KL Divergence of Reciprocal Chains

George Stamatescu, Langford B White, *Senior Member, IEEE,*

***Abstract*—This letter presents new results on the application of Kullback-Leibler (KL) divergence to discrete parameter reciprocal chain (RC) models. The first result is an algorithm for computing the KL divergence between two RC models with different dynamics, defined on the same finite state space. This algorithm is based on the Markov bridge (MB) characterisation of a RC. The second and third results are derivations of approximate models by determining the closest dynamics (in the KL sense) of an inhomogeneous Markov chain (MC) model and a homogeneous MC model to a given RC model.**

I. Introduction

1) Reciprocal processes and KL divergence: Reciprocal processes are a non-causal generalisation of Markov processes which arose from a problem studied by Schrödinger in 1931, which we now recount. Suppose a Markov process is observed have an initial distribution $\pi(0)$ and, at a later time T, a distribution $\pi(T)$ that differs from what is predicted by the *a priori* model of its dynamics. What are the dynamics for this observed process which are closest in some suitable sense to the *a priori* model? In modern terms, this problem is formulated as a minimisation of the KL divergence between the two Markov processes [1]-[4]. In his original solution to the problem, Schrödinger introduced what we now call reciprocal dynamics, since he correctly deduced that the Markov process with the marginals $\pi(0)$ and $\pi(T)$ must have the same two-sided non-causal dynamics as the *a priori* process in order to minimise the KL divergence. Thus there is a long history of applications involving KL divergence and Markovian reciprocal processes. In this letter, we instead apply the concepts of entropy and KL divergence to non-Markovian reciprocal processes, which to our knowledge, is the first such attempt.

2) RC background : Recently, finite state reciprocal processes, which we shall call *reciprocal chains* (RC), have been of interest (e.g. [5]-[8]). A significant advance, which is also of considerable practical importance, was made in [6] where the characterisation of a RC in terms of a finite number of Markov bridges (MB) was given. In [6], this formulation allowed the derivation of optimal smoothers for RC measured in noise. An important application of RC is target tracking (e.g. [5], [7]), where the "intent" of a target to move between a given origin and a destination can be better modelled than with Markov chain (MC) based models. The RC based target motion model was generalised in [9] to include "waypoints".

3) Information theory background: There is a long history of studying the information theoretic properties of MC models [10]. The main applications are essentially asymptotic (for large sequence lengths) and fundamental to these studies is the idea of the entropy rate of a MC model, first presented in [13]. For details on the history and applications of entropy or KL divergence to MC models see [10]. Since RC models are inherently of finite (and specified) length sequences, asymptotic aspects of information theoretic concepts such as entropy and KL divergence are not relevant here. However, we can use these concepts to advance the understanding of RC models *vis à vis* MC models both from a theoretical and practical viewpoint.

4) Contributions: Apart from providing a computationally efficient algorithm for evaluating the KL divergence between two RC models, we can also use the underlying formulation to derive analytical formulae for the best approximating MC models (inhomogeneous and homogeneous) to a given RC model. In this letter we only consider models defined on the same finite state space. This essentially means that the processes concerned have the same number of states. However we note that in the context of MC, the problem of approximating a MC by another defined on a smaller state space, the so-called state reduction problem, has been well studied (e.g. [11], [12]). The techniques developed therein should also be relevant to RC state reduction, using the results presented in this note.

The layout of the letter is as follows : In sec. II we review the form of an RC model and briefly describe the MB characterisation. Further details are provided in [6]. In sec. III we derive the analytical formulae for the entropy and KL divergence for RC models, describe the algorithms for computing these quantities and conclude with a discussion of the potential applications. Section V considers the problems of approximating a given RC model by (i) an inhomogeneous MC model, and (ii) a homogeneous MC model.

II. Reciprocal Chain Models

Consider a random process $\mathcal{X} = \{x_t : t = 0, 1, \dots, T\}$ with $T \geq 2$, taking values in a finite set $\mathcal{S}$ of cardinality $N \geq 2$, which can be taken to be, without loss of generality, $\mathcal{S} = \{1, \dots, N\}$. Then $\mathcal{X}$ is said to be as *reciprocal chain* (RC) if

$$\mathbf{P}\left\{x_t | x_s, \forall\, s \neq t\right\} = \mathbf{P}\left\{x_t | x_{t-1}, x_{t+1}\right\}, \qquad (1)$$

for each $t = 1, \dots, T-1$. All Markov chains (MC) are reciprocal, but a RC is not, in general, a MC [14]. A RC is completely specified by its set of three-point transition functions (1) and a joint distribution over its endpoints,

$$\mathbf{P}\left\{x_0 = i, x_T = j\right\} = \Pi_{0\mathrm{T}}(i, j), \quad i, j \in \mathcal{S}\ . \qquad (2)$$

In [6], a characterisation for a RC was given in terms of a set of Markov bridges (MB). A MB is a Markov process conditioned on the value of its final state x_T. More specifically, in the finite (N) state case, a RC can be viewed equivalently as

George Stamatescu and Langford B White are with the School of Electrical and Electronic Engineering, The University of Adelaide, North Tce, Adelaide 5005, Australia. {George.Stamatescu, Lang.White}@adelaide.edu.au

N Markov bridges, each corresponding to one of the possible final states taken by x_t. This result allows the realisation of RCs and has enabled the development of estimation and detection algorithms. We use this result again in the present paper since it allows us to express the joint distribution of the states of a RC in a simplified manner, which is directly relevant to calculating entropy and KL divergence. Using direct Bayes' conditioning and the reciprocal property (1) we have,

$$\mathbf{P}\left\{x_0,\ldots,x_T\right\} = \mathbf{P}\left\{x_1,x_T\right\}\left(\prod_{t=2}^{T-1}\mathbf{P}\left\{x_t|x_{t-1},x_T\right\}\right) \quad (3)$$

This expression is equivalent to the chain rule for MC, but with the terms inside the product in (3) being the state transitions for a MB ending at x_T. The relationship between reciprocal processes and MB was studied in more general problems (not specifically discrete parameter and state) by Jamison in [14], where he also considered reciprocal processes constructed from Markov transitions of a process $\{z_t\}$. Specifically, for transition probabilities $\mathbf{P}\left\{z_{t+1}=j|z_t=i\right\}=A_{ij}(t)$ that are non-negative and sum to one, there are RC with three point transitions (1), for $t=0,\ldots,T-1$, given by

$$\mathbf{P}\left\{x_t=j|x_{t-1}=i,x_{t+1}=k\right\}=\frac{A_{ij}(t-1)A_{jk}(t)}{\sum_\ell A_{i,\ell}(t-1)A_{\ell,k}(t)}.$$

We call a Markov chain $\{z_t\}$ with transitions given by $\{A_{ij}(t)\}$ and initial distribution $\mathbf{P}\left\{z_0\right\}$, the *base process*. The method for determining the transitions of a MB ending in a state $x_T=k$, given the transitions $A_{ij}(t)$, is given by

$$B_{ij}^k(t)=\frac{A_{ij}(t)F_{jk}(t+1,T)}{F_{ik}(t,T)}\ , \quad (4)$$

where for $t<s$, we define

$$F_{ij}(t,s)=\mathbf{P}\left\{z_s=j|z_t=i\right\}=\big(A(t)\ldots A(s-1)\big)_{ij}.$$

Thus we can construct and realise RC from MB given a set of Markov transitions. Note also that the MB transitions need not be derived from a base process, they may also be constructed from arbitrary three-point transition functions (1), as described in [6].

III. Entropy and KL Divergence

The entropy of a discrete random variable X taking values in some finite set $\mathcal{S}$, with probability mass function $p(x)=\mathbf{P}\left\{X=x\right\}$, is defined to be

$$H_X=-\sum_{x\in\mathcal{S}}p(x)\log p(x)$$

We adopt the definition that $0\times\log(0)=0$, in the case $p(x)=0$. This is the definition of entropy first given by Shannon [13], for historical details see [10]. KL divergence, also known as the relative entropy, is a measure of the distance between two distributions. For two probability mass functions $p(x)$ and $q(x)$ that are defined on the same alphabet $\mathcal{S}$, the KL divergence (KLD) is defined as in [10],

$$\mathcal{D}(p||q)=\sum_{x\in\mathcal{S}}p(x)\log\frac{p(x)}{q(x)}, \quad (5)$$

provided the support of $p(x)$ is contained in the support of $q(x)$, i.e. there is no $x\in\mathcal{S}$ for which $p(x)\neq 0$ and $q(x)=0$. Thus the KLD (5) is then finite. If the support of $p(x)$ is not contained in the support of $(q(x)$, then (5) is defined to be $+\infty$. The KLD is a measure of how close two probability distributions are, although it is not a metric in the usual sense [10].

IV. Entropy and KL divergence for RCs

Extending the definitions of entropy and KLD to random processes on finite intervals is straightforward, the difference being that the distributions $p(x)$ and $q(x)$ are now over a collection of random variables, defined on the same state space, namely the values taken by all possible state sequences $\{x_0,...,x_T\}\in\{1,...,N\}^{T+1}$. Thus if we are interested in evaluating the entropy or KL divergence of a process on a fixed length interval taking values on even a small state space, we require a more efficient alternative to total enumeration.

A. Entropy of a RC

Let $\mathcal{X}=\{x_0,\ldots,x_T\}$ be a RC with known joint endpoint distrbution $\Pi_{0\mathrm{T}}$ and MB transition matrices $B_{ij}^k(t)$ as defined in sec. II, and let $p(\mathcal{X})$ be the joint probability distribution of the process. The entropy for an RC is given by

$$\begin{aligned}
H_{\mathcal{X}} &= -\sum_{\mathcal{X}}p(\mathcal{X})\log p(\mathcal{X})\\
&= -\sum_{\mathcal{X}}p(x_0,x_T)\prod_{t=1}^{T-1}p(x_t|x_{t-1},x_T)\\
&\quad\times\log\big(p(x_0,x_T)\prod_{s=1}^{T-1}p(x_s|x_{s-1},x_T)\big)\\
&= -\sum_{\mathcal{X}}p(x_0,x_T)\Bigg(\prod_{t=1}^{T-1}p(x_t|x_{t-1},x_t)\log p(x_0,x_t)\\
&\quad+\prod_{t=1}^{T-1}p(x_t|x_{t-1},x_T)\sum_{s=1}^{T-1}\log p(x_s|x_{s-1},x_T)\Bigg)\\
&= H_{0,T}+\sum_{x_0,x_T}p(x_0,x_T)H_{1|0,T}\\
&\quad+\sum_{x_0,x_T}p(x_0,x_T)\sum_{t=2}^{T-1}\sum_{x_{t-1}}p(x_{t-1}|x_0,x_T)H_{t|t-1,T}
\end{aligned}$$

In the above derivation we have used, in the following order, the MB characterisation of the RC joint distribution (3), properties of the logarithm, the law of total probability and properties of Markov chains. Note also that we have defined

$$\begin{aligned}
H_{0,T} &= -\sum_{x_0,x_T}p(x_0,x_T)\log p(x_0,x_T)\\
H_{t|t-1,T} &= -\sum_{x_t}p(x_t|x_{t-1},x_T)\log p(x_t|x_{t-1},x_T).
\end{aligned}$$

B. KL Divergence between two RCs

Now let us consider two RCs defined as above on the same state space, with joint endpoint distributions Π^1, Π^2 and two sets of different MB transition matrices $B^k_{ij}(t;1)$ and $B^k_{ij}(t;2)$. Let $p_1(\mathcal{X})$ and $p_2(\mathcal{X})$ be the joint probability distributions of each process, then the KLD between the two RCs is given by

$$\begin{aligned}\mathcal{D}(p_1\|p_2) &= \mathcal{D}_{0,T}(p_1\|p_2) \\ &+ \sum_{x_0,x_T} p_1(x_0,x_T)\Bigg(\mathcal{D}_{1|0,T}(p_1\|p_2) \\ &+ \sum_{t=1}^{T-2}\sum_{x_{t-1}} p_1(x_{t-1}|x_0,x_T)\mathcal{D}_{t|t-1,T}(p_1\|p_2)\Bigg) \quad (6)\end{aligned}$$

where we have defined

$$\mathcal{D}_{0,T}(p_1\|p_2) = -\sum_{x_0,x_T} p_1(x_0,x_T)\log\frac{p_1(x_0,x_T)}{p_2(x_0,x_T)}$$

$$\mathcal{D}_{t|t-1,T}(p_1\|p_2) = -\sum_{x_t} p_1(x_t|x_{t-1},x_T)\log\frac{p_1(x_t|x_{t-1},x_T)}{p_2(x_t|x_{t-1},x_T)}$$

The derivation follows the same procedure as for the entropy derivation, with only the terms inside the logarithm changing. The algorithm (6) is of complexity $\mathcal{O}(N^3T)$, since the terms $\mathcal{D}_{t|t-1,T}(p_1\|p_2)$ can be precomputed also in $\mathcal{O}(N^3T)$.

C. Discussion

As stated in the introduction, the notion of a reciprocal three-point transition (1) arose as part of the solution to Schrödinger's original problem. It is known that in this case the KLD between any two Markov processes on a finite interval with the same three point dynamics (1), reduces to the KLD between their endpoint distributions. This is true for non-Markov RCs as well.

A special case worth consideration is the KLD between an RC and its *base process*. If we consider p_2 in expression (6) to represent the base process of the RC process p_1, the expression reduces to the KLD between the endpoint distributions $p_1(x_0,x_T)$ and $p_2(x_0,x_T)$, since both processes have the same internal dynamics, or three-point transitions (see e.g. [1]). As the KLD provides a measure of the difference between the dynamics of both processes, it may be suitable for studying whether there is a benefit in using the higher complexity reciprocal model when the base process is known. This was the case in [5], where partially observed sequences of an RC model were used to represent target trajectories hidden in noise and targets were tracked using both the RC and base process models. The KLD between two RC provides an alternative indicator of tracking performance between the competing models, compared to the empricial method used in [5].

As discussed in the introduction, the problem of state reduction via an information-theoretic approach has been studied in the MC case (e.g. [11], [12]). This problem is of interest in the RC case, especially since state estimation algorithms for hidden RC of higher computational complexity than standard hidden MC estimators; $\mathcal{O}(N^3T)$ compared with $\mathcal{O}(N^2T)$. The KLD algorithm for RC is a step towards an algorithm for state reduction.

V. Markov Approximations to RC Models

In this section we derive the exact analytical expressions for the MC which best approximates a non-Markov RC, in terms of minimising the KLD between the joint probability distributions of the processes. Henceforth when we refer to the 'closest' Markov approximation we mean in KL sense.

The first result is an expression for the closest time-inhomogeneous MC approximation to an RC, meaning the transition probabilities are time dependent. Following this we present the expression for the closest homogeneous MC approximation to an RC, which has a nice interpretation as a time average of the closest inhomogeneous MC. .

A. Inhomogeneous MC Approximation

Now let us consider the KL divergence between an RC with joint distribution $p(x_0,\dots x_T)$ and a stationary MC with joint distribution $q(x_0,\dots x_T)$.

$$\begin{aligned}\mathcal{D}(p\|q) &= \sum_{\mathcal{X}} p(\mathcal{X})\log\frac{p(\mathcal{X})}{q(\mathcal{X})} \\ &= \sum_{\mathcal{X}} p(\mathcal{X})\log p(\mathcal{X}) - \sum_{\mathcal{X}} p(\mathcal{X})\log q(\mathcal{X})\end{aligned}$$

The first term is simply the entropy of the RC as before, so all that remains is to calculate the cross entropy term. While the MB characterisation can also be used for a MC, we simply use the standard chain rule for Markov chains. Thus we obtain for the cross entropy term

$$\begin{aligned}H^{p,q}_{\mathcal{X}} &= -\sum_{\mathcal{X}} p(\mathcal{X})\log q(\mathcal{X}) \\ &= -\sum_{x_0,x_T} p(x_0,x_T)\log q(x_0) \\ &+ \sum_{x_0,x_T} p(x_0,x_T)\Bigg(\sum_{x_1} p(x_1|x_0,x_T)\log q(x_1|x_0) \\ &+ \sum_{t=1}^{T-2}\sum_{x_t} p(x_t|x_0,x_T)\sum_{x_{t+1}} p(x_{t+1}|x_t,x_T)\log q(x_{t+1}|x_t) \\ &+ \sum_{x_{T-1}} p(x_{T-1}|x_0,x_T)\log q(x_T|x_{T-1})\Bigg) \quad (7)\end{aligned}$$

We now formulate the Lagrangian to minimise this divergence with respect to the transitions of $q(x_0,\dots,x_t)$, which we denote as $M_{ij}(t)$. The constraint is simply that the $M_{ij}(t) = q(x_{t+1}=j|x_t=i)$ be valid Markov transitions, that is, that the probabilities are non-negative and sum to one.

$$\mathcal{L} = \mathcal{D}(p\|q) + \sum_{m,t}\lambda_{m,t}\Big(\sum_{n=1} M_{m,n}(t) - 1\Big).$$

Then

$$\frac{\partial\mathcal{L}}{\partial M_{ij}(t)} = \frac{\partial H^{p,q}_{\mathcal{X}}}{\partial M_{ij}(t)} + \lambda_{i,t}$$

From (7)

$$\frac{\partial\log q(x_{t+1}=j|x_t=i)}{\partial M_{ij}(t)} = \frac{\partial\log M_{ij}(t)}{\partial M_{ij}(t)} = \frac{1}{M_{ij}(t)}$$

Then, for $0 < t < T-1$, the cross entropy term in the minimisation step becomes

$$\begin{aligned}\frac{\partial H_{\mathcal{X}}^{p,q}}{\partial M_{ij}(t)} &= -\sum_{x_0,x_T} p(x_0,x_T)p(x_t^i|x_0,x_T)p(x_{t+1}^j|x_t^i,x_T)\frac{1}{M_{ij}(t)} \\ &= -\frac{1}{M_{ij}(t)}\sum_{x_T} p(x_{t+1}^j|x_t^i,x_T)\sum_{x_0} p(x_t^i,x_0,x_T) \\ &= -\frac{1}{M_{ij}(t)}\sum_{x_T} p(x_{t+1}^j|x_t^i,x_T)p(x_t^i,x_T) \\ &= -\frac{1}{M_{ij}(t)}p(x_{t+1}^j|x_t^i)p(x_t^i)\end{aligned}$$

Where we have used the notation x_t^i to denote $x_t = i$ for shorthand. We obtain the same expressions for $M_{ij}(t)$ at $t = 0$ and $t = T-1$. Setting the partial derivatives of the Lagrangian to zero, we have that

$$M_{ij}(t) = \frac{p(x_{t+1} = j|x_t = i)p(x_t = i)}{\lambda_{i,t}}$$

and by enforcing the constraint,

$$\lambda_{i,t} = p(x_t = i)$$

Therefore we find that

$$M_{ij}(t) = p(x_{t+1} = j|x_t = i), \tag{8}$$

which are the two point transitions of the RC, a result which makes intuitive sense. In the special case of a RC constructed with an endpoint distribution resulting in the process being Markov, then the derived expression will be its usual Markov transitions, meaning the divergence will be equal to zero.

B. Homogeneous MC Approximation

If we now consider the case of finding a homogeneous approximation to any RC, we can extend the analysis from above as follows. In this case $M_{ij}(t) = A_{ij},\ \forall\ t$, and with corresponding constraints on A_{ij}, the Lagrangian is now

$$\mathcal{L} = \mathcal{D}(p\|q) + \sum_m \lambda_m(\sum_{n=1} A_{m,n} - 1)$$

If we again take partial derivatives, and note that once more we are minimising only the cross entropy terms,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial A_{ij}} &= \frac{\partial \mathcal{D}(p\|q)}{\partial A_{ij}} + \frac{\partial \sum_m \lambda_m(\sum_{n=1} A_{m,n} - 1)}{\partial A_{ij}} \\ &= -\sum_{x_0,x_T} p(x_0,x_T)\sum_{t=1}^{T-1} p(x_t^i|x_0,x_T)p(x_{t+1}^j|x_t^i,x_T)\frac{1}{A_{ij}} \\ &\quad + \sum_{x_T} p(x_0^i,x_T)p(x_1^j|x_0^i,x_T)\frac{1}{A_{ij}} + \lambda_i \\ &= -\frac{1}{A_{ij}}\sum_{t=1}^{T-1} p(x_{t+1}^j|x_t^i)p(x_t^i) + p(x_1^j|x_0^i)p(x_0^i) + \lambda_i \\ &= -\frac{1}{A_{ij}}\sum_{t=0}^{T-1} p(x_{t+1}^j|x_t^i)p(x_t^i) + \lambda_i\end{aligned}$$

Again setting the partial derivatives of the Lagrangian to zero, we have that

$$A_{ij} = \frac{\sum_{t=0}^{T-1} p(x_{t+1} = j|x_t = i)p(x_t = i)}{\lambda_i},$$

and by evaluating the constraint, we have

$$\lambda_i = \sum_{t=0}^{T-1} p(x_t = i),$$

so we find that the Lagrange multipliers λ_i are a time-average of the marginals of state i. Explicitly,

$$A_{ij} = \sum_{t=0}^{T-1} p(x_{t+1} = j|x_t = i)\frac{p(x_t = i)}{\sum_{s=0}^{T-1} p(x_s = i)} \tag{9}$$

This result builds nicely on the previous result, since the homogeneous transitions probabilities are clearly a time average of the inhomogeneous transition probabilities $p(x_t + 1 = j|x_t = i)$, weighted by a time average of the probability of being in state $x_t = i$.

In summary, we have derived analytic expressions for lower complexity MC model approximations for an RC, based on a minimisation of the KLD. Possible applications arising from these results are mentioned in the following section.

VI. Conclusion and Future Work

This letter has presented an efficient algorithm for the computation of Kullback-Leibler divergence (KLD) between reciprocal chains (RC), including both Markov and non-Markov chains. The algorithm is of $\mathcal{O}(N^3T)$ complexity. Analytical expressions have been derived for Markov chain (MC) approximations to an RC via a minimisation of the KLD between the RC model and its MC approximation, both inhomogeneous and homogeneous.

Applications for state space reduction and improved modelling of targets using the KLD algorithm and associated information theoretic concepts have been proposed and merit further investigation. Important applications arising from the MC model approximations of RC are lower complexity suboptimal estimation for hidden reciprocal chains (HRC). As mentioned, the optimal HRC estimators are of $\mathcal{O}(N^3T)$ complexity, where as a HMC filter is of $\mathcal{O}(N^2T)$.

Acknowledgements

The support of the Defence Science and Technology Group, Australia, and the Data to Decisions CRC is gratefully acknowledged. We would also like to thank Riley Bruce-Doust for useful discussions.

References

[1] P. Dai Pra, "A stochastic control approach to reciprocal diffusion processes", *Appl. Math. Optim.*, vol. 23, pp. 313-329, 1991.

[2] A. Beghi, "On the relative entropy of discrete-time Markov processes with given end-point densities," *IEEE Transactions on Information Theory*, vol. 42, pp. 15291535, 1996.

[3] M. Pavon and F. Ticozzi, "Discrete-Time Classical and Quantum Markovian Evolutions: Maximum Entropy Problems on Path Space," *Journal of Mathematical Physics*, vol. 51, no. 4, 2010.

[4] T. T. Georgiou and M. Pavon, "Positive Contraction Mappings for Classical and Quantum Schrödinger systems," *Journal of Mathematical Physics*, vol. 56, no. 3, p. 033301, 2015.
[5] G. S. Stamatescu, L. B. White and R. Bruce-Doust, "Track extraction with Hidden Reciprocal Chain Models", submitted: arXiv:1605.04046
[6] L. B. White and F. Carravetta, "Optimal smoothing for finite state hidden reciprocal processes", *IEEE Trans. Automatic Control*, v. 56, no. 9, pp. 2156-2161, 2011.
[7] M. Fanaswala and V. Krishnamurthy, "Detection of Anomalous Trajectory Patterns in Target Tracking via Stochastic Context-Free Grammars and Reciprocal Process Models", *IEEE J. Sel. Topics in Signal Processing*, vol.7, no. 1, pp. 76-90, 2013.
[8] F. Carravetta, "Nearest-neighbour modelling of reciprocal chains," *Stoch. Int. J. Probab. Stoch. Processes*, vol. 80, pp. 525-584, 2008.
[9] L. B. White and F. Carravetta, "New Normalised Bayesian smoothers for signals modelled by non-causal compositions of reciprocal chains," *Proc. 2014 IEEE Workshop on Statistical Signal Processing (SSP)*, Gold Coast, Aus., 2014, pp. 205-208.
[10] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
[11] K. Deng, Y. Sun, P. G. Mehta and S. P. Meyn,"An information-theoretic framework to aggregate a Markov chain," *2009 American Control Conference*, St. Louis, MO, 2009, pp. 731-736.
[12] M. Vidyasagar, "Reduced-order modeling of Markov and hidden Markov processes via aggregation," *49th IEEE Conference on Decision and Control (CDC)*, Atlanta, GA, 2010, pp. 1810-1815.
[13] C.E Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.,* 27: pp. 379-423, 623-656, 1948.
[14] B. Jamison, "Reciprocal Processes", *Z. Warscheinlichkeitstheorie verw. Gebiete*, 30, pp. 65-86, 1974.