Information-theoretic coordinate subset and partition selection of multivariate Markov chains via submodular optimization

Zheyuan Lai[1]

Supervisor: Prof. Michael C. H. Choi[1]

[1]Department of Statistics and Data Science, National University of Singapore

arXiv preprint: https://arxiv.org/abs/2503.23340

Introduction

- **Setup**: d-dimensional Markov chains on product spaces: $\mathscr{X} = \mathscr{X}^{(1)} \times \ldots \times \mathscr{X}^{(d)}$
- **Subset selection**: Find a lower-dimensional coordinate subset that *preserves maximal information* or *stays closest to equilibrium*
- **Partition selection**: Find an optimal partition of coordinates that the factorized chain is closest to the original chain

Preliminaries

Information theory of Markov chains (Choi et al., 2024)

- Discrete time, discrete state, π-stationary Markov chain
- Leave-S-out transition matrix: $P_\pi^{(-S)}(x^{(-S)}, y^{(-S)})$ (Keep-S-in: $P_\pi^{(S)}(x^{(S)}, y^{(S)})$)
- Shannon's entropy: $H(\pi) = -\sum_{x \in \mathscr{X}} \pi(x) \ln \pi(x)$
- Entropy rate of P: $H(P) = -\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{X}} \pi(x) P(x,y) \ln P(x,y)$
- KL divergence from L to M: $D_{\mathrm{KL}}^{\pi}(M \| L) = \sum_{x \in \mathscr{X}} \pi(x) \sum_{y \in \mathscr{X}} M(x,y) \ln \frac{M(x,y)}{L(x,y)}$

Submodularity and k-submodularity

- Submodularity: "diminishing gains" $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$
- k-submodularity: generalization to multivariable $f(\mathbf{S}) + f(\mathbf{T}) \geq f(\mathbf{S} \sqcup \mathbf{T}) + f(\mathbf{S} \sqcap \mathbf{T})$
- k-submodularity $=$ pairwise monotonicity $+$ orthant submodularity
  - pairwise monotonicity: $\Delta_{e,i} f(\mathbf{S}) + \Delta_{e,j} f(\mathbf{S}) \geq 0$
  - orthant submodularity: $\Delta_{e,i} f(\mathbf{S}) \geq \Delta_{e,i} f(\mathbf{T}), \mathbf{S} \preceq \mathbf{T}$
- Sum of submodular mono. incr. $\Rightarrow$ k-submodular
- Transformations into monotonicity:
  - Non-monotone submodular $\Rightarrow$ Monotone submodular minus modular
  - Orthant submodular $\Rightarrow$ k-submodular minus modular
- Submodular structure in Markov chains (Choi et al., 2024)
  - $S \mapsto H(P^{(S)})$ is submodular but non-monotone
  - $S \mapsto D_{\mathrm{KL}}(P \| P^{(S)} \otimes P^{(-S)})$ is submodular and symmetric but non-monotone

Submodular optimization under cardinality constraint

- Greedy algorithm (Nemhauser et al., 1978):
  - $(1 - e^{-1})$-approximation guarantee for monotone submodular objective
  - no theoretical guarantee for non-monotone submodular objective
- Distorted greedy algorithm (Harshaw et al., 2019): theoretical lower bound for maximizing a monotone submodular function minus a modular function

Algorithm 2 **Distorted greedy algorithm for maximizing the difference between a monotonically non-decreasing submodular function and a modular function**

Require: monotonically non-decreasing submodular g with $g(\emptyset) \geq 0$, non-negative modular c, cardinality m, ground set U

```
Initialize S_0 ← ∅
for i = 0 to m − 1 do
    e_i ← argmax_{e∈U} { (1 − 1/m)^{m−(i+1)} (g(S_i ∪ {e}) − g(S_i)) − c({e}) }
    if (1 − 1/m)^{m−(i+1)} (g(S_i ∪ {e_i}) − g(S_i)) − c({e_i}) > 0 then
        S_{i+1} ← S_i ∪ {e_i}
    else
        S_{i+1} ← S_i
    end if
end for
Output: S_m.
```

- Generalized distorted greedy algorithm (this paper): theoretical lower bound for maximizing a k-submodular function minus a modular function

Algorithm 3 **Generalized distorted greedy algorithm for maximizing the difference of k-submodular function and a modular function**

Require: k-submodular monotonically non-decreasing g with $g(\emptyset) \geq 0$, non-negative modular c with $c(\emptyset) = 0$, cardinality m, ground set U, $\mathbf{V} = (V_1, \ldots, V_k) \in (k+1)^U$.

```
Initialize S_0 = (S_{0,1}, ..., S_{0,k}) ← ∅
for i = 0 to m − 1 do
    (j*, e*) ← argmax_{j∈[[k]], e∈V_j \ S_{i,j}} { (1 − 1/m)^{m−(i+1)} Δ_{e,j} g(S_i) − c({e}) }
    if (1 − 1/m)^{m−(i+1)} Δ_{e*,j*} g(S_i) − c({e*}) > 0 then
        S_{i+1,j*} ← S_{i,j*} ∪ {e*}
    else
        S_{i+1,j*} ← S_{i,j*}
    end if
    for l ≠ j* do
        S_{i+1,l} ← S_{i,l}
    end for
end for
Output: S_m = (S_{m,1}, ..., S_{m,k}).
```

Subset selection

Submodular maximization of entropy rate

- Motivation: find the coordinate subset with the most information
- Problem setup: $\max_{|S| \leq m} H(P^{(S)})$, non-monotone submodular maximization
- Apply distorted greedy algorithm with theoretical lower bound:
  - product form π: $S \mapsto H(P^{(S)}) = H(\pi^{(S)} \boxtimes P^{(S)}) - H(\pi^{(S)})$
  - general π: monotone submodular g and modular c

$$g(S) = H(P^{(S)}) - \beta + \sum_{e \in S} (H(P^{(-e)}) - H(P))$$

$$c(S) = -\beta + \sum_{e \in S} (D(P \| P^{(e)} \otimes P^{(-e)}) - H(P^{(e)}))$$

$$\beta \leq -\sum_{i=1}^{d} \log |\mathscr{X}^{(i)}|$$

Supermodular minimization of distance to stationarity

- Motivation: find the coordinate subset that is closest to equilibrium $\Rightarrow$ design of efficient discrete-space MCMC samplers
- Problem setup: $\max_{|S| \leq m} -D_{\mathrm{KL}}(P^{(S)} \| \Pi^{(S)})$, submodular with product form π
- Apply distorted greedy algorithm with theoretical lower bound

$$g(S) = -D_{\mathrm{KL}}(P^{(S)} \| \Pi^{(S)}) + \sum_{e \in S} (D_{\mathrm{KL}}(P \| P^{(e)} \otimes P^{(-e)}) + D_{\mathrm{KL}}(P^{(e)} \| \Pi^{(e)}))$$

$$c(S) = \sum_{e \in S} (D_{\mathrm{KL}}(P \| P^{(e)} \otimes P^{(-e)}) + D_{\mathrm{KL}}(P^{(e)} \| \Pi^{(e)}))$$

Partition selection

k-submodular maximization of distance to factorizability

- Motivation: select a coordinate partition $\mathbf{S}$ to determine a factorized chain being the "furthest" from the original chain P in terms of KL divergence
- Problem setup: $\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} f(\mathbf{S}) = D_{\mathrm{KL}}(P \| P^{(S_1)} \otimes \ldots \otimes P^{(S_k)} \otimes P^{(-\cup_{i=1}^{k} S_i)})$ orthant submodular but not pairwise monotone
- Apply generalized distorted greedy algorithm with guaranteed lower bound

$$\begin{aligned} g(\mathbf{S}) = f(\mathbf{S}) - \beta + \sum_{i=1}^{k} \sum_{e \in S_i} \Big[& D_{\mathrm{KL}}(P^{(V_i)} \| P^{(V_i \setminus \{e\})} \otimes P^{(e)}) \\ & - D_{\mathrm{KL}}(P^{(-\mathrm{supp}(\mathbf{V}) \setminus \{e\})} \| P^{(-\mathrm{supp}(\mathbf{V}))} \otimes P^{(e)}) \Big] \end{aligned}$$

$$\begin{aligned} c(\mathbf{S}) = -\beta + \sum_{i=1}^{k} \sum_{e \in S_i} \Big[& D_{\mathrm{KL}}(P^{(V_i)} \| P^{(V_i \setminus \{e\})} \otimes P^{(e)}) \\ & - D_{\mathrm{KL}}(P^{(-\mathrm{supp}(\mathbf{V}) \setminus \{e\})} \| P^{(-\mathrm{supp}(\mathbf{V}))} \otimes P^{(e)}) \Big] \end{aligned}$$

$$\beta \leq -\sum_{i=1}^{k} \sum_{e \in V_i} \Big(H(P^{(-\mathrm{supp}(\mathbf{V}) \setminus \{e\})}) + H(P^{(e)}) \Big)$$

Numerical experiments

- d-dimensional Markov chain associated with the Curie–Weiss model and the Bernoulli–Laplace level model
- Toy experiment on MCMC with Curie–Weiss model
- Experiments on the algorithms introduced throughout the paper

References

- M. C. H. Choi, Y. Wang, and G. Wolfer. Geometry and factorization of multivariate Markov chains with applications to MCMC acceleration. arXiv:2404.12589.
- C. Harshaw, M. Feldman, J. Ward, and A. Karbasi. Submodular maximization beyond non-negativity: Guarantees, fast algorithms, and applications. ICML 2019.
- G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. I. *Math. Programming*, 14(3):265–294, 1978.