Pattern Recognition Letters
journal homepage: www.elsevier.com

On the choice of Number of Superstates in the Aggregation of Markov Chains

Amber Srivastava, Raj K. Velicheti, Srinivasa M. Salapaka**

[a]*Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, IL, 61801, USA*

ABSTRACT

Many studies involving large Markov chains require determining a smaller representative (aggregated) chain. Each *superstate* in the representative chain represents a *group of related* states in the original Markov chain. Typically, the choice of number of superstates in the aggregated chain is ambiguous, and based on the limited prior know-how. In this paper we present a structured methodology of determining the best candidate for the number of superstates. We achieve this by comparing aggregated chains of different sizes. To facilitate this comparison we first develop a new quantity called *heterogeneity* of a superstate, and subsequently use it to establish a notion of *marginal return* of an aggregated chain. In particular, our notion of marginal return captures the decrease in the *heterogeneity* upon a unit increase in the number of superstates in the aggregated chain. We use Maximum Entropy Principle to justify the notion of marginal return, as well as our quantification of heterogeneity. Through simulations on synthetic Markov chains, where the number of superstates are known apriori, we show that the aggregated chain with the largest marginal return identifies this number. In case of Markov chains that model real-life scenarios we show that the aggregated model with the largest marginal return identifies an inherent structure unique to the scenario being modelled; thus, substantiating the efficacy of our proposed methodology.

1. Introduction

Markov chains provide a mathematical model to study many real-world stochastic processes, such as population dynamics, cruise control, transportation systems, and queueing networks (Gagniuc, 2017). However, Markov chain models for several complex systems such as applications originating in network analysis (Srikant, 2004), neuroscience (Quinn et al., 2011), and economics (Zhang and Yin, 2004) require large number of states where analyzing them is challenging and inefficient; thus, laying foundation for Markov chain aggregation techniques.

The two-fold objective of the aggregation problem is to (a) group *similar* states in the Markov chain and represent them as a single *superstate* in the aggregated model, and (b) determine the state transition probability matrix of the aggregated model. Prior works in literature (Aldhaheri and Khalil, 1991; Rached et al., 2004; Beck et al., 2009; Vidyasagar, 2012; Xu et al., 2013; Deng et al., 2011) aim at developing an appropriate *dissimilarity* metric to compare the differently-sized state transition probability matrices of the Markov chain and its aggregated model. The algorithms proposed in (Xu et al., 2014; Geiger, 2014; Deng et al., 2011) determine the aggregated models at a *pre-specified* number of superstates such that the dissimilarity is minimized. However, there is scant literature on determining the appropriate number of superstates in the aggregated chain. One of the recent works (Sledge and Pri'ncipe, 2019) provides an aggregation algorithm for the class of nearly completely decomposable (NCD) (Ando and Fisher, 1963) Markov chains, that also estimates the number of superstates underlying these particular NCD chains.

In this work we provide a methodology to estimate the number of superstates underlying any general Markov chain. Our proposed approach is independent of the algorithm being used to obtain the aggregated representations of the given Markov chain. We devise a method to quantitatively compare the aggregated models of different sizes that describe the same Markov chain. We demonstrate its utility in determining an appropriate choice for the number of superstates among the *given* aggregated models. More precisely, we develop and quantify a notion of *marginal return* that compares the aggregated mod-

**Corresponding author
e-mail: salapaka@illinois.edu (Srinivasa M. Salapaka)

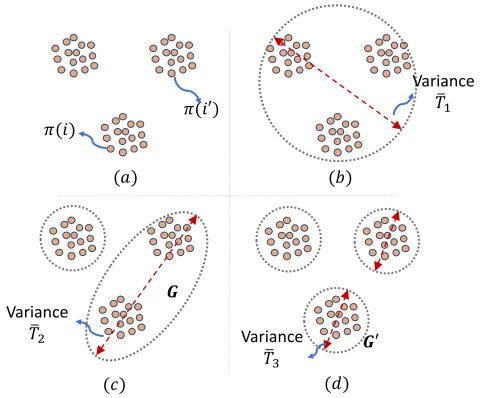

Fig. 1: (a) Represents the states in the original Markov chain. (b) Aggregated model with 1 superstates, i.e. all states in the same group. (c) Illustrates the 2 superstates, i.e., 2 groups of related states. (c) Illustrates the 3 distinct groups of related states.

els of different sizes. We demonstrate via simulations that the aggregated model with the largest marginal return estimates the superstates *underlying* the original Markov chain. The corresponding size of this aggregated model provides an appropriate choice for the number of superstates.

The primary goal in the aggregation of Markov chain is akin to the data-clustering problem. The former identifies groups of *similar* states, whereas the latter determines the groups (or clusters) of *similar* data points. Thus, one can view the problem of determining the appropriate number of superstates to be analogous to estimating the *true* (or natural) number of clusters in a dataset. The latter is a well-studied problem in the data clustering literature (Tibshirani et al., 2001; Feng and Hamerly, 2007; Pelleg et al., 2000; Sugar and James, 2003; Hamerly and Elkan, 2004; Kalogeratos and Likas, 2012; Srivastava et al., 2019). However, the methods proposed in the context of data clustering are not directly applicable to the aggregation of Markov chains. This is due to the inherent differences in the two problems. The data clustering problem identifies groups (or clusters) of similar data points lying in the euclidean space where the cost function is typically 2-norm. On the other hand, the Markov chain aggregation requires determining the groups of similar states whose transition probabilities lie in the probability space, and the cost is typically in terms of the relative entropy of the transition probabilities. Nonetheless, the methods proposed in the context of data clustering can still be helpful in devising *procedures* to compare aggregated chains of different sizes. These procedures can then be used to estimate the appropriate number of superstates. In fact, our paper builds on the prior work done in (Srivastava et al., 2019) that develops and quantifies a notion of *persistence* of a clustering solution, and utilizes it to estimate the natural number of clusters in the dataset. Below we provide a qualitative and quantitative descriptions of our notions of marginal return and heterogeneity.

Qualitative description of marginal return: As stated above we exploit the work done in (Srivastava et al., 2019) to qualitatively elucidate our notion of marginal return and the heterogeneity of a superstate. We first give a brief description of the idea presented in (Srivastava et al., 2019). Thereafter, we abstract it to the aggregation of Markov chains. Consider the dataset $\mathscr{X} = \{x_i\}$ illustrated in the Figure 1(a). It is apparent from the figure that the dataset comprises of 3 groups of similar datapoints, i.e., 3 *natural* clusters. Figures 1(b), 1(c), and 1(d) illustrate the corresponding clustering solutions obtained with 1, 2, and 3 number of clusters, respectively. $\bar{T}_1$, $\bar{T}_2$, and $\bar{T}_3$ denote the *variance* (or, spread of data points) within the largest cluster in the respective clustering solutions.

Note that there is a significant reduction in the variance of the largest cluster from $\bar{T}_2$ in Figure 1(c) (with 2 clusters) to $\bar{T}_3$ in Figure 1(d) (with 3 clusters). Contrastingly, the variance $\bar{T}_1$ in Figure 1(b) (with 1 cluster) is comparable to the variance $\bar{T}_2$ in Figure 1(c) (with 2 clusters). The significant reduction in the variance from $\bar{T}_2$ to $\bar{T}_3$ coincides with the progression from the clustering solution with 2 clusters (in Figure 1(c)) to the clustering solution with 3 clusters (in Figure 1(d)) - where 3 is the *natural* number of clusters in Figure 1(a). The work done in (Srivastava et al., 2019) exploits the above observation. It first quantifies variance in terms of the largest eigenvalue of the *cluster covariance matrix* (Rose, 1998). Subsequently, it proposes that the clustering solution with *true* number of clusters exhibits *a significant drop in the variance of the largest cluster* when compared to the clustering solution with one less cluster. (Srivastava et al., 2019) refers this to as the clustering solution with true number of clusters being more *persistent* than other clustering solutions. The simulations in (Srivastava et al., 2019) demonstrate that the above idea significantly outperforms popular benchmark methods (Feng and Hamerly, 2007; Pelleg et al., 2000; Sugar and James, 2003; Hamerly and Elkan, 2004; Kalogeratos and Likas, 2012) in literature when applied to diverse synthetic and standard datasets.

We now extend the above idea to the aggregation of Markov chains. Analogous to the *variance* of a cluster in (Srivastava et al., 2019) we establish the *heterogeneity* of a superstate in the context of Markov chains; wherein, heterogeneity quantifies the dissimilarity within the group of related states (i.e., states represented by the same superstate). Subsequently, we define the *marginal return* of an aggregated model as the *reduction in the heterogeneity* of the largest superstate in comparison to the aggregated model with one less superstate. We illustrate the idea further as follows. Consider a set of K aggregated models each with different number $k \in \{1,\ldots,K\}$ of superstates. Analogous to the above case of data clustering we propose that *the aggregated model with the largest marginal return estimates the true number k_t of superstates*. The idea is that there is a significant drop in the heterogeneity of the largest superstate in the aggregated models $\bar{M}_{k_t-1}$ with k_t-1 superstate in comparison to the aggregated model $\bar{M}_{k_t}$ with k_t superstates. This could be owing to several reasons. For instance, the largest superstate in $\bar{M}_{k_t-1}$ combines *distinct* groups of *similar* states in the original Markov chain resulting into much larger heterogeneity, whereas the aggregated model $\bar{M}_{k_t}$ possibly identifies each group of *similar* states distinctly thus exhibiting much lesser heterogeneity. As in the case of data-clustering, the simulations on multiple synthetic and real-world Markov chains demonstrate the efficacy of the above proposed methodology.

Quantifying heterogeneity and marginal return: The quantification of persistence of clustering solutions and the subsequent characterization of natural number of clusters in (Sri-

vastava et al., 2019) is motivated from the Deterministic Annealing (DA) algorithm presented in (Rose, 1998). Deterministic Annealing (DA) is a Maximum Entropy Principle (Jaynes, 2003) based clustering algorithm. It operates by determining a single cluster at large values of annealing parameter T. As the annealing proceeds (i.e., T decreases) the number of distinct clusters increases at specific instances referred to as *phase transitions* (Rose, 1998). The notion of persistence and the subsequent characterization of natural clusters in (Srivastava et al., 2019) are motivated by the analytical conditions on the annealing parameter T that govern the phase transitions in DA.

The work done in (Xu et al., 2014) presents a deterministic annealing (DA) approach to the aggregation of Markov chains. Here, the problem formulation and the associated cost functions are fundamentally distinct from the data clustering problem in (Rose, 1998); however, the aggregation algorithm presented in (Xu et al., 2014) undergoes similar phase transition phenomenon. In particular, the algorithm begins with determining an aggregated model with a single superstate at large values of annealing parameter $T(\to\infty)$. As the annealing proceeds (i.e., T decreases), the number of superstates in the model increases at specific critical values $T = T_{\rm cr}$ resulting into a phase transition. In this paper, we explicitly determine the analytical conditions on the annealing parameter T at which the phase transition phenomenon occurrs in (Xu et al., 2014). Thereafter, analogous to the above case of clustering we quantify our notion of marginal return and the heterogeneity of a superstate based on these analytical conditions. Further, our characterization of the best choice for the number of superstates in the aggregated model is also motivated from the phase transition phenomenon. We substantiate on the exact expressions of heterogeneity and marginal return in the later sections.

Our simulations demonstrate that the marginal return value $\nu(k_t)$ is as high as 8 to 60 times the second largest value for Markov chains where the true number k_t is discernible from the heatmap of the associated transition matrices Π. For Markov chains where true number k_t is not discernible from the heatmaps, i.e. the underlying true number k_t of superstates is not visually apparent, the marginal return value $\nu(k_t)$ is still the largest and as high as 1.2 to 5 times the second largest value. In our simulations on Markov chains that model real-life scenarios - (a) emotion transitions in brain networks, and (b) letter bigram dataset, we observe that marginal return provides meaningful insights. In particular, it captures the distinct types of emotions (positive and negative) in the former, and the different types of English alphabets (vowels and consonants) in latter. This paper is organized as follows. Section 2 summarizes the MEP-based aggregation framework in (Xu et al., 2014), and derives conditions for phase transitions. Section 3 presents the main results of the paper. In Section 4 we present simulations on synthetic and real-world Markov chains.

2. Aggregation of a Markov Chain

In this section we briefly illustrate the Markov chain aggregation framework adopted in (Xu et al., 2014). As stated above this framework forms the foundation for our quantification of

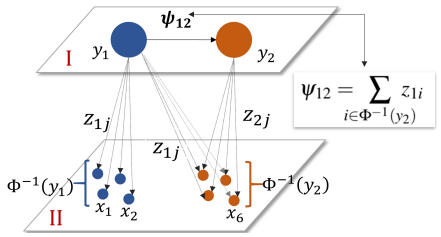

Fig. 2: The plane I represents the aggregated model for the Markov chain in the plane II. Each superstate y_j in I is associated to each state x_i in II via the weight z_{ji}. The set of states represented by superstate y_j is given by $\Phi^{-1}(y_j)$.

heterogeneity of a superstate, and the marginal return of an aggregated model. Consider a Markov chain (X,Π) with state space $X = \{x_i : 1 \le i \le N\}$, and the transition probability matrix $\Pi = (\pi_{ij}) \in \mathbb{R}^{N\times N}$. The objective is to determine an aggregated representative chain (Y,Ψ,Φ) with $M \ll N$ (super-) states, where $Y = \{y_j : 1 \le j \le M\}$ denotes the state space, $\Psi = (\psi_{jk}) \in \mathbb{R}^{M\times M}$ denotes the transition probability matrix, and $\Phi : X \to Y$ is the associated *partition* function such that the state $x_i \in X$ is represented by the superstate $\Phi(x_i) \in Y$.

The framework operates by associating a distribution vector $z(j) = (z_{j1},\dots,z_{jN}) \in \mathbb{R}^N$ to each superstate $y_j \in Y$. This distribution vector $z(j)$ is an auxiliary variable that facilitates determining the transition probability matrix Ψ of the aggregated chain. It captures the relation of the superstate y_j to all the N states in X, and consequently helps determine y_j's relation (transition probability) with other superstates in Y. Please refer to Figure 2 for a graphical interpretation of $z(j)$. Without loss of generality $\sum_{k=1}^{N} z_{jk} = 1$ and $z_{jk} \ge 0$ for all k. In the original Markov chain (X,Π) the transition probability vector $\pi(i) = (\pi_{i1},\dots,\pi_{iN})$ captures the relation of the state x_i to all the N states in X. Thus, the distance $d(x_i,y_j)$ between the state x_i and the superstate y_j is given by the *relative entropy* between probability vector $\pi(i)$ and distribution vector $z(j)$, i.e.,

$$d(x_i,y_j) = \sum_{k=1}^{N} \pi_{ik} \log \frac{\pi_{ik}}{z_{jk}} \quad \forall \ 1 \le i \le N, 1 \le j \le M. \tag{1}$$

Thereafter, the framework determines a set of M distribution vectors $\{z(j)\}_{j=1}^{M}$ such that the cumulative distance of each state x_i to its closest superstate $y_j := \Phi(x_i)$ is minimized, i.e., it solves the optimization problem

$$\begin{aligned} \min_{Z,\Phi} \quad & D(\Pi,Z,\Phi) = \sum_{i=1}^{N} \rho_i \, d(x_i,\Phi(x_i)) \\ \text{subject to} \quad & z_{jk} \ge 0 \ \forall \ j,k \text{ and } z(j)^\top \mathbf{1}_N = 1 \ \forall \ j, \end{aligned} \tag{2}$$

where ρ_i is a given relative weight of the state x_i (for instance, $\rho_i = 1/N$). Subsequently, (Xu et al., 2014) determines the transition matrix $\Psi = (\psi_{jk})$ of the aggregated chain in terms of the distribution vectors $\{z(j)\}_{j=1}^{M}$. In particular, the transition probability ψ_{jk} from the superstate y_j to y_k is the cumulative sum of the weights z_{ji}'s from superstate y_j to all states x_i that are represented by the superstate y_k (see Figure 2), i.e.,

$$\psi_{jk} = \sum_{x_i \in \Phi^{-1}(y_k)} z_{ji} \ \forall \ 1 \le j,k \le M. \tag{3}$$

The partition function $\Phi : X \to Y$ in (2) associates the state x_i to a *particular* superstate $y_j = \Phi(x_i) \in Y$. In its solution methodology, (Xu et al., 2014) replaces this partition function by *soft* partition weights $p_{j|i} \in [0,1]$ that determine the association of the state x_i to different superstates $\{y_j \in Y\}$. Without loss of generality $\sum_j p_{j|i} = 1$ and $p_{j|i} \geq 0$ for all i, j. Thereafter, Maximum Entropy Principle (MEP) is utilized to design the partition weights $\{p_{j|i}\}$. In particular, (Xu et al., 2014) determines the most *unbiased* partition weights $\{p_{j|i}\}$ and the distribution vectors $\{z(j)\}$ that maximize the corresponding Shannon entropy H such that the expected cost function $\mathbb{E}_p[D]$ attains a pre-specified value d_0, i.e, the framework solves

$$\begin{aligned} \max_{\{p_{j|i}\},\{z(j)\}} \quad & H = -\sum_{i=1}^{N} \rho_i \sum_{j=1}^{M} p_{j|i} \log p_{j|i} \\ \text{subject to} \quad & \mathbb{E}_p[D] := \sum_{i=1}^{N} \rho_i \sum_{j=1}^{M} p_{j|i} d(x_i, y_j) = d_0 \\ & z_{jk} \geq 0,\ z(j)\mathbf{1}_N = 1\ \forall j,k. \end{aligned} \tag{4}$$

Minimizing (local) the Lagrangian $\mathscr{L} = (\mathbb{E}_p[D] - d_0) - TH$ (where T is the Lagrange parameter) of above optimization problem with respect to $\{p_{j|i}\}$ and $\{z(j)\}$ results into

$$p_{j|i} = \frac{\exp\{-(1/T)d(x_i,y_j)\}}{\sum_k \exp\{-(1/T)d(x_i,y_k)\}}, \quad Z = P'\Pi \tag{5}$$

$$\text{where } [P]_{ij} = \frac{\rho_i p_{j|i}}{\sum_t \rho_t p_{j|t}} \text{ and } Z = [z(1), \ldots, z(M)]^\top. \tag{6}$$

The resulting Lagrangian $\mathscr{L}(Z)$ is given by

$$\mathscr{L}(Z) = -T \sum_{i=1}^{N} \rho_i \log \sum_{j=1}^{M} \exp\Big\{ -\frac{1}{T} \sum_{k=1}^{N} \pi_{ik} \log \frac{\pi_{ik}}{z_{jk}} \Big\}. \tag{7}$$

It is known from the sensitivity analysis (Jaynes, 2003) that the large (small) values of d_0 in (4) is analogous to large (small) values of the Lagrange parameter T. In fact, the optimization problem in (4) is repeatedly solved at decreasing values of T (or equivalently, decreasing values of d_0). At large values of $T \to \infty$, the Lagrangian $\mathscr{L}$ is dominated by the convex function $-H$, the partition weights $\{p_{j|i}\}$ in (5) are uniformly distributed ($p_{j|i} = 1/M$), and all the distributions $\{z(j)\}$ in (5) are *co-incident*, i.e., effectively *one* distinct superstate is obtained at large values of T. As T decreases further the Lagrangian is more and more dominated by the cost function $\mathbb{E}_p[D]$, the partition weights $\{p_{j|i}\}$ are no longer uniform, and the coincident distributions $\{z(j)\}$ split into distinct groups, i.e., the effective number of distinct superstates increases.

In this article we exploit this heirarchical splitting to define the notions of heterogeneity and the marginal return. Insights from the simulations of the algorithm in (Xu et al., 2014) demonstrate that there exist certain critical temperature $T = T_{\text{cr}}$ values at which the solution undergoes the phenomenon of *phase transition*. As briefly illustrated in Section 1, this phenomenon is characterized by the increase in the number of distinct distributions in $\{z(j)\}$; or, equivalently increase in the number of distinct superstates. These critical temperatures T_{cr}'s occur when the critical point Z^* of the Lagrangian $\mathscr{L}$, given by $\frac{\partial \mathscr{L}(Z^*)}{\partial Z} = 0$ is no longer the local minima, that is, when for at least one perturbation direction $\Psi = [\psi_1, \ldots, \psi_M]^\top \in \mathbb{R}^{M \times N}$, the Hessian $\mathscr{H}(Z^*, P^*, \Psi, T) :=$

$$\begin{aligned} \lim_{\varepsilon \to 0} \frac{\partial^2 \mathscr{L}(Z^* + \varepsilon\Psi)}{\partial \varepsilon^2} = & \sum_{j=1}^{M} q_j \psi_j^\top (\Lambda_T(j) - \frac{1}{T} C_T(j)) \psi_j \\ & + T \sum_{i=1}^{N} \Big(\sum_{j=1}^{M} p_{j|i} \Big[\frac{\pi(i).}{z^*(j)} \Big]^\top \psi_j \Big)^2, \end{aligned} \tag{8}$$

is no longer positive. Here, $q_j = \sum_{i=1}^{N} \rho_i p_{j|i}$.

$$\begin{aligned} \Lambda_T(j) &= \text{diag}\Big\{ \frac{\sum_{i=1}^{N} p_{i|j} \pi(i).}{z^*(j).^2} \Big\}, \\ C_T(j) &= \sum_{i=1}^{N} p_{i|j} \Big[\frac{\pi(i) - z^*(j).}{z^*(j)} \Big] \Big[\frac{\pi(i) - z^*(j).}{z^*(j)} \Big]^\top, \end{aligned} \tag{9}$$

and $p_{i|j} = (\rho_i p_{j|i} / \sum_t \rho_t p_{j|t})$ is the posterior distribution. We derive the expression for critical temperature T_{cr} as below.

Theorem 1. *The value of critical temperature T_{cr} at which the the Hessian $\mathscr{H}(Z^*, P^*, \Psi, T)$ in (8) is no longer positive for some perturbation direction Ψ, and Z^* in (5) undergoes phase transition is given by*

$$T_{cr} := \max_{1 \leq j \leq M} \Big[T_{cr,j} \Big], \text{ where } T_{cr,j} = \lambda_{\max}\big(\mathscr{C}_T(j)\big), \text{ and} \tag{10}$$

$$\mathscr{C}_T(j) = \sum_{i=1}^{N} [P]_{ij} \Big[\Theta^\top \frac{\pi(i) - z^*(j).}{z^*(j)} \Big] \Big[\Theta^\top \frac{\pi(i) - z^*(j).}{z^*(j)} \Big]^\top. \tag{11}$$

Here, $\lambda_{\max}(\cdot)$ is the largest eigenvalue, $\frac{a.}{b}$ denotes element-wise division of vectors a and b, and $\Theta \in \mathbb{R}^{N \times N-1}$ is a parameter that arises due to the constraint $z(j)^\top \mathbf{1}_N = 1\ \forall\ 1 \leq j \leq M$ in (2) and (4); see Appendix A for exact Θ, and further details.

Proof. Please refer to the Appendix A. □

Interpreting T_{cr}, $T_{cr,j}$, and $\mathscr{C}_T(j)$: As stated earlier, the annealing temperature $T := T_{cr}$ (computed in (10)-(11)) marks the increase in the number of distinct superstates in the aggregated chain, i.e., it characterizes the phase transitions in the DA-based algorithm (Xu et al., 2014). However, the expressions in (10)-(11) are also further interpretable beyond the current context of phase transitions. For instance, $\mathscr{C}_T(j)$ in (11) is a *soft co-variance matrix* of the posterior distribution $[P]_{ij}$ corresponding to the superstate $y_j \in Y$. Since, $T_{cr,j}$ in (10) is the maximum eigenvalue of $\mathscr{C}_T(j)$ it captures the *maximum variance* within the transition probabilities $\{\pi(i)\}$ of the states $\{x_i\}$ represented by the superstate y_j. In other words, $T_{cr,j}$ can be interpreted as a *measure of heterogeneity* within the states represented by the superstate y_j, and T_{cr} can be interpreted as the maximum heterogeneity among all the superstates $\{y_j \in Y\}$.

The soft co-variance matrix $\mathscr{C}_T(j)$ in (11) depends on the partition weights $[P]_{ij}$ and distribution vectors $Z = P^\top \Pi$ in (6) that result from the aggregation algorithm in (Xu et al., 2014) − making the above quantification $T_{cr,j}$ of heterogeneity dependent on the algorithm. In the following section, we adapt $\mathscr{C}_T(j)$ to incorporate aggregated chains irrespective of the algorithm used to determine them. In particular, for a given aggregated

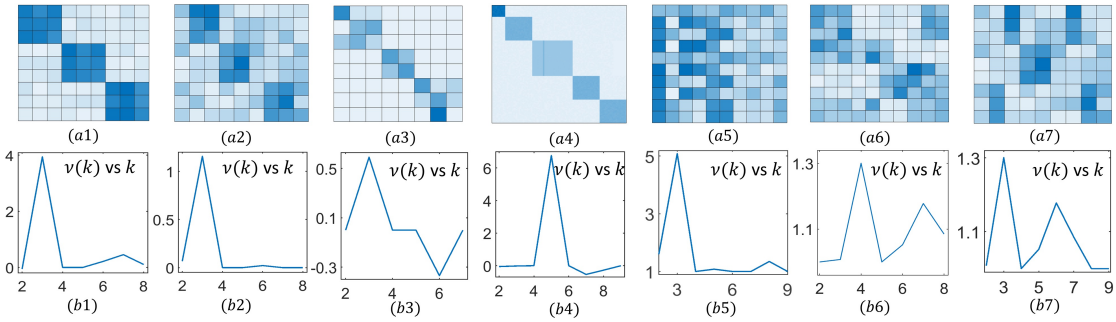

Fig. 3: Illustrates the efficacy of Algorithm 1 in estimating k_t. (a1)-(a4) demonstrate the heatmaps for the transition matrices of NCD Markov chains generated such that $N = 9$, $k_t = 3$ in (a1)-(a2), $N = 8$, $k_t = 3$ in (a3), and $N = 100$, $k_t = 5$ in (a4). (b1)-(b4) are the corresponding persistence plots t hat clearly indicate $\nu(k_t) > \nu(k)$ $\forall\, k \neq k_t$. (a5)-(a7) illustrate Markov chains generated by considering multiple copies of random vectors $\{\xi_i \in \mathbb{R}^N\}_{i=1}^{k_t}$; where $N = 10$, $k_t = 3$ in (a5), $N = 10$, $k_t = 4$ in (a6), and $N = 9$, $k_t = 3$ in (a7). (b5)-b(7) are the corresponding persistence plot that accurately estimate k_t.

representation (Y,Ψ,Φ) of a Markov chain (X,Π), we replace the soft partitions $[P]_{ij}$ in $\mathscr{C}_T(j)$ with the hard-partitions prescribed by the partition function $\Phi : X \to Y$. Subsequently, we provide the expression for *marginal return*, and the Algorithm 1 that estimates a choice for the number of superstates to be considered in the aggregated chain.

3. Covariance Matrix and Marginal Return

Consider a Markov chain (X,Π) with states $X = \{x_i : 1 \leq i \leq N\}$, and the transition probability matrix $\Pi = (\pi_{ij}) \in \mathbb{R}^{N\times N}$, its aggregated representative chain (Y,Ψ,Φ) with states $Y = \{y_j : 1 \leq j \leq M\}$, the transition probability matrix $\Psi = (\psi_{jk}) \in \mathbb{R}^{M\times M}$, where $M \ll N$, and the partition function $\Phi : X \to Y$ such that the $x_i \in X$ is represented by the $y_j := \Phi(x_i) \in Y$.

Definition 1. The *co-variance matrix* $C_X^{\Phi}(j)$ corresponding to the superstate $y_j \in Y$ in the aggregated chain is given by

$$C_X^{\Phi}(j) := \sum_{i=1}^{N} [Q]_{ij} \Big[\Theta^\top \frac{(\pi(i) - w(j))_\cdot}{w(j)}\Big] \Big[\Theta^\top \frac{(\pi(i) - w(j))_\cdot}{w(j)}\Big]^\top, \quad (12)$$

$$\text{where } [Q]_{ij} = \begin{cases} 1, & \text{if } x_i \in \Phi^{-1}(y_j) \\ 0, & \text{otherwise} \end{cases}, \quad Q := [Q]_{ij}, \quad (13)$$

$W = [w(1),\ldots,w(M)]^\top = Q^\top\Pi$ denotes the distribution vector as obtained in (5), $\Theta \in \mathbb{R}^{N\times N-1}$ corresponds to the constraint $z(j)^\top \mathbf{1}_N = 1\ \forall\ 1 \leq j \leq M$ in (2), $\frac{a_\cdot}{b}$ denotes element-wise division of vectors a and b, and $[\cdot]^\top$ denotes the transpose.

Definition 2. The *marginal return* $\nu(k)$ of the aggregated model with k number of superstates among the given K aggregated representations $\{(Y_k,\Psi_k,\Phi_k) : |Y_k| = k\}_{k=1}^K$ of a Markov chain (X,Π) is given by

$$\nu(k) := \log \bar{T}_{k-1} - \log \bar{T}_k, \text{ where} \quad (14)$$

$$\bar{T}_l := \max_{1\leq j\leq l} [\bar{T}_{l,j}] \text{ for } l \in \{k-1,k\}, \bar{T}_{l,j} = \lambda_{\max}(C_X^{\Phi_l}(j)) \quad (15)$$

is heterogeneity of superstate y_j, and $\lambda_{\max}$ is largest eigenvalue.

As illustrated in Section 1, the aggregated model with largest marginal return provides an appropriate choice for the number k_t of superstate underlying a Markov chain, i.e.

$$k_t := \arg\max_{1\leq k\leq K} \nu(k), \quad (16)$$

The following algorithm computes the marginal return of the K representations $\{(Y_k,\Psi_k,\Phi_k) : |Y_k| = k\}_{k=1}^K$ of a Markov chain (X,Π), and estimates the number k_t of superstates.

Algorithm 1 Marginal return and Number of superstates

Input: (X,Π), $\{(Y_k,\Psi_k,\Phi_k) : |Y_k| = k\}_{k=1}^K$; **Output:** $\nu(k)$, k_t
for $k = 1$ *to* K **do**
  From partition $\Phi_k : X \to Y$ determine Q in (13)
  Compute $W := [\cdots w(j) \cdots]^\top = Q^\top\Pi$ and $\bar{T}_k$ using (15).
compute $\nu(k)$ in (14) $\forall\ 2 \leq k \leq K$, and $k_t := \arg\max_k \nu(k)$.

4. Simulations

In this section we demonstrate the efficacy of marginal return in comparing different aggregated models, and estimating true number k_t of superstates underlying a given Markov chain. We use Algorithm 1 that takes in the aggregated models at different number k of superstates as inputs, and outputs the corresponding marginal return and k_t. To demonstrate the generality of our method, we use aggregated models obtained from different aggregation algorithms. We use the algorithm presented in (Geiger, 2014) on the first four example simulations in Figure 3, and (Xu et al., 2014) on the remaining examples. We would like to remark that both the algorithms perform well and result into meaningful aggregations at different number of superstates.

NCD Markov Chains: The transition matrix Π for the NCD Markov chains (Ando and Fisher, 1963) presumes the following structure $\Pi = \Pi^* + \varepsilon C$, where Π^* is a block diagonal matrix and C adds a perturbation of the range ε. Naturally, the number k_t of superstates for such Markov chains can be approximated to be the number of block diagonals in Π^*. Figures 3(a1)-a(2) illustrate the heatmaps of the transition probability matrix for two such Markov chains obtained at different levels of perturbation to a three block diagonal matrix Π^*. Figures 3(b1)-(b2) illustrate the corresponding persistence $\nu(k)$ versus k plots which correctly identifies the true number $k_t = 3$ of superstates in both cases. Note that, even though the transition matrix Π in Figure 3(a2) is highly perturbed, our method accurately identifies number of superstates thereby, demonstrating the robustness of the marginal return $\nu(k)$.

Figure 3(a3) illustrates a specialized NCD Markov chain with Courtois Transition Matrices (Elsayad, 2002). These Markov chains are well studied in literature for their slow convergence to steady state. As is evident from the above heatmap

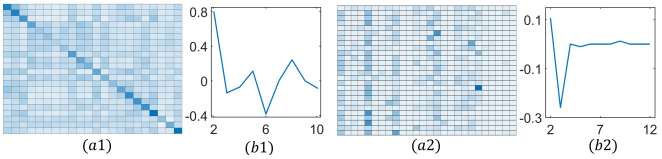

Fig. 4: Marginal return analysis in realistic Markov chains. (a1) Transition matrix for brain emotion transitions where each state corresponds to an emotion. (b1) Maximum $\nu(k)$ observed at $k_t = 2$ suggesting two types of underlying emotions. (a2) Transition matrix for letter bigram dataset. Each state correspond to an English alphabet. (b2) Largest $\nu(k)$ at $k_t = 2$ indicating two types of alphabets in English language - vowels and consonants.

in Figure 3(a3), the transition matrix comprises of 3 dominant block diagonal. Thus, the corresponding Markov chain comprises of $k_t = 3$ number of superstates. This is captured in our marginal return analysis in the Figure 3(b3). Figure 3(a4) demonstrates the state transition matrix of a large ($N = 100$ states) NCD Markov chain constituting 5 underlying block diagonals (one each of size 10×10 and 30×30, and three of size 20×20). Figure 3(b4) confirms largest marginal return at $k = 5$ number of superstates, and thus, appropriately estimates k_t.

Randomly generated Markov Chains: In the following simulations we consider N state Markov chains where the transition matrices $\Pi = [\xi_{i_1}, \ldots, \xi_{i_N}]^\top + \varepsilon C$ are generated from k_t random distribution vectors $\{\xi_i\}_{i=1}^{k_t}$, and εC introduces random perturbations. Naturally, k_t estimates true number of superstates underlying the above Markov chains. Figure 3(a5) illustrates one such scenario where $\Pi \in \mathbb{R}^{10\times 10}$ is generated by perturbing multiple copies of $\{\xi_i\}_{i=1}^{3}$, and the marginal return analysis in Figure 3(b5) accurately determines $k_t = 3$ as the estimate for the true number of superstates. Similarly, Figures 3(a6) and 3(a7) illustrate the transition matrices $\Pi \in \mathbb{R}^{10\times 10}$ generated from 4 and 3 random distribution vectors, respectively, and the marginal return plots in Figure 3(b6) and 3(b7) accurately estimate the underlying true number of superstates. Note that even though the perturbations are significantly large in Figures 3(a5)-(a7), marginal return $\nu(k)$ accurately estimates the number k_t of superstates underlying the original Markov chains.

Emotion Transitions in Brain Network: We consider the example Markov chain that models the transition between different emotional states of a person (Thornton and Tamir, 2017). The Figure 4(a1) illustrates the transition matrix Π comprising $N = 22$ states - each corresponding to an emotion in the set $X = \{$*anxious, jittery, irritable, vigorous, alert, lively happy, attentive, intense, full-of-pep, excited, distressed, strong, stirred-up, nervous, upset, touchy, bold, temperamental, quiet, talkative, insecure*$\}$. We obtain different aggregated representations of Π using the algorithm presented in (Xu et al., 2014), and observe that the marginal return $\nu(k)$ is the largest for $k_t = 2$ (see Figure 4(b1)). This indicates the presence of two superstates underlying the original Markov chain in Figure 4(a1). We note that the largest marginal return $\nu(k)$ at $k_t = 2$ is in accordance with the neuro-science literature (Thornton and Tamir, 2017) which also classifies emotions $X = X_p \sqcup X_n$ into positive X_p and negative X_n emotions, where $X_p = \{$vigorous, alert, lively, happy, attentive, intense, full-of-pep, excited, strong, bold, quiet$\}$, and $X_n = \{$anxious, jittery, irritable, vigorous, distressed, stirred-up, nervous, upset, touchy, temperamental, insecure, talkative$\}$. Thus, marginal return captures the number of distinct type of emotions in X. Similar analysis can be done on different brain network models (Vidaurre et al., 2017) to qualify, for instance, the network states responsible for either *sensory-motor systems*, or for *higher-order cognition* (such as language).

Table 1: State aggregations of letter bi-gram data obtained at different number of superstates. Obtained via aggregation algorithm in (Xu et al., 2014).

k	State Partitioning
2	$\{aeiouy\}, \{bcdfghjklmnpqrstvwxz\}$
3	$\{aeiouy\}, \{cdghknrstvwxz\}, \{bfjlmpq\}$
4	$\{aeiouy\}, \{hnrsvxz\}, \{cdgkwt\}, \{bfjlmpq\}$
5	$\{aeiouy\}, \{bflmp\}, \{cdgktw\}, \{hnrlvxz\}, \{jq\}$

Letter Bigram dataset: We finally investigate an example from Natural Language Processing. We consider the letter bigram dataset (Norvig, 2013) that enumerates the number of transitions from one alphabet to another based on the Google Corpus Data (collection of $97,565$ distinct words, which were mentioned $743,842,922,321$ times). We represent the entire dataset as a Markov chain where each state represents an individual alphabet. The transition matrix $\Pi \in \mathbb{R}^{26\times 26}$ (see Figure 4(a2)) of the Markov chain captures the frequency of transition from one alphabet (state) to another. Our idea is to demonstrate that *marginal return* captures a "meaningful" number of superstates to be considered in the aggregated representation of the above Markov chain. Table 1 illustrates aggregated models obtained at different number of superstates. Since the alphabets are of two types, i.e., either vowels or consonants, one may find it reasonable to aggregate the original Markov chain into two superstates - one each for vowels and consonants. The above intuitive argument is reinforced by the largest value of marginal return $\nu(k)$ that is obtained at $k_t = 2$ (see Figure 4(b2)). Note that the corresponding aggregated model at $k_t = 2$ (as illustrated in Table 1) partitions the alphabets into either vowel sounds $\{aeiouy\}$, or the consonants (where y is a (semi)-vowel). We also emphasize that the $k_t = 2$ superstates is not apparent from the transition matrix Π in Figure 4(a2), however, the notion of marginal return successfully captures it.

Remark. The notion of marginal return is useful in evaluating the performance of a proposed aggregation algorithm. This can be done as follows. Consider a Markov chain where the number k_t of superstates are known apriori. Use Algorithm 1 to obtain the marginal return $\nu(k)$ corresponding to the representative chains obtained from the proposed aggregation algorithm. The performance of the aggregation algorithm is assessed based on whether the largest marginal return coincides with k_t or not.

5. Conclusion

Model-reduction techniques usually rely on the pre-specified size of the reduced model. Thus, there arises a need to methodically estimate this size. In the context of Markov chain aggregation, we devise the notions of heterogeneity and marginal return. We demonstrate that the proposed notions are insightful in comparing aggregated models of different sizes, and estimating the number of superstates underlying the original Markov chain. The ideas and methods presented in this paper are extendable to several related model-reduction problems. For in-

stance, the graph clustering problem (Xu et al., 2014), that requires aggregating the nodes of a large graph into *supernodes*, or the co-clustering problem (Dhillon et al., 2003), that aggregates a given matrix $\mathscr{X} \in \mathbb{R}^{N_1 \times N_2}$ to a smaller representative matrix $\mathscr{Y} \in \mathbb{R}^{M_1 \times M_2}$. The appropriate size of the representative graph in the former, or the aggregated matrix Y in the latter can be estimated using similar ideas as elucidated in our work.

Acknowledgments

We thank Bernhard Geiger, Rana Ali Amjad, Clemens Bloechl for sharing code illustrated in (Amjad et al., 2019).

Funding: Our work is supported by NSF grant ECCS (NRI) 18-3063, DOE Subaward (WPI) 10809-GR, and the UIUC-ZJUI Center for Adaptive, Resilient Cyber-Physical Manufacturing Networks.

Appendix A.

Perturbation $\Psi = [\psi_1, \ldots, \psi_M] \in \mathbb{R}^{M \times N}$. Let $\psi_j = \Sigma K_j$, where $\Sigma \in \mathbb{R}^{N \times N}$ and $\Phi \mathbf{1}_N = 0$ for admissible perturbation of $z(j)$ (i.e. $(z(j) + \varepsilon \psi_j)\mathbf{1}_N = 1$ in (2)).

Lemma 1. *The Hessian $\mathscr{H}(Z^*, P^*, \Psi, T)$ in (8) loses rank when $det\big[\Sigma^\top \big(\Lambda_T(j) - \frac{1}{T} C_T(j)\big)\Sigma\big] = 0$ for some $1 \le j \le M$, where $\Lambda_T(j)$ and $C_T(j)$ are as defined in (9).*

Proof. Motivated from (Rose, 1998). $\mathscr{H}(Z^*, P^*, \Psi, T)$ is positive for all perturbation Ψ if and only if its first part (see (8)) is positive. 'If' part is straightforward. For the 'only if' part we show that $\exists$ a Ψ such that second term in $\mathscr{H}$ vanishes. Let $J = \{j_1, \ldots, j_r\}$ denote the superstates with distribution z_{j_0}. Select Ψ such that $\psi_j = 0 \; \forall \; j \notin J$ and $\sum_{j \in J} \psi_j = 0$. For this perturbation the second term vanishes and $\mathscr{H}$ depends only on its first term. Thus, we establish the 'only if' part. Replacing the ψ_j in (8) with ΣK_j we obtain the condition stated in the Lemma. □

Lemma 2. *The condition $det\big[\Sigma^\top \big(\Lambda_T(j) - \frac{1}{T} C_T(j)\big)\Sigma\big] = 0$ for some $1 \le j \le M$ is attained at temperature value $T = T_{cr} := \lambda_{\max}\big(\mathscr{C}_T(j)\big)$, where $\mathscr{C}_T(j)$ is given in (11).*

Proof. Rank(Σ)$= N-1$. Let $\Upsilon \in \mathbb{R}^{N \times N-1}$ such that Range(Υ) = Range(Σ). Let $W_j \in \mathbb{R}^{N-1}$ so that $\Upsilon W_j = \Sigma K_j$. From Lemma 1 $\exists \; W_j : W_j \Upsilon^\top \big(\Lambda_T(j) - \frac{1}{T} C_T(j)\big)\Upsilon W_j = 0$. Let $H_0 = \Upsilon^\top \Lambda_T(j) \Upsilon$, $H_1 = \Upsilon^\top C_T(j) \Upsilon$. From Theorem 12.19 in (Laub, 2005) $\exists$ $G \in \mathbb{R}^{N \times N-1}$ s.t. $G^\top H_0 G = I$ and $G^\top H_1 G = \mathscr{C}_T(j)$ where $G = L^{-\top} P$, $LL^\top = H_0$, $P^\top [L^{-1} H_1 L^{-\top}] P = \mathscr{C}_T(j)$. Let $W_j = G\omega_j$ for some $\omega_j \in \mathbb{R}^{N-1}$, $\Rightarrow \omega_j^\top (G^\top H_0 G - \frac{1}{T} G^\top H_1 G)\omega_j = 0 \Rightarrow$ $\det\big(I - \frac{1}{T}\mathscr{C}_T(j)\big) = 0 \Rightarrow T_{cr} = \lambda_{\max}\big(\mathscr{C}_T(j)\big)$. Since $G^\top H_1 G = \mathscr{C}_T(j)$ we have that $\Theta = \Upsilon G$. □

References

Aldhaheri, R.W., Khalil, H.K., 1991. Aggregation of the policy iteration method for nearly completely decomposable markov chains. IEEE Transactions on Automatic Control 36, 178–187.

Amjad, R.A., Blöechl, C., Geiger, B.C., 2019. A generalized framework for kullback-leibler markov aggregation. IEEE Transactions on Automatic Control .

Ando, A., Fisher, F.M., 1963. Near-decomposability, partition and aggregation, and the relevance of stability discussions. International Economic Review 4, 53–67.

Beck, C.L., Lall, S., Liang, T., West, M., 2009. Model reduction, optimal prediction, and the mori-zwanzig representation of markov chains, in: Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, IEEE. pp. 3282–3287.

Deng, K., Mehta, P.G., Meyn, S.P., 2011. Optimal kullback-leibler aggregation via spectral theory of markov chains. IEEE Transactions on Automatic Control 56, 2793–2808.

Dhillon, I.S., Mallela, S., Modha, D.S., 2003. Information-theoretic co-clustering, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 89–98.

Elsayad, A.L., 2002. Numerical solution of markov chains .

Feng, Y., Hamerly, G., 2007. Pg-means: learning the number of clusters in data, in: Advances in neural information processing systems, pp. 393–400.

Gagniuc, P.A., 2017. Markov chains: from theory to implementation and experimentation. John Wiley & Sons.

Geiger, B.C., 2014. Markov state space aggregation via the information bottleneck method. Schedae Informaticae 23.

Hamerly, G., Elkan, C., 2004. Learning the k in k-means, in: Advances in neural information processing systems, pp. 281–288.

Jaynes, E.T., 2003. Probability theory: The logic of science. Cambridge university press.

Kalogeratos, A., Likas, A., 2012. Dip-means: an incremental clustering method for estimating the number of clusters, in: Advances in neural information processing systems, pp. 2393–2401.

Laub, A.J., 2005. Matrix analysis for scientists and engineers. volume 91. Siam.

Norvig, P., 2013. English letter frequency counts: Mayzner revisited or etaoin srhldcu. Dostopno na http://www. norvig. com/mayzner. html.[obiskano 2016-08-07] .

Pelleg, D., Moore, A.W., et al., 2000. X-means: Extending k-means with efficient estimation of the number of clusters., in: Icml, pp. 727–734.

Quinn, C.J., Coleman, T.P., Kiyavash, N., Hatsopoulos, N.G., 2011. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. Journal of computational neuroscience 30, 17–44.

Rached, Z., Alajaji, F., Campbell, L.L., 2004. The kullback-leibler divergence rate between markov sources. IEEE Transactions on Information Theory 50, 917–921.

Rose, K., 1998. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proceedings of the IEEE 86, 2210–2239.

Sledge, I.J., Pri'ncipe, J.C., 2019. An information-theoretic approach for automatically determining the number of state groups when aggregating markov chains, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 3612–3616.

Srikant, R., 2004. The mathematics of Internet congestion control. Springer Science & Business Media.

Srivastava, A., Baranwal, M., Salapaka, S., 2019. On the persistence of clustering solutions and true number of clusters in a dataset, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5000–5007.

Sugar, C.A., James, G.M., 2003. Finding the number of clusters in a dataset: An information-theoretic approach. Journal of the American Statistical Association 98, 750–763.

Thornton, M.A., Tamir, D.I., 2017. Mental models accurately predict emotion transitions. Proceedings of the National Academy of Sciences 114, 5982–5987.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, 411–423.

Vidaurre, D., Smith, S.M., Woolrich, M.W., 2017. Brain network dynamics are hierarchically organized in time. Proceedings of the National Academy of Sciences 114, 12827–12832.

Vidyasagar, M., 2012. A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. IEEE Transactions on Automatic Control 57, 2464–2477.

Xu, Y., Salapaka, S.M., Beck, C.L., 2013. A distance metric between directed weighted graphs, in: 52nd IEEE Conference on Decision and Control, IEEE. pp. 6359–6364.

Xu, Y., Salapaka, S.M., Beck, C.L., 2014. Aggregation of graph models and markov chains by deterministic annealing. IEEE Transactions on Automatic Control 59, 2807–2812.

Zhang, Q., Yin, G., 2004. Nearly-optimal asset allocation in hybrid stock investment models. Journal of Optimization theory and Applications 121, 419–444.