Identifying Sparse Low-Dimensional Structures in Markov Chains: A Nonnegative Matrix Factorization Approach

MAHSA GHASEMI, ABOLFAZL HASHEMI, HARIS VIKALO, AND UFUK TOPCU

AMERICAN CONTROL CONFERENCE
JULY 1, 2020

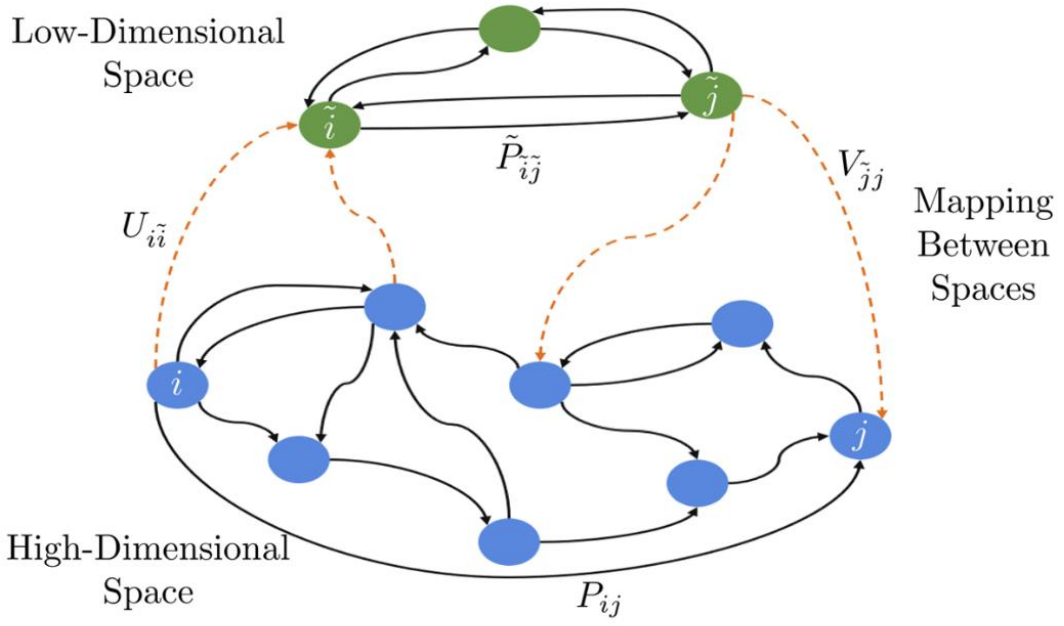
Model Reduction for Markov Chains

- Markov chains: a modeling framework for study of stochastic systems

- Applications in control, machine learning, and computational biology

- Large-scale models in practical settings

- Abstraction using structural properties
  - A nonnegative matrix factorization approach
  - Efficient solution using block coordinate gradient descent



Model Reduction for Markov chains

Markov Chains (MC)

An MC is a tuple $\mathcal{MC} = (S, \mu_{init}, P)$ where

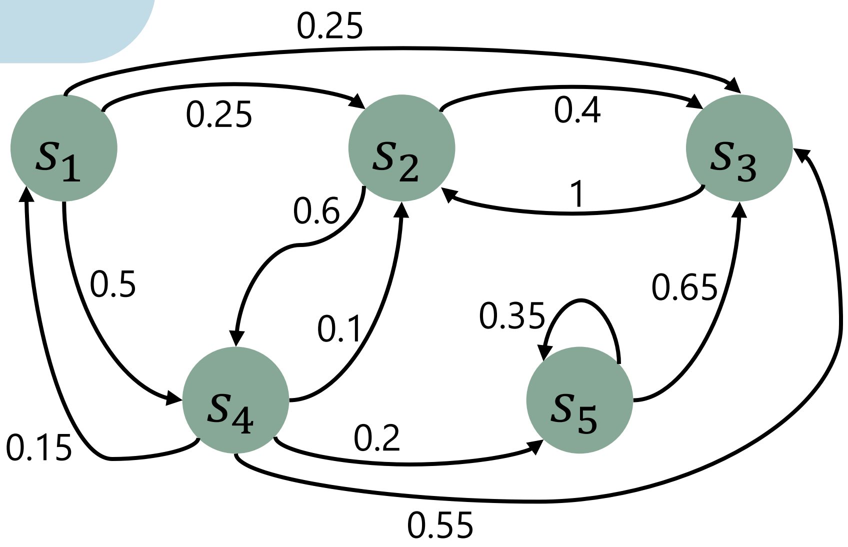
- S is a finite set of states with cardinality $|S| = n$
- μ_{init} is an initial distribution over the states
- $P : S \times S \to [0,1] \subseteq \mathbb{R}$ is a probability transition function such that for all $s \in S, \sum_{s' \in S} P(s, s') = 1$

A finite path is a finite sequence of states $\sigma = x_0 x_1 x_2 \dots x_T$, such that

- x_0 is in the support of μ_{init}, and
- $P(x_{t-1}, x_t) > 0$ for all $t \in \{1, 2, \dots, T\}$.

The probability of observing σ is

$$Pr(\sigma) = \mu_{init}(x_0) \prod_{t=1}^{T} P(x_{t-1}, x_t).$$



An example MC

Characterization of Low-Dimensional Structure

Nonnegative rank of a Markov chain: Smallest $k \in \mathbb{N}$ such that

$$\Pr(X_{t+1}|X_t) = \sum_{l=1}^{k} \underbrace{f_l(X_t)}_{\text{left Markov features}} \underbrace{g_l(X_{t+1})}_{\text{right Markov features}},$$

where $f_1, f_2, \ldots, f_k$ and $g_1, g_2, \ldots, g_k$ are mappings from S to $\mathbb{R}_+$.

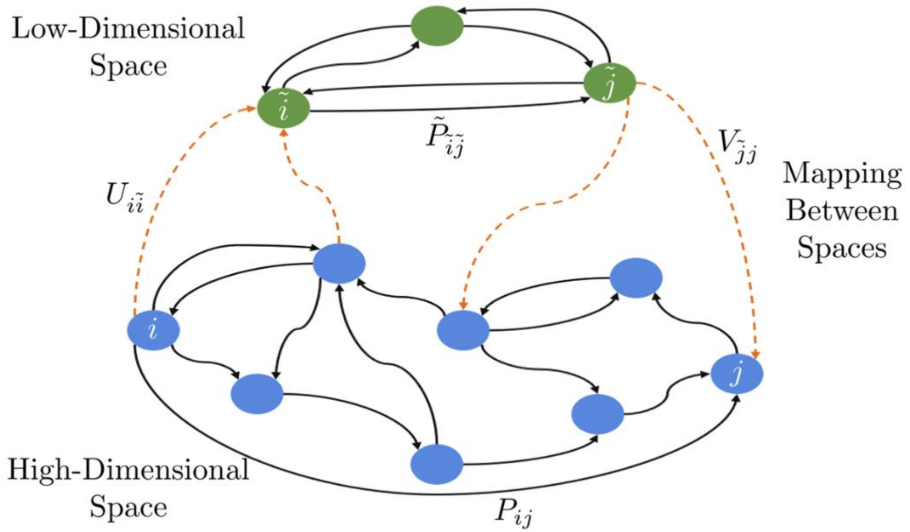
Goal: Given that a Markov chain with n states has a nonnegative rank of $k \ll n$, design an algorithm to find a low-dimensional representation, i.e., the features.

Formulation as Matrix Factorization

Proposition:[1] The nonnegative rank of a Markov chain is k if and only if there exists $U \in \mathbb{R}_+^{n\times k}$, $\tilde{P} \in \mathbb{R}_+^{k\times k}$, $V \in \mathbb{R}_+^{k\times n}$ such that

$$P = U\tilde{P}V,$$

where U, $\tilde{P}$, and V are stochastic matrices.



Problem Formulation: Given a Markov chain $\mathcal{MC} = (S, \mu_{init}, P)$, find a kernel space and kernel transition, denoted by $(\tilde{S}, \tilde{P})$, along with sparse mappings (U, V) such that the following decomposition property holds:

$$P = U\tilde{P}V.$$

Efficient Multi-Step Transition

Probability of going from state s_i at time step t to state s_j, in m time steps, is

$$\Pr(X_{t+m} = s_j | X_t = s_i) = p_{ij}^{(m)}, \text{ where } p_{ij}^{(m)} = [P^m]_{ij}.$$

Assume a perfect low-rank decomposition $P = U\tilde{P}V$ and let $K = VU\tilde{P}$. Then,

$$\Pr(X_{t+m} | X_t) = \sum_{l_1=1}^{k} \sum_{l_2=1}^{k} U_{X_t, l_1} [\tilde{P} K^{m-1}]_{l_1 l_2} V_{l_2, X_{t+m}}.$$

→ Reducing the computational complexity from $\mathcal{O}(mn^2)$ to $\mathcal{O}(mk^2)$.

Matrix Factorization as an Optimization Task

$$\min_{U\geq 0, \tilde{P}\geq 0, V\geq 0} \mathcal{D}(P, U\tilde{P}V)$$

$$\text{s.t.} \quad \sum_{j=1}^{k} U_{ij} = 1, \quad \|u_i\|_0 \leq s_i^{(u)}, \forall i \in [n],$$

$$\sum_{j=1}^{k} \tilde{P}_{\ell j} = 1, \quad \forall \ell \in [k],$$

$$\sum_{j=1}^{n} V_{\ell j} = 1, \quad \|v_\ell\|_0 \leq s_\ell^{(v)}, \forall \ell \in [k].$$

sparsity constraints

stochastic matrix

$$\min_{U\geq 0, \tilde{P}\geq 0, V\geq 0} \frac{1}{2}\|P - U\tilde{P}V\|_F^2 + \lambda_u \|U\|_1 + \lambda_v \|V\|_1$$

$$\text{s.t.} \quad U\mathbf{1} = \mathbf{1}, \quad \tilde{P}\mathbf{1} = \mathbf{1}, \quad V\mathbf{1} = \mathbf{1}.$$

regularization parameters

promoting sparsity

Block Coordinate Gradient Descent (BCGD)

- **Input parameters:** regularization parameters λ_u, λ_v, step sizes $\alpha_t, \beta_t, \gamma_t$
- Initialize U_0 randomly
- For $t = 0, 1, 2, \ldots, T-1$, iteratively perform:
  - $U_{t+1} \longleftarrow$ Given $(U_t, \tilde{P}_t, V_t)$, optimize with respect to U
  - $\tilde{P}_{t+1} \longleftarrow$ Given $(U_{t+1}, \tilde{P}_t, V_t)$, optimize with respect to $\tilde{P}$
  - $V_{t+1} \longleftarrow$ Given $(U_{t+1}, \tilde{P}_{t+1}, V_t)$, optimize with respect to V

$$\boxed{\begin{aligned} &\bullet\ \nabla f(V_t) = -\tilde{P}_{t+1}^\top U_{t+1}^\top (P - U_{t+1}\tilde{P}_{t+1}V_t) \\ &\bullet\ V_{t+\frac{1}{2}} = V_t - \gamma_t \nabla f(V_t) \\ &\bullet\ V_{t+1} = \Pi_{\triangle_n}\left(\mathcal{T}_{\frac{\lambda_v}{2}}(V_{t+\frac{1}{2}})\right) \end{aligned}}$$

projection to simplex

shrinkage-thresholding operator

Convergence Analysis and Computational Complexity

Theorem: If the step sizes are selected according to:

$$\alpha_t = \frac{C_1 \|\nabla f(U_t)\|_F^2}{\|\nabla f(U_t)\tilde{P}_t V_t\|_F^2}, \qquad \beta_t = \frac{C_2 \|\nabla f(V_t)\|_F^2}{\|U_{t+1}\nabla f(\tilde{P}_t)V_t\|_F^2},$$

$$\gamma_t = \frac{C_3 \|\nabla f(\tilde{P}_t)\|_F^2}{\|U_{t+1}\tilde{P}_{t+1}\nabla f(V_t)\|_F^2}, \qquad C_1, C_2, C_3 \in (0, 2),$$

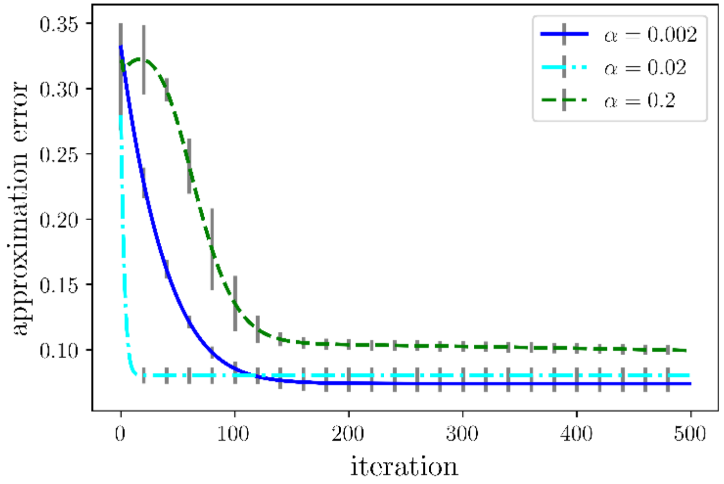
then, BCGD converges to a stationary point.

Complexity: BCGD algorithm requires $\mathcal{O}(nkT)$ computations.

Effect of Step Size on Convergence

Setting:

- A transition matrix of size 100 × 100 with rank 25
- 500 iterations of BCGD
- 10 independent runs for each instance



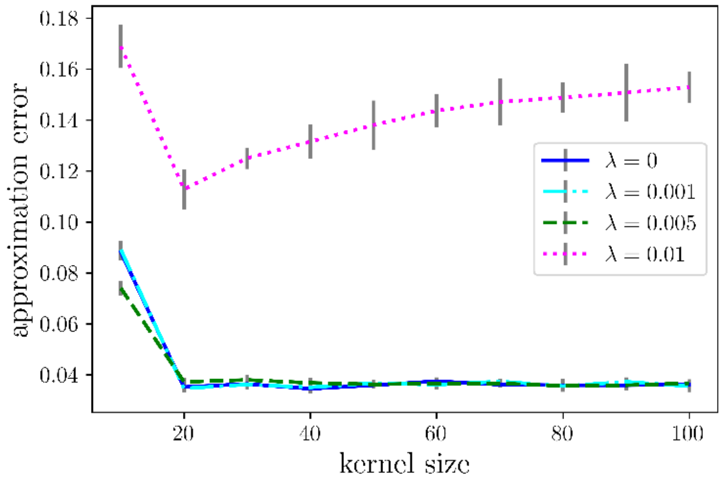
Results:

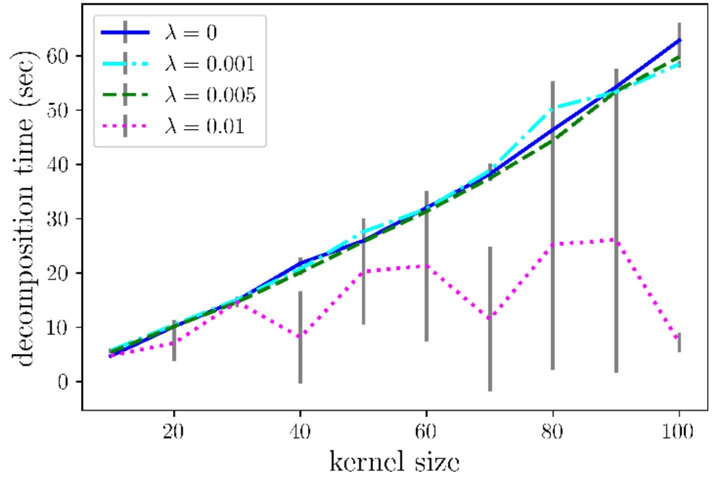
- Lower approximation error for smaller step sizes
- Algorithm diverging for step sizes over 0.2

Effect of Regularization Parameter on Performance

Results:

- Relation between approximation error and the size of the kernel transition
- Trade-off between lower approximation error and higher sparsity of the mappings
- Linearity of the running time with respect to the kernel size
- Negligible effect of regularization on the running time





Conclusion and Future Directions

Conclusion:

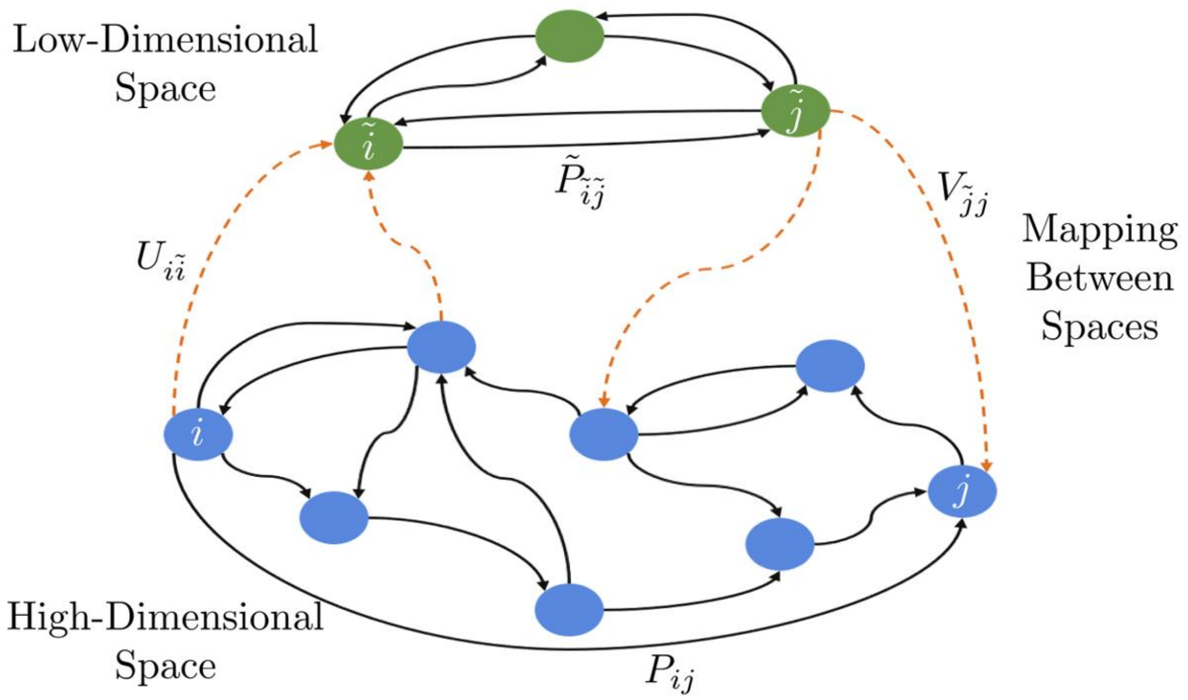
- Proposed a nonnegative matrix factorization formulation for learning sparse low-dimensional structures in Markov chains

- Developed an efficient iterative scheme based on block coordinate gradient descent

Future Directions:

- Extending the proposed formulation to model reduction of Markov decision processes

- Evaluating the abstract representation in terms of the performance in different downstream analyses



Thank you!

Identifying Sparse Low-Dimensional Structures in Markov Chains: A Nonnegative Matrix Factorization Approach

Mahsa Ghasemi, Abolfazl Hashemi, Haris Vikalo, and Ufuk Topcu