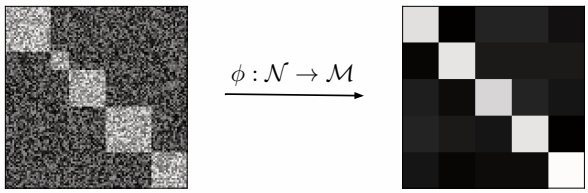
TECHNISCHE UNIVERSITÄT DARMSTADT

Spectral methods for Markov chain aggregation

Markov chains are a versatile tool for modelling sequential data. Once such a model has been obtained, a frequent and general challenge one is facing is that of *model reduction*. While retaining a 'sufficient' degree of information, the model should be simplified as much as possible.
An example of a concrete application are models of protein or RNA folding. The temporal dynamics of such molecules can be simulated, which yields time course data of 3D positions of the constituent atoms/particles. These time course data however are not readily comprehensible, however. For this among other reasons, it is a valid approach to construct a discrete-state Markov model for these data. To that end, one needs to employ a discretisation scheme on the (continuous) state space. Because it is unknown in advance what a suitable discretisation might be, a typical strategy is to choose a fairly fine-grained mesh. To arrive at an informative, yet simple model, one then has to approximate this with a more coarse-grained Markov chain.
Mathematically speaking, we start with a Markov chain on a fine-grained state space $\mathcal{N} = \{1, ..., n\}$. We wish to project this model to a more coarse-grained one, $\mathcal{M} = \{1, ..., m\}$, where $m < n$. We hence search for a partition function $\phi : \mathcal{N} \to \mathcal{M}$ that yields the 'best' $\mathcal{M}$-approximation (in a sense we have to strictly define). We arrive at a transition matrix $\mathbf{Q}$ in $\mathcal{M}$ by *lumping* or *aggregating* together states from $\mathcal{N}$.



$\mathbf{P} \in \mathbb{R}^{n \times n}$
$\mathcal{N} = \{1, ..., n\}$

$\mathbf{Q} \in \mathbb{R}^{m \times m}$
$\mathcal{M} = \{1, ..., m\}$
$m < n$

This project is concerned with methods to arrive at this representation $\mathbf{Q}$ by finding a suitable partition function ϕ. Much work on this has already been done, where a particularly apt starting point can be found in [1]. The *m-ary partitioning* scheme presented in [1] should be compared with the Perron Cluster Cluster Analysis approach (PCCA) [2]. These methods will be tested not only on synthetic data, but also on data from protein or RNA simulations, where a Markov model has been already constructed. Construction of these (specifically termed Markov State Models) is described in [3]. If time permits, further analysis should examine the connections of these spectral approaches to the information bottleneck method presented in [4].

Prerequisites are good knowledge of linear algebra; a basic understanding of probability theory is desirable. Coding should be done in python.

For further information, please contact Lukas Köhs.

Fachbereich 18
Elektrotechnik und Informationstechnik
Bioinspirierte Kommunikationssysteme

Department 18
Electrical Engineering and Information Technology
Bioinspired Communication Systems

Prof. Dr. Heinz Koeppl
Head of lab

Lukas Köhs
Project supervisor

Rundeturmstraße 12
64283 Darmstadt

Phone: +49 6151 16 - 57 243
lukas.koehs@bcs.tu-darmstadt.de
https://www.bcs.tu-darmstadt.de

November 2019

References

[1] K. Deng, P. G. Mehta, and S. P. Meyn. Optimal Kullback-Leibler Aggregation via Spectral Theory of Markov Chains. *IEEE Transactions on Automatic Control*, 56(12):2793–2808, December 2011.

[2] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 315(1-3):39–59, August 2000.

[3] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105, May 2011.

[4] B. C. Geiger, T. Petrov, G. Kubin, and H. Koeppl. Optimal Kullback–Leibler Aggregation via Information Bottleneck. *IEEE Transactions on Automatic Control*, 60(4):1010–1022, April 2015.