the case where the dictionary size is \( p \), the number of documents is \( p \), and the minimum document length is \( \Omega(n/p) \).

We notice that Theorem 2.2 of [11] assumes all documents have equal length. Here, \( n_i \)'s are different; but since they are at the same order, the proof of Theorem 2.2 of [11] also applies. Moreover, our regularity conditions look different from those of Theorem 2.2 of [11], as we have adapted them to the state aggregation setting.

We now prove Theorem 2. Fix \( 1 \leq j \leq p \). Note that  
\[ \hat{u}_j^T = \hat{p}_j^T \hat{V}(\hat{V}^T \hat{V})^{-1}, \quad u_j^T = p_j^T V(V^T V)^{-1}. \]
As a result,
\[
\|\hat{u}_j - u_j\|_1 \leq \| (\hat{p}_j - p_j)^T V (\hat{V}^T \hat{V})^{-1} \|_1 \\
+ \| p_j^T V [(\hat{V}^T \hat{V})^{-1} - (V^T V)^{-1}] \|_1 \\
+ \| p_j^T (\hat{V} - V)(\hat{V}^T \hat{V})^{-1} \|_1 \\
\equiv (I_1) + (I_2) + (I_3).
\]

We shall bound the three terms separately. Consider \((I_1)\). Let \(V_k\) be the \(k\)-th column of \(V\). We first give a bound for \(\| (\hat{p}_j - p_j)^T V_k \|\). Let \(y_1, \ldots, y_{n_j}\) be \textit{iid} drawn from a multinomial distribution with the number of trials equal to 1 and the probability vector equal to \(p_j\). Then,
\[
(\hat{p}_j - p_j)^T V_k \equiv \frac{1}{n_j} \sum_{m=1}^{n_j} V_k^T y_m.
\]

Each \(\| V_k^T y_m \|\) is bounded by \(\| V_k \|_\infty = O(1/p)\). Moreover, the variance of \(V_k^T y_m\) is bounded by \(V_k^T \text{diag}(p_j)V_k \leq C/p^2\), where we have used the fact that \(\| p_k \|_\infty \leq \max_k \| V_k \|_\infty = O(1/p)\). Using the Bernstein inequality, we know that with probability \(1 - o(1/p^2)\),
\[
(\| \hat{V} - V \|_F)^2 = (1 + \frac{p^2}{n\sqrt{n}}) \sqrt{p \log(n)}/n.\]

In Theorem 2.2 of [11], it provides a row-wise error bound on \(\hat{V}\), that is, with probability \(1 - o(n^{-1})\), for all \(j \in [p]\)
\[
\| \hat{v}_j - v_j \|_1 \leq \| v_j \|_1 \cdot err_n.
\]

As a result, we can get an upper bound on \(\| \hat{V} - V \|_F^2\):
\[
\| \hat{V} - V \|_F^2 = \sum_{j=1}^p \| \hat{v}_j - v_j \|_1^2 \leq \sum_{j=1}^p \| \hat{v}_j - v_j \|_1^2 \leq err_n^2 \sum_{j=1}^p \| v_j \|_1^2 \leq r \cdot err_n^2 \sum_{j=1}^p \| v_j \|_1^2 \leq Cerr_n^2 \| \hat{V} \|_F^2 \leq C_{errn}^2.
\]

In particular, it implies \(\| \hat{V} - V \| \leq \| \hat{V} - V \|_F \leq C_{errn}/\sqrt{p}\). We notice that
\[
\| \hat{V}^T \hat{V} - V^T V \| \\
\leq 2\| (\hat{V} - V)^T V \| + \| (\hat{V} - V)^T (\hat{V} - V) \| \\
\leq 2\| \hat{V} - V \| \| V \| \\
\leq 3\| \hat{V} - V \|_F \| V \| \\
\leq C_{errn}/p,
\]

where in the last inequality we have used \(\| \hat{V} \| \leq C / \sqrt{p}\). As a result,
\[
\| (\hat{V}^T \hat{V})^{-1} - (V^T V)^{-1} \| \\
= \| (V^T V)^{-1} \| \| (V^T \hat{V})^{-1} - (V^T V)^{-1} \| \| V^T \hat{V} - V^T V \| \\
\leq C_{errn} \cdot err_n.
\]
In particular, \( \| (\hat{\mathbf{V}}^\top \hat{\mathbf{V}})^{-1} \| \lesssim \| (\mathbf{V}^\top \mathbf{V})^{-1} \| \leq Cp \). Combining it with (11), we find that
\[
| (I_1) | \leq \| (\hat{\mathbf{p}}_j - \mathbf{p}_j)^\top \mathbf{V} \|_1 \| (\hat{\mathbf{V}}^\top \hat{\mathbf{V}})^{-1} \|_1 \\
\leq \sqrt{r} \| (\hat{\mathbf{p}}_j - \mathbf{p}_j)^\top \mathbf{V} \|_1 \| (\hat{\mathbf{V}}^\top \hat{\mathbf{V}})^{-1} \|
\leq C \sqrt{\frac{p \log(n)}{n}}. \tag{15}
\]

Consider \((I_2)\). It is seen that
\[
| (I_2) | \leq \| \mathbf{p}_j^\top \mathbf{V} \|_1 \| (\hat{\mathbf{V}}^\top \hat{\mathbf{V}})^{-1} - (\mathbf{V}^\top \mathbf{V})^{-1} \|_1 \\
\leq \sqrt{r} \| \mathbf{p}_j^\top \mathbf{V} \|_1 \| (\hat{\mathbf{V}}^\top \hat{\mathbf{V}})^{-1} - (\mathbf{V}^\top \mathbf{V})^{-1} \|
\leq Cerr_n. \tag{16}
\]

Consider \((I_3)\). It is seen that
\[
| (I_3) | \leq \| \hat{\mathbf{p}}_j^\top (\hat{\mathbf{V}} - \mathbf{V}) \|_1 \| (\hat{\mathbf{V}}^\top \hat{\mathbf{V}})^{-1} \|_1 \\
\leq \sqrt{r} \| \hat{\mathbf{p}}_j^\top (\hat{\mathbf{V}} - \mathbf{V}) \|_1 \| (\hat{\mathbf{V}}^\top \hat{\mathbf{V}})^{-1} \|
\leq Cp \cdot \| \hat{\mathbf{p}}_j^\top (\hat{\mathbf{V}} - \mathbf{V}) \|_1 \\
\leq Cp \cdot (\| \mathbf{p}_j^\top (\hat{\mathbf{V}} - \mathbf{V}) \|_1 + \| (\hat{\mathbf{p}}_j - \mathbf{p}_j)^\top \mathbf{V} \|_1
+ \| (\hat{\mathbf{p}}_j - \mathbf{p}_j)^\top (\hat{\mathbf{V}} - \mathbf{V}) \|_1)
\]

The third two term is dominated by the first two terms. We already have a bound for \( \| (\hat{\mathbf{p}}_j - \mathbf{p}_j)^\top \mathbf{V} \|_1 \). Additionally, by (12),
\[
\| \mathbf{p}_j^\top (\hat{\mathbf{V}} - \mathbf{V}) \|_1 \leq \sum_{i=1}^{p} \mathbf{p}_j(i) \| \hat{\mathbf{v}}_i - \mathbf{v}_i \|_1
\leq err_n \sum_{i=1}^{p} \mathbf{p}_j(i) \| \mathbf{v}_i \|_1
\leq Cerr_n \max_i \| \mathbf{v}_i \|_1 \\
\leq Cp^{-1} err_n. \tag{17}
\]

Combining (11) and (17) gives
\[
| (I_3) | \leq C \left( err_n + \sqrt{\frac{p \log(n)}{n}} \right) \leq Cerr_n. \tag{18}
\]

The claim follows from (15), (16) and (18).