Information-theoretic analysis of multivariate Markov chains: subset selection and minimax factorization

Zheyuan Lai*[1]

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

Supervisor: Professor Michael Choi†

Abstract

We study multivariate Markov chains on product state spaces through an information-theoretic lens. On the one hand, we study the problem of optimally projecting the transition matrix of a finite ergodic multivariate Markov chain onto a lower-dimensional state space. Specifically, we seek to construct a projected Markov chain that optimizes various information-theoretic criteria under cardinality constraints. We formulate these tasks as best subset selection problems over multivariate Markov chains and leverage the (k-)submodular (or (k-)supermodular) structure of the objective functions to develop efficient greedy-based algorithms with theoretical guarantees. On the other hand, we study the minimax factorization problem of multivariate Markov chains, where we seek to find the optimal factorizable transition matrix that minimizes the maximum information-theoretic distance to the transition matrices of the original family of Markov chains. We show that this problem can be formulated as a convex optimization problem through strong duality and provide provable algorithms. Finally, we present numerical experiments associated with Curie-Weiss and Bernoulli-Laplace models to demonstrate the effectiveness of our proposed methods.

Contents

*Email: zheyuan_lai@u.nus.edu
†Email: mchchoi@nus.edu.sg

1 Introduction

Motivation. Multivariate Markov chains on product spaces $\mathcal{X} = \mathcal{X}^{(1)} \times \ldots \times \mathcal{X}^{(d)}$ with $d \in \mathbb{N}$ arise naturally throughout stochastic modeling, Markov chain Monte Carlo (MCMC), and interacting particle systems. In high dimensions when d is large, it is natural—both for analysis and for algorithm design—to (i) propose a subset Markov chain which preserves the most information or is closest to equilibrium, and (ii) approximate a complex transition matrix P by a simpler model that *factorizes* across groups of coordinates. This paper develops an information-theoretic framework, associated structure theorems, and algorithms for subset selection of a single Markov chain and minimax factorization of a family of Markov chains.

Related works. We build on three lines of work: information projection for Markov chains, minimax information aggregation, and (robust) submodular optimization over partitions. Choi et al. (2024) view factorization as minimizing the KL divergence between an original chain and the set of factorizable chains; Lacker (2025) introduces an independent projection for diffusion processes via relative entropy minimization over product measures; and Geiger and Temmel (2014) study lumping of Markov chains from combinatorial and information-theoretic perspectives. For minimax information aggregation, Haussler (1997); Gushchin and Zhdanov (2006) analyze minimax optimization under KL and general f-divergences for probability measures, while Hafez-Kolahi et al. (2022) cast minimax excess risk as a zero-sum game between a learner and Nature. For submodular optimization over partitions, Nemhauser et al. (1978) and Ward and Živnỳ (2016) give greedy algorithms with guarantees for submodular and k-submodular partition functions; Orlin et al. (2018) address robust submodular optimization via bilevel formulations; Bogunovic et al. (2017) propose algorithms for non-uniform partitions; and Staib and Jegelka (2019) leverage continuous submodularity for robust budget allocation.

Structure. The remainder of the paper is organized as follows. Section 2 fixes notation and introduces background knowledge: Section 2.1 summarizes key information-theoretic results in Markov chain theory; Section 2.2 reviews submodularity and k-submodularity and discusses submodular functions arising in the information-theoretic study of multivariate Markov chains; Section 2.3 covers submodular optimization algorithms with guarantees; and Section 2.4 gives examples of multivariate Markov chains on product state spaces. We then divide the discussion into two parts. Part I studies optimization problems concerning entropy rate (Section 3), distance to factorizability (Section 4), distance to independence (Section 5), distance to stationarity (Section 6), and distance to factorizability over a fixed set (Section 7), with numerical illustrations in Section 8. Part II reformulates minimax factorization as a concave maximization problem via strong duality (Section 9) and interprets it as a two-player zero-sum game (Section 10); we then present a projected subgradient method (Section 11) and a subgradient–greedy algorithm (Section 12) to solve the minimax and max–min–max problems, followed by numerical experiments in Section 13.

2 Preliminaries

2.1 Some information-theoretic properties of multivariate Markov chains

Throughout this paper, we consider a finite d-dimensional state space described by $\mathcal{X} = \mathcal{X}^{(1)} \times \ldots \times \mathcal{X}^{(d)}$. We write $[\![d]\!] = \{1, 2, \ldots, d\}$. For $S \subseteq [\![d]\!]$, we write $\mathcal{X}^{(S)} = \times_{i \in S} \mathcal{X}^{(i)}$. We denote by $\mathcal{L}(\mathcal{X})$ to be the set of transition matrices on $\mathcal{X}$, and $\mathcal{P}(\mathcal{X}) = \{\pi \mid \min_{x \in \mathcal{X}} \pi(x) > 0\}$ to be the set of probability masses with support on $\mathcal{X}$. Let $\pi \in \mathcal{P}(\mathcal{X})$ be any given probability distribution, and denote $\mathcal{L}(\pi) \subseteq \mathcal{L}(\mathcal{X})$ as the set of π-reversible transition matrices on $\mathcal{X}$, where a transition matrix $P \in \mathcal{L}(\mathcal{X})$ is said to be π-reversible if the detailed balance condition holds such that $\pi(x)P(x, y) = \pi(y)P(y, x)$ for all $x, y \in \mathcal{X}$. Additionally, we say that $P \in \mathcal{L}(\mathcal{X})$ is π-stationary if it satisfies $\pi = \pi P$.

We now recall the definition of the tensor product of transition matrices and probability masses, see e.g. Exercise 12.6 of (Levin and Peres, 2017). Define, for $M_l \in \mathcal{L}(\mathcal{X}^{(l)})$, $\pi_l \in \mathcal{P}(\mathcal{X}^{(l)})$, $x^l, y^l \in \mathcal{X}^{(l)}$ for $l \in \{i, j\}, i \neq j \in [\![d]\!]$,

$$(M_i \otimes M_j)((x^i, x^j), (y^i, y^j)) := M_i(x^i, y^i) M_j(x^j, y^j),$$
$$(\pi_i \otimes \pi_j)(x^i, x^j) := \pi_i(x^i)\pi_j(x^j).$$

A transition matrix $P \in \mathcal{L}(\mathcal{X})$ is said to be in a product form if there exists $M_i \in \mathcal{L}(\mathcal{X}^{(i)})$ for $i \in [\![d]\!]$ such that $P = \otimes_{i=1}^{d} M_i$ can be expressed as a d-fold tensor product. A probability mass π is said to be in a product form if there exists $\pi_i \in \mathcal{P}(\mathcal{X}^{(i)})$ such that $\pi = \otimes_{i=1}^{d} \pi_i$.

We then recall the definition of leave-S-out and keep-S-in transition matrices of a given transition matrix P, see Section 2.2 of (Choi et al., 2024). Let $\pi \in P(\mathcal{X})$, $P \in \mathcal{L}(\mathcal{X})$, and $S \subseteq [\![d]\!]$. For any $(x^{(-S)}, y^{(-S)}) \in \mathcal{X}^{(-S)} \times \mathcal{X}^{(-S)}$, we define the **leave-$S$-out** transition matrix to be $P_\pi^{(-S)}$ with entries given by

$$P_\pi^{(-S)}(x^{(-S)}, y^{(-S)}) := \frac{\sum_{(x^{(S)}, y^{(S)}) \in \mathcal{X}^{(S)} \times \mathcal{X}^{(S)}} \pi(x^1, \ldots, x^d) P((x^1, \ldots, x^d), (y^1, \ldots, y^d))}{\sum_{x^{(S)} \in \mathcal{X}^{(S)}} \pi(x^1, \ldots, x^d)}.$$

The **keep-S-in** transition matrix of P with respect to π is

$$P_\pi^{(S)} := P_\pi^{(-[\![d]\!] \setminus S)} \in \mathcal{L}(\mathcal{X}^{(S)}).$$

In the special case of $S = \{i\}$ for $i \in [\![d]\!]$, we write

$$P_\pi^{(-i)} = P_\pi^{(-\{i\})}, \quad P_\pi^{(i)} = P_\pi^{(\{i\})}.$$

When P is π-stationary, we omit the subscript π and write directly $P^{(-S)}, P^{(S)}$. We also apply the convention of $P^{(\emptyset)} = P^{(-[\![d]\!])} = 1$.

We proceed to recall the Shannon entropy of a probability distribution and the entropy rate of the transition matrix, see Section 1 of (Polyanskiy and Wu, 2025). For $\pi \in \mathcal{P}(\mathcal{X})$, its **Shannon entropy** is defined as

$$H(\pi) := -\sum_{x \in \mathcal{X}} \pi(x) \ln \pi(x),$$

where the standard convention of $0 \ln 0 := 0$ applies. For π-stationary $P \in \mathcal{L}(\mathcal{X})$, the **entropy rate** of P is defined as

$$H(P) := -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \pi(x) P(x, y) \ln P(x, y),$$

where the standard convention of $0 \ln 0 := 0$ applies.

We shall also recall the definition of **KL divergence** between Markov chains (Definition 2.1 of (Choi et al., 2024)) and the distance to independence (Definition 2.2 of (Choi et al., 2024)). For given $\pi \in \mathcal{P}(\mathcal{X})$ and transition matrices $M, L \in \mathcal{L}(\mathcal{X})$, we define the KL divergence from L to M with respect to π as

$$D_{\mathrm{KL}}^{\pi}(M \| L) := \sum_{x \in \mathcal{X}} \pi(x) \sum_{y \in \mathcal{X}} M(x, y) \ln \left(\frac{M(x, y)}{L(x, y)} \right),$$

where the convention of $0 \ln \frac{0}{a} := 0$ applies for $a \in [0, 1]$. Note that π need not be the stationary distribution of L or M. In particular, when M, L are assumed to be π-stationary, we write

$$D(M \| L) := D_{\mathrm{KL}}^{\pi}(M \| L),$$

which can be interpreted as the KL divergence rate from L to M. Given $P \in \mathcal{L}(\mathcal{X})$, we define the **distance to independence** of P with respect to D_{KL}^{π} to be

$$\mathbb{I}^{\pi}(P) := \min_{L_i \in \mathcal{L}(\mathcal{X}^{(i)}),\ \forall i \in [\![d]\!]} D_{\mathrm{KL}}^{\pi}(P \| \otimes_{i=1}^{d} L_i) = D_{\mathrm{KL}}^{\pi}(P \| \otimes_{i=1}^{d} P_\pi^{(i)}).$$

We write

$$\mathbb{I}(P) = \mathbb{I}^{\pi}(P)$$

if P is π-stationary.

We recall the partition lemma for KL divergence of Markov chains (see Theorem 2.4 of (Choi et al., 2024)).

Theorem 2.1 (Partition lemma). *Let $\pi \in \mathcal{P}(\mathcal{X})$, $P, L \in \mathcal{L}(\mathcal{X})$ and suppose $S \subseteq [\![d]\!]$, we have:*

$$D_{\mathrm{KL}}^{\pi}(P \| L) \geq D_{\mathrm{KL}}^{\pi^{(S)}}(P^{(S)} \| L^{(S)}).$$

We then define the **averaging operation** $\overline{P}(\mathbf{w})$ of a transition probability matrix P. We define $\mathcal{S}_n$ as the n-probability-simplex such that

$$\mathcal{S}_n = \left\{ \mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}_+^n;\ \sum_{i=1}^n w_i = 1 \right\}.$$

Given a set of π-stationary transition probability matrices $\mathcal{B} = \{P_1, \ldots, P_n\}$, we define the transition probability matrix weighted by $\mathbf{w} = (w_1, \ldots, w_n) \in \mathcal{S}_n$ as $\overline{P}(\mathbf{w})$ by

$$\overline{P} = \overline{P}(\mathbf{w}) := \sum_{i=1}^n w_i P_i.$$

We see that $\overline{P}$ is also π-stationary because

$$\pi\overline{P} = \pi\left(\sum_{i=1}^n w_i P_i\right) = \sum_{i=1}^n w_i(\pi P_i) = \sum_{i=1}^n w_i \pi = \pi.$$

We project each P_i onto $S \in 2^{[\![d]\!]}$ and denote the weighted projection as

$$\overline{P}(S, \mathbf{w}) := \sum_{i=1}^n w_i P_i^{(S)}.$$

As a result, we have

$$\overline{P}^{(S)} = \left(\sum_{i=1}^n w_i P_i\right)^{(S)} = \sum_{i=1}^n w_i P_i^{(S)} = \overline{P}(S, \mathbf{w}),$$

which means that the averaging operation commutes with the projection operation.

We then prove a Pythagorean identity related to the averaging operation and the KL divergence of transition matrices.

Lemma 2.2. *For given $\mathbf{w} \in \mathcal{S}_n$, $\pi \in \mathcal{P}(\mathcal{X})$, $P_i, Q \in \mathcal{L}(\mathcal{X})$ for $i \in [\![n]\!]$ where P_i are all π-stationary, we choose mutually disjoint sets $S_1, \ldots, S_m$ with $\sqcup_{i=1}^m S_i = [\![d]\!]$, and the following identity holds:*

$$\sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^m Q^{(S_j)}) = \sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^m \overline{P}^{(S_j)}) + \sum_{j=1}^m D_{\mathrm{KL}}^{\pi^{(S_j)}}(\overline{P}^{(S_j)} \| Q^{(S_j)}). \tag{1}$$

In particular, we have the following minimization result:

$$\min_{Q;\ Q = \otimes_{j=1}^m Q^{(S_j)}} \sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| Q) = \sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^m \overline{P}^{(S_j)}).$$

Proof. Inspired by Theorem 2.22 of (Choi et al., 2024), we note that

$$\begin{aligned}
&\sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^m Q^{(S_j)}) \\
&= \sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^m \overline{P}^{(S_j)}) + \sum_{i=1}^n w_i \sum_{x,y} \pi(x) P_i(x,y) \ln \frac{\otimes_{j=1}^m \overline{P}^{(S_j)}(x,y)}{\otimes_{j=1}^m Q^{(S_j)}(x,y)} \\
&= \sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^m \overline{P}^{(S_j)}) + \sum_{j=1}^m \sum_{i=1}^n w_i \sum_{x^{(S_j)}, y^{(S_j)}} \pi^{(S_j)}(x^{(S_j)}) P_i^{(S_j)}(x^{(S_j)}, y^{(S_j)}) \ln \frac{\overline{P}^{(S_j)}(x^{(S_j)}, y^{(S_j)})}{Q^{(S_j)}(x^{(S_j)}, y^{(S_j)})} \\
&= \sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^m \overline{P}^{(S_j)}) + \sum_{j=1}^m D_{\mathrm{KL}}^{\pi^{(S_j)}}(\overline{P}^{(S_j)} \| Q^{(S_j)}),
\end{aligned}$$

where the last equality comes from the fact that the averaging and projection operation commutes. □

As a corollary, in the special case of $m = 2$ with $S_1 = S$, $S_2 = [\![d]\!]\backslash S$, we see that

Corollary 2.3. *For given* $\mathbf{w} \in \mathcal{S}_n$, $\pi \in \mathcal{P}(\mathcal{X})$, $P_i, Q \in \mathcal{L}(\mathcal{X})$ *for* $i \in [\![n]\!]$ *where* P_i *are all* π*-stationary,* $S \in 2^{[\![d]\!]}$*, the following identity holds:*

$$\sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| Q^{(S)} \otimes Q^{(-S)}) = \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \overline{P}^{(S)} \otimes \overline{P}^{(-S)}) + D_{\mathrm{KL}}^{\pi^{(S)}}(\overline{P}^{(S)} \| Q^{(S)}) + D_{\mathrm{KL}}^{\pi^{(-S)}}(\overline{P}^{(-S)} \| Q^{(-S)}). \tag{2}$$

In particular, we have the following minimization result:

$$\min_{Q;\ Q = Q^{(S)} \otimes Q^{(-S)}} \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| Q) = \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \overline{P}^{(S)} \otimes \overline{P}^{(-S)}).$$

2.2 Background and examples of submodular functions

We first recall the definition of a submodular function (Ward and Živnỳ, 2016). Given a finite nonempty ground set U, a set function $f : 2^U \to \mathbb{R}$ defined on subsets of U is called **submodular** if for all $S, T \subseteq U$,

$$f(S) + f(T) \geq f(S \cap T) + f(S \cup T).$$

f is said to be **supermodular** if $-f$ is submodular, and f is said to be **modular** if f is both submodular and supermodular.

Next, we recall a result that states the complement of a submodular function is still submodular:

Lemma 2.4. *If* $S \mapsto f(S)$ *is submodular, then* $S \mapsto f(U\backslash S)$ *is submodular.*

Proof. We choose $S \subseteq T \subseteq U$ and $e \in U\backslash T$, then

$$\begin{aligned} &\big(f(U\backslash(S \cup \{e\})) - f(U\backslash S)\big) - \big(f(U\backslash(T \cup \{e\})) - f(U\backslash T)\big) \\ &\quad = \big(f(U\backslash T) - f(U\backslash(T \cup \{e\}))\big) - \big(f(U\backslash S) - f(U\backslash(S \cup \{e\}))\big) \geq 0 \end{aligned}$$

since $S \mapsto f(S)$ is submodular and $U\backslash T \subseteq U\backslash S$, and hence $S \mapsto f(U\backslash S)$ is submodular. □

We call a submodular function $f : 2^U \to \mathbb{R}$ **symmetric** if $f(A) = f(U\backslash A)$ for all $A \subseteq U$.

A multivariate generalization of submodularity is known as k-submodularity (Ene and Nguyen, 2022) where $k \in \mathbb{N}$. In particular, 1-submodular function is equivalent to submodular function. Let $f : (k+1)^U \to \mathbb{R}$ be a set function. The function f is said to be k-**submodular** if

$$f(\mathbf{S}) + f(\mathbf{T}) \geq f(\mathbf{S} \sqcap \mathbf{T}) + f(\mathbf{S} \sqcup \mathbf{T}) \quad \forall\, \mathbf{S}, \mathbf{T} \in (k+1)^U,$$

where $\mathbf{S} \sqcap \mathbf{T}$ is the k-tuple whose i-th set is $S_i \cap T_i$ and $\mathbf{S} \sqcup \mathbf{T}$ is the k-tuple whose i-th set is $(S_i \cup T_i) \setminus \Big(\bigcup_{j \neq i}(S_j \cup T_j)\Big)$. A function f is said to be k-**supermodular** if $-f$ is k-submodular.

For $\mathbf{S} = (S_1, \ldots, S_k), \mathbf{T} = (T_1, \ldots, T_k) \in (k+1)^U$, we write $\mathbf{S} \preceq \mathbf{T}$ if and only if $S_i \subseteq T_i\ \forall i \in [\![k]\!]$. A function f is said to be **monotonically non-decreasing** (resp. **non-increasing**) if

$$f(\mathbf{S}) \leq (\text{resp. } \geq)\, f(\mathbf{T}) \quad \forall\, \mathbf{S} \preceq \mathbf{T}.$$

Let $\Delta_{e,i} f(\mathbf{S})$ be the marginal gain of adding e to the i-th set of $\mathbf{S}$:

$$\Delta_{e,i} f(\mathbf{S}) := f(S_1, \ldots, S_i \cup \{e\}, \ldots, S_k) - f(S_1, \ldots, S_i, \ldots, S_k).$$

Note that f being monotonically non-decreasing is equivalent to $\Delta_{e,i} f(\mathbf{S}) \geq 0$ for all $\mathbf{S} \in (k+1)^U$, $i \in [\![k]\!]$, and $e \notin \mathrm{supp}(\mathbf{S})$, where we define $\mathrm{supp}(\mathbf{S}) := \cup_{i=1}^{k} S_i$. A function f is said to be **pairwise monotonically non-decreasing** (resp. **non-increasing**) if

$$\Delta_{e,i} f(\mathbf{S}) + \Delta_{e,j} f(\mathbf{S}) \geq (\text{resp. } \leq)\, 0$$

for all $\mathbf{S} \in (k+1)^U$, $e \notin \mathrm{supp}(\mathbf{S})$, and $i, j \in [\![k]\!]$ such that $i \neq j$. A function f is said to be **orthant submodular** (resp. **orthant supermodular**) if

$$\Delta_{e,i} f(\mathbf{S}) \geq (\text{resp. } \leq) \Delta_{e,i} f(\mathbf{T}) \tag{3}$$

for all $i \in [\![k]\!]$ and $\mathbf{S}, \mathbf{T} \in (k+1)^U$ such that $\mathbf{S} \preceq \mathbf{T}$, $e \notin \mathrm{supp}(\mathbf{T})$.

The following result that we recall characterizes k-submodularity (Theorem 7 of (Ward and Živnỳ, 2016)).

Theorem 2.5 (Characterization of k-submodularity)**.** *A function f is k-submodular (resp. k-supermodular) if and only if f is both orthant submodular (resp. supermodular) and pairwise monotonically non-decreasing (resp. non-increasing).*

The next two results relates the sum of individually supermodular or submodular functions to k-supermodularity or k-submodularity respectively.

Lemma 2.6. *Let $F : (k+1)^U \to \mathbb{R}$ defined to be*

$$F(\mathbf{S}) = F(S_1, \ldots, S_k) := \sum_{i=1}^{k} F_i(S_i)$$

be the sum of k monotonically non-increasing and supermodular functions $(F_i)_{i=1}^k$ with $F_i : 2^U \to \mathbb{R}$ for all $i \in [\![k]\!]$. Then F is k-supermodular.

Proof. Throughout this proof, let $i \neq j \in [\![k]\!]$. First, we seek to prove that F is pairwise monotonically non-increasing, in which case we aim to show $\Delta_{e,i}F(\mathbf{S}) + \Delta_{e,j}F(\mathbf{S}) \leq 0$ for $e \notin \text{supp}(\mathbf{S})$:

$$\Delta_{e,i}F(\mathbf{S}) + \Delta_{e,j}F(\mathbf{S}) = (F_i(S_i \cup \{e\}) - F_i(S_i)) + (F_j(S_j \cup \{e\}) - F_j(S_i)) \leq 0,$$

given that F_i, F_j are both monotonically non-increasing. Next, we seek to show that F is orthant supermodular, in which case we aim to show that $\Delta_{e,i}F(\mathbf{S}) \leq \Delta_{e,i}F(\mathbf{T})$ for any $\mathbf{S} \preceq \mathbf{T}$ and $e \notin \text{supp}(\mathbf{T})$:

$$\Delta_{e,i}F(\mathbf{S}) - \Delta_{e,i}F(\mathbf{T}) = (F_i(S_i \cup \{e\}) - F_i(S_i)) - (F_i(T_i \cup \{e\}) - F_i(T_i)) \leq 0,$$

given that F_i is supermodular. Therefore, F is k-supermodular given that it is pairwise monotonically non-increasing and orthant supermodular using Theorem 2.5. □

Corollary 2.7. *Let $G : (k+1)^U \to \mathbb{R}$ defined to be*

$$G(\mathbf{S}) = G(S_1, \ldots, S_k) := \sum_{i=1}^{k} G_i(S_i)$$

be the sum of k monotonically non-decreasing and submodular functions $(G_i)_{i=1}^k$ with $G_i : 2^U \to \mathbb{R}$ for all $i \in [\![k]\!]$. Then G is k-submodular.

Proof. By applying Lemma 2.6 to $-G$, we see that $-G$ is k-supermodular, which is equivalent to G being k-submodular. □

The next result, that we shall apply in subsequent sections, transforms a non-monotone submodular f to a monotonically non-decreasing submodular g (Proposition 14.18 of (Korte and Vygen, 2008)).

Theorem 2.8 (Transform a non-monotone submodular f to a monotone submodular g)**.** *Let $f : 2^U \to \mathbb{R}$ be a submodular function and $\beta \in \mathbb{R}$, then $g : 2^U \to \mathbb{R}$ defined by*

$$g(S) := f(S) - \beta + \sum_{e \in S} (f(U \backslash \{e\}) - f(U))$$

is submodular and monotonically non-decreasing.

We aim to prove a generalized version of Theorem 2.8, that transforms a given constrained orthant submodular function into a k-submodular function. Suppose that we are given $\mathbf{V} \in (k+1)^U$. Then, constrained to $\mathbf{V}$, we can transform an orthant submodular function into a k-submodular function.

Theorem 2.9. *Let $f : (k+1)^U \to \mathbb{R}$ be an orthant submodular function, $\beta \in \mathbb{R}$ and $\mathbf{V} \in (k+1)^U$. then $g : (k+1)^U \preceq \mathbf{V} \to \mathbb{R}$ with*

$$g(\mathbf{S}) := f(\mathbf{S}) - \beta + \sum_{i=1}^{k} \sum_{e \in S_i} (f(V_1, \ldots, V_i \backslash \{e\}, \ldots, V_k) - f(V_1, \ldots, V_i, \ldots, V_k))$$

is k-submodular and monotonically non-decreasing.

Proof. Suppose that $\mathbf{S} \preceq \mathbf{T}$, $i \in [\![k]\!]$, and $e \in V_i \backslash T_i$. Since f is orthant submodular, we have $\Delta_{e,i} f(\mathbf{S}) \geq \Delta_{e,i} f(\mathbf{T})$, and hence

$$\begin{aligned}\Delta_{e,i} g(\mathbf{S}) &= \Delta_{e,i} f(\mathbf{S}) + f(V_1, \ldots, V_i \backslash \{e\}, \ldots, V_k) - f(V_1, \ldots, V_i, \ldots, V_k) \\ &\geq \Delta_{e,i} f(\mathbf{T}) + \Delta_{e,i} \sum_{j=1}^{k} \sum_{u \in T_j} (f(V_1, \ldots, V_j \backslash \{u\}, \ldots, V_k) - f(V_1, \ldots, V_j, \ldots, V_k)) \\ &= \Delta_{e,i} g(\mathbf{T}).\end{aligned}$$

This gives g is orthant submodular.

To prove the orthant monotonicity, we choose $\mathbf{S} \in (k+1)^U$, $i \in [\![k]\!]$, and $e \in V_i \backslash S_i$. From the orthant submodularity of f, since $S_i \subseteq V_i \backslash \{e\}$, we have

$$\Delta_{e,i} g(\mathbf{S}) = \Delta_{e,i} f(\mathbf{S}) - (f(V_1, \ldots, V_i, \ldots, V_k) - f(V_1, \ldots, V_i \backslash \{e\}, \ldots, V_k)) \geq 0.$$

Therefore g is monotonically non-decreasing, which implies that g is pairwise monotonically non-decreasing, and hence g is k-submodular. □

We then show some examples of submodular structures that arise in the information theory of Markov chains.

Theorem 2.10 (Submodularity of some information-theoretic functions in Markov chain theory)**.** *Let $\mathbf{w} \in \mathcal{S}_n$, $S \subseteq [\![d]\!]$, $P, P_i \in \mathcal{L}(\mathcal{X})$ be π-stationary transition matrices for $i \in [\![n]\!]$. We have*

1. *(Submodularity of the entropy rate of P) The mapping $S \mapsto H(P^{(S)})$ is submodular.*
2. *(Submodularity of the distance to $(S, [\![d]\!] \backslash S)$-factorizability of P) The mapping $S \mapsto D_{\mathrm{KL}}^{\pi}(P \| P^{(S)} \otimes P^{(-S)})$ is submodular.*
3. *(Supermodularity and monotonicity of the distance to independence) The mapping $S \mapsto \mathbb{I}(P^{(S)})$ is monotonically non-decreasing and supermodular.*
4. *(Submodularity of the entropy rate of $\overline{P}$) The mapping $S \mapsto H(\overline{P}^{(S)})$ is submodular.*
5. *(Submodularity of the weighted distance to $(S, [\![d]\!] \backslash S)$-factorizability of $\mathcal{B}$) The mapping $S \mapsto \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \overline{P}^{(S)} \otimes \overline{P}^{(-S)})$ is submodular.*

Proof. From Proposition 2.33 of (Choi et al., 2024), item (1), item (2), and item (3) hold. Since the map $S \mapsto H(P^{(S)})$ is submodular, the map $S \mapsto H(\overline{P}^{(S)})$ is submodular since $\overline{P}^{(S)}$ is the projection of $\overline{P}$ onto subset S, which proves item (4). Since

$$\sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \overline{P}^{(S)} \otimes \overline{P}^{(-S)}) = H(\overline{P}^{(S)}) + H(\overline{P}^{(-S)}) - \sum_{i=1}^{n} w_i H(P_i),$$

we can conclude that $S \mapsto \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \overline{P}^{(S)} \otimes \overline{P}^{(-S)})$ is submodular because both the map $S \mapsto H(\overline{P}^{(S)})$ and the map $S \mapsto H(\overline{P}^{(-S)})$ are submodular (by Lemma 2.4). □

Next, we investigate the map $S \mapsto \mathbb{I}(P^{(-S)})$, and show that it is monotonically non-increasing and supermodular.

Theorem 2.11 (Supermodularity and monotonicity of the distance to independence of $P^{(-S)}$)**.** *The mapping $S \mapsto \mathbb{I}(P^{(-S)})$ is monotonically non-increasing and supermodular.*

Proof. We first prove the monotonicity. Suppose $S \subseteq T \subseteq [\![d]\!]$, then $[\![d]\!] \backslash T \subseteq [\![d]\!] \backslash S$, hence according to the partition lemma (Theorem 2.1), we have:

$$\mathbb{I}(P^{(-S)}) = D(P^{(-S)} \| \otimes_{i \in [\![d]\!] \backslash S} P^{(i)}) \geq D(P^{(-T)} \| \otimes_{i \in [\![d]\!] \backslash T} P^{(i)}) = \mathbb{I}(P^{(-T)}),$$

therefore, $S \mapsto \mathbb{I}(P^{(-S)})$ is monotonically non-increasing.

We then look into the supermodularity of this map. Since

$$\mathbb{I}(P^{(-S)}) = \sum_{i \in [\![d]\!] \backslash S} H(P^{(i)}) - H(P^{(-S)}),$$

then $\mathbb{I}(P^{(-S)})$ is supermodular because $H(P^{(-S)})$ is submodular in view of Lemma 2.4 and Lemma 2.6. □

2.3 Some submodular optimization algorithms

To maximize a monotonically non-decreasing submodular function, one can apply a heuristic greedy algorithm (see Section 4 of (Nemhauser et al., 1978)) with $(1-e^{-1})$-approximation guarantee. For non-monotone submodular functions, we recall a local search algorithm (see Theorem 3.4 of (Feige et al., 2011)) in Algorithm 1 that comes along with an approximation guarantee.

Algorithm 1 Local Search Algorithm (Feige et al., 2011)

Require: Ground set U with $|U| = d$, submodular function f, positive $\epsilon > 0$
1: Initialize $S \leftarrow \{e\}$, where $f(\{e\})$ is the maximum over all singletons $e \in U$
2: **while** $\exists\, a \in U \backslash S$ such that $f(S \cup \{a\}) \geq (1 + \epsilon/d^2) f(S)$ **do**
3: $\quad S \leftarrow S \cup \{a\}$
4: **end while**
5: **if** $\exists\, a \in S$ such that $f(S \backslash \{a\}) \geq (1 + \epsilon/d^2) f(S)$ **then**
6: $\quad S \leftarrow S \backslash \{a\}$
7: $\quad$ Go back to line 2
8: **end if**
9: **Output**: $f(S)$ and $f(U \backslash S)$

Theorem 2.12 (Approximation guarantee of Algorithm 1)**.** *Algorithm 1 is a $\left(\frac{1}{3} - \frac{\epsilon}{d}\right)$-approximation algorithm for maximizing non-negative submodular functions, and $\left(\frac{1}{2} - \frac{\epsilon}{d}\right)$-approximation algorithm for maximizing non-negative symmetric submodular functions. The time complexity of Algorithm 1 is $\mathcal{O}\left(\frac{1}{\epsilon} d^3 \log d\right)$.*

In this paper, it turns out that some functions we are interested in optimizing can be written as a difference of a submodular function and a modular function. In this section, we shall consider maximizing the difference of a monotonically non-decreasing submodular g and a modular c on the ground set U with cardinality constraint being at most $m \in \mathbb{N}$. Precisely, we consider the problem

$$\max_{S \subseteq U;\ |S| \leq m} g(S) - c(S),$$

and

$$\mathrm{OPT} = \mathrm{OPT}(g, c, U, m) := \operatorname*{arg\,max}_{S \subseteq U;\ |S| \leq m} g(S) - c(S).$$

Under this setting, a distorted greedy algorithm (Algorithm 2) has been proposed along with a theoretical lower bound (Harshaw et al., 2019).

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
1: Initialize $S_0 \leftarrow \emptyset$
2: **for** $i = 0$ to $m - 1$ **do**
3: $\quad e_i \leftarrow \operatorname*{arg\,max}_{e \in U} \left\{ \left(1 - \frac{1}{m}\right)^{m-(i+1)} \left(g(S_i \cup \{e\}) - g(S_i)\right) - c(\{e\}) \right\}$
4: $\quad$ **if** $\left(1 - \frac{1}{m}\right)^{m-(i+1)} \left(g(S_i \cup \{e_i\}) - g(S_i)\right) - c(\{e_i\}) > 0$ **then**
5: $\quad\quad S_{i+1} \leftarrow S_i \cup \{e_i\}$
6: $\quad$ **else**
7: $\quad\quad S_{i+1} \leftarrow S_i$
8: $\quad$ **end if**
9: **end for**
10: **Output:** S_m.

Theorem 2.13 (Lower bound for distorted greedy algorithm)**.** *Algorithm 2 provides the following lower bound:*

$$g(S_m) - c(S_m) \geq (1 - e^{-1}) g(\mathrm{OPT}) - c(\mathrm{OPT}),$$

where S_m is the final output set.

Let $\mathbf{V} \in (k+1)^U$, and consider maximizing the difference of a monotonically non-decreasing k-submodular g and a modular c on the ground set U with cardinality constraint being at most $m \in \mathbb{N}$. Precisely, we consider the problem

$$\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} g(\mathbf{S}) - c(\mathbf{S}), \tag{4}$$

and

$$\mathbf{OPT} = \mathbf{OPT}(g, c, U, \mathbf{V}, m) := \operatorname*{arg\,max}_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} g(\mathbf{S}) - c(\mathbf{S}).$$

We propose a generalized distorted greedy algorithm (Algorithm 3) for solving (4), which is of independent interest.

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 $\mathbf{S}_0 = (S_{0,1}, \ldots, S_{0,k}) \leftarrow \emptyset$
for $i = 0$ to $m-1$ **do**
  $(j^*, e^*) \leftarrow \operatorname*{arg\,max}_{j \in [\![k]\!], e \in V_j \setminus S_{i,j}} \left\{ \left(1 - \frac{1}{m}\right)^{m-(i+1)} \Delta_{e,j} g(\mathbf{S}_i) - c(\{e\}) \right\}$
  if $\left(1 - \frac{1}{m}\right)^{m-(i+1)} \Delta_{e^*,j^*} g(\mathbf{S}_i) - c(\{e^*\}) > 0$ **then**
    $S_{i+1,j^*} \leftarrow S_{i,j^*} \cup \{e^*\}$
  else
    $S_{i+1,j^*} \leftarrow S_{i,j^*}$
  end if
  for $l \neq j^*$ **do**
    $S_{i+1,l} \leftarrow S_{i,l}$
  end for
end for
Output: $\mathbf{S}_m = (S_{m,1}, \ldots, S_{m,k})$.

The rest of this section is devoted to giving a lower bound for the generalized distorted greedy algorithm. We assume that g is monotonically non-decreasing, k-submodular, $g(\emptyset) \geq 0$, while c is non-negative, modular and $c(\emptyset) = 0$.

In order to prove the lower bound for the generalized distorted greedy algorithm, we first define the distorted objective function $\Phi_i : (k+1)^U \to \mathbb{R}$, for $m \in \mathbb{N}$ and $0 \leq i \leq m-1$, that

$$\Phi_i(\mathbf{S}) := (1 - m^{-1})^{m-i} g(\mathbf{S}) - c(\mathbf{S}).$$

We also denote $\Psi_i : (k+1)^U \times [\![k]\!] \times U \to \mathbb{R}$ that

$$\Psi_i(\mathbf{S}, j, e) := \max\{0, (1 - m^{-1})^{m-(i+1)} \Delta_{e,j} g(\mathbf{S}) - c(\{e\})\}.$$

Lemma 2.14. *The difference of the distorted objective function of two iterations can be written as*

$$\Phi_{i+1}(\mathbf{S}_{i+1}) - \Phi_i(\mathbf{S}_i) = \Psi_i(\mathbf{S}_i, j^*, e^*) + \frac{1}{m}\left(1 - \frac{1}{m}\right)^{m-(i+1)} g(\mathbf{S}_i).$$

Proof. Similar to Lemma 1 of (Harshaw et al., 2019), we can show

$$\begin{aligned}
\Phi_{i+1}(\mathbf{S}_{i+1}) - \Phi_i(\mathbf{S}_i) &= \left(1 - \frac{1}{m}\right)^{m-(i+1)} g(\mathbf{S}_{i+1}) - c(\mathbf{S}_{i+1}) - \left(1 - \frac{1}{m}\right)^{m-i} g(\mathbf{S}_i) + c(\mathbf{S}_i) \\
&= \left(1 - \frac{1}{m}\right)^{m-(i+1)} g(\mathbf{S}_{i+1}) - c(\mathbf{S}_{i+1}) - \left(1 - \frac{1}{m}\right)^{m-(i+1)} \left(1 - \frac{1}{m}\right) g(\mathbf{S}_i) + c(\mathbf{S}_i) \\
&= \left(1 - \frac{1}{m}\right)^{m-(i+1)} (g(\mathbf{S}_{i+1}) - g(\mathbf{S}_i)) - (c(\mathbf{S}_{i+1}) - c(\mathbf{S}_i)) \\
&\quad + \frac{1}{m}\left(1 - \frac{1}{m}\right)^{m-(i+1)} g(\mathbf{S}_i).
\end{aligned}$$

If $(1-m^{-1})^{m-(i+1)}\Delta_{e^*,j^*}g(\mathbf{S})-c(\{e^*\})>0$, then e^* is added to the solution set. In the algorithm we have $e^*\in V_{j^*}\backslash S_{i,j^*}$, $g(\mathbf{S}_{i+1})-g(\mathbf{S}_i)=\Delta_{e^*,j^*}g(\mathbf{S}_i)$, $c(\mathbf{S}_{i+1})-c(\mathbf{S}_i)=c(\{e^*\})$, hence

$$\Phi_{i+1}(\mathbf{S}_{i+1})-\Phi_i(\mathbf{S}_i)=\Psi_i(\mathbf{S}_i,j^*,e^*)+\frac{1}{m}\left(1-\frac{1}{m}\right)^{m-(i+1)}g(\mathbf{S}_i).$$

If $(1-m^{-1})^{m-(i+1)}\Delta_{e^*,j^*}g(\mathbf{S})-c(\{e_i\})\leq 0$, the algorithm does not add e^* into the solution set, hence $\mathbf{S}_{i+1}=\mathbf{S}_i$. In this case, we also have

$$\Phi_{i+1}(\mathbf{S}_{i+1})-\Phi_i(\mathbf{S}_i)=0+\frac{1}{m}\left(1-\frac{1}{m}\right)^{m-(i+1)}g(\mathbf{S}_i)=\Psi_i(\mathbf{S}_i,j^*,e^*)+\frac{1}{m}\left(1-\frac{1}{m}\right)^{m-(i+1)}g(\mathbf{S}_i).$$

Summarizing these two cases, we see that

$$\Phi_{i+1}(\mathbf{S}_{i+1})-\Phi_i(\mathbf{S}_i)=\Psi_i(\mathbf{S}_i,j^*,e^*)+\frac{1}{m}\left(1-\frac{1}{m}\right)^{m-(i+1)}g(\mathbf{S}_i).$$

□

Lemma 2.15. *A lower bound for Ψ_i is*

$$\Psi_i(\mathbf{S}_i,j^*,e^*)\geq\frac{1}{m}\left(\left(1-\frac{1}{m}\right)^{m-(i+1)}\big(g(\mathbf{OPT})-g(\mathbf{S}_i)\big)-c(\mathbf{OPT})\right).$$

Proof. For $j\in[\![k]\!]$, let

$$\begin{aligned}U_{i,j}&:=(V_j\backslash S_{i,j})\cap\mathrm{OPT}_j,\\ U_i&:=\bigcup_{j=1}^{k}U_{i,j},\\ \mathbf{U}_i&:=(U_{i,1},U_{i,2},\ldots,U_{i,k}),\end{aligned}$$

and hence

$$S_{i,j}\cup U_{i,j}=S_{i,j}\cup\mathrm{OPT}_j. \tag{5}$$

We then have

$$\begin{aligned}m\Psi_i(\mathbf{S}_i,j^*,e^*)&=m\max_{j\in[\![k]\!],e\in V_j\backslash S_{i,j}}\left\{0,\left(1-\frac{1}{m}\right)^{m-(i+1)}\Delta_{e,j}g(\mathbf{S}_i)-c(\{e\})\right\}\\ &\geq|\mathrm{supp}(\mathbf{OPT})|\max_{j\in[\![k]\!],e\in U_{i,j}}\left\{0,\left(1-\frac{1}{m}\right)^{m-(i+1)}\Delta_{e,j}g(\mathbf{S}_i)-c(\{e\})\right\}\\ &\geq|U_i|\max_{j\in[\![k]\!],e\in U_{i,j}}\left\{\left(1-\frac{1}{m}\right)^{m-(i+1)}\Delta_{e,j}g(\mathbf{S}_i)-c(\{e\})\right\}\\ &\geq\sum_{j=1}^{k}\sum_{e\in U_{i,j}}\left(\left(1-\frac{1}{m}\right)^{m-(i+1)}\Delta_{e,j}g(\mathbf{S}_i)-c(\{e\})\right)\\ &=\left(1-\frac{1}{m}\right)^{m-(i+1)}\sum_{j=1}^{k}\sum_{e\in U_{i,j}}\Delta_{e,j}g(\mathbf{S}_i)-c(\mathbf{U}_i)\\ &\geq\left(1-\frac{1}{m}\right)^{m-(i+1)}\sum_{j=1}^{k}\sum_{e\in U_{i,j}}\Delta_{e,j}g(\mathbf{S}_i)-c(\mathbf{OPT}),\end{aligned}$$

where the last inequality follows from the fact that c is non-negative. Then, the desired result follows if we show that

$$\sum_{j=1}^{k}\sum_{e\in U_{i,j}}\Delta_{e,j}g(\mathbf{S_i})\geq g(\mathbf{OPT})-g(\mathbf{S}_i).$$

Since g is orthant submodular, by Lemma 1.1 of (Lee et al., 2010), we have

$$\sum_{e\in U_{i,j}} \Delta_{e,j} g(\mathbf{S_i}) \geq g(S_{i,1},\ldots,S_{i,j-1},S_{i,j}\cup U_{i,j},S_{i,j+1},\ldots,S_k) - g(\mathbf{S}_i),$$

and hence it further suffices to prove

$$\sum_{j=1}^{k} g(S_{i,1},\ldots,S_{i,j-1},S_{i,j}\cup U_{i,j},S_{i,j+1},\ldots,S_k) \geq g(\mathbf{OPT}) + (k-1)g(\mathbf{S}_i). \tag{6}$$

Since g is k-submodular, then

$$g(\mathbf{X}) + g(\mathbf{Y}) \geq g(\mathbf{X}\sqcup\mathbf{Y}) + g(\mathbf{X}\sqcap\mathbf{Y}),$$

for any $\mathbf{X},\mathbf{Y}\in(k+1)^U$. We seek to apply this definition to update each of the k coordinates by adding $(U_{i,j})_{j=1}^k$ sequentially. For the first step, we have

$$\begin{aligned}
&g(S_{i,1}\cup U_{i,1},S_{i,2},\ldots,S_{i,k}) + g(S_{i,1},S_{i,2}\cup U_{i,2},S_{i,3},\ldots,S_{i,k})\\
&\geq g((S_{i,1}\cup U_{i,1})\backslash(\cup_{l\neq 1}^k S_{i,l}\cup U_{i,2}),(S_{i,2}\cup U_{i,2})\backslash(\cup_{l\neq 2}^k S_{i,l}\cup U_{i,1}),S_{i,3},\ldots,S_{i,k}) + g(\mathbf{S}_i)\\
&= g(S_{i,1}\cup U_{i,1},S_{i,2}\cup U_{i,2},S_{i,3},\ldots,S_{i,k}) + g(\mathbf{S}_i),
\end{aligned}$$

where the last equality uses the fact that with $n\in[\![k]\!]$,

$$(S_{i,n}\cup U_{i,n}) = (S_{i,n}\cup U_{i,n})\backslash(\cup_{l\neq n}^k (S_{i,l}\cup U_{i,l})).$$

In the n-th step with $n\in[\![k]\!]$, we thus have

$$\begin{aligned}
&g(S_{i,1}\cup U_{i,1},\ldots,S_{i,n}\cup U_{i,n},\ldots,S_{i,k}) + g(S_{i,1},\ldots,S_{i,n},S_{i,n+1}\cup U_{i,n+1},\ldots,S_{i,k})\\
&\geq g(S_{i,1}\cup U_{i,1},\ldots,S_{i,n+1}\cup U_{i,n+1},\ldots,S_{i,k}) + g(\mathbf{S}_i).
\end{aligned}$$

Repeating the above analysis leads to

$$\sum_{j=1}^{k} g(S_{i,1},\ldots,S_{i,j-1},S_{i,j}\cup U_{i,j},S_{i,j+1},\ldots,S_k) \geq g(\mathbf{S}_i\sqcup\mathbf{U}_i) + (k-1)g(\mathbf{S}_i).$$

Finally, using the assumption that g is monotonically non-decreasing and $\mathbf{OPT}\preceq\mathbf{S}_i\sqcup\mathbf{U}_i$ in view of (5), we have

$$\sum_{j=1}^{k} g(S_{i,1},\ldots,S_{i,j-1},S_{i,j}\cup U_{i,j},S_{i,j+1},\ldots,S_k) \geq g(\mathbf{OPT}) + (k-1)g(\mathbf{S}_i),$$

and hence (6) holds. □

Finally, we prove a lower bound for the generalized distorted greedy algorithm:

Theorem 2.16 (Lower bound for generalized distorted greedy algorithm)**.** *Algorithm 3 provides the following lower bound:*

$$g(\mathbf{S}_m) - c(\mathbf{S}_m) \geq (1-e^{-1})g(\mathbf{OPT}) - c(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1},\ldots,S_{m,k})$ *is the final output set.*

Proof. According to our assumptions, we have

$$\Phi_0(\mathbf{S}_0) = \left(1-\frac{1}{m}\right)^m g(\emptyset) - c(\emptyset) \geq 0$$

and

$$\Phi_m(\mathbf{S}_m) = \left(1-\frac{1}{m}\right)^0 g(\mathbf{S}_m) - c(\mathbf{S}_m) = g(\mathbf{S}_m) - c(\mathbf{S}_m).$$

Therefore, we have

$$g(\mathbf{S}_m) - c(\mathbf{S}_m) \geq \Phi_m(\mathbf{S}_m) - \Phi_0(\mathbf{S}_0) = \sum_{i=0}^{m-1} \Phi_{i+1}(\mathbf{S}_{i+1}) - \Phi_i(\mathbf{S}_i). \tag{7}$$

We apply Lemma 2.14 and 2.15 to yield

$$\begin{aligned}\Phi_{i+1}(\mathbf{S}_{i+1}) - \Phi_i(\mathbf{S}_i) &= \Psi_i(\mathbf{S}_i, j^*, e^*) + \frac{1}{m}\left(1 - \frac{1}{m}\right)^{m-(i+1)} g(\mathbf{S}_i) \\ &\geq \frac{1}{m}\left(1 - \frac{1}{m}\right)^{m-(i+1)} g(\mathbf{OPT}) - \frac{1}{m} c(\mathbf{OPT}).\end{aligned}$$

We plug the above bound into (7) to obtain

$$\begin{aligned}g(\mathbf{S}_m) - c(\mathrm{supp}(\mathbf{S}_m)) &\geq \sum_{i=0}^{m-1} \left[\frac{1}{m}\left(1 - \frac{1}{m}\right)^{m-(i+1)} g(\mathbf{OPT}) - \frac{1}{m} c(\mathbf{OPT})\right] \\ &= \left[\frac{1}{m}\sum_{i=0}^{m-1}\left(1 - \frac{1}{m}\right)^{i}\right] g(\mathbf{OPT}) - c(\mathbf{OPT}) \\ &= \left(1 - \left(1 - \frac{1}{m}\right)^{m}\right) g(\mathbf{OPT}) - c(\mathbf{OPT}) \\ &\geq (1 - e^{-1}) g(\mathbf{OPT}) - c(\mathbf{OPT}).\end{aligned}$$

□

2.4 Examples of multivariate Markov chains

2.4.1 Curie-Weiss model

We aim to generate a d-dimensional Markov chain from the Curie-Weiss model. We consider a discrete d-dimensional hypercube state space given by

$$\mathcal{X} = \{-1, +1\}^d.$$

Let the Hamiltonian function be that of the Curie-Weiss model (see Chapter 13 of (Bovier and Den Hollander, 2016)) on $\mathcal{X}$ with interaction coefficients $\frac{1}{2^{|j-i|}}$ and external magnetic field $h = 1$, that is, for $x = (x^1, \ldots, x^d) \in \mathcal{X}$,

$$\mathcal{H}(x) = -\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{2^{|j-i|}} x^i x^j - h\sum_{i=1}^{d} x^i.$$

We consider a Glauber dynamics with a simple random walk proposal targeting the Gibbs distribution at temperature $T = 10$. At each step we pick uniformly at random one of the d coordinates and flip it to the opposite sign, along with an acceptance-rejection filter, that is,

$$P(x, y) = \begin{cases} \frac{1}{d} e^{-\frac{1}{T}(\mathcal{H}(y) - \mathcal{H}(x))_+}, & \text{if } y = (x^1, x^2, \ldots, -x^i, \ldots, x^d), i \in [\![d]\!], \\ 1 - \sum_{y;\ y \neq x} P(x, y), & \text{if } x = y, \\ 0, & \text{otherwise,} \end{cases}$$

where for $m \in \mathbb{R}$ we denote $m_+ := \max\{m, 0\}$ as the non-negative part of m. The stationary distribution of P is the Gibbs distribution at temperature T given by

$$\pi(x) = \frac{e^{-\frac{1}{T}\mathcal{H}(x)}}{\sum_{z \in \mathcal{X}} e^{-\frac{1}{T}\mathcal{H}(z)}}.$$

2.4.2 Bernoulli-Laplace level model

We aim to generate a d-dimensional Markov chain from the Bernoulli-Laplace level model. We consider a $(d+1)$-dimensional Bernoulli-Laplace level model as described in Section 4.2 of (Khare and Zhou, 2009). Let

$$\mathcal{X} = \{x = (x^1, \ldots, x^{d+1}) \in \mathbb{N}_0^{d+1};\ x^1 + \ldots + x^{d+1} = N\}$$

be the state space, where x^i can be interpreted as the number of "particles" of type i out of the total number $N = d$. The stationary distribution of such Markov chain, π, is given by the multivariate hypergeometric distribution described in Lemma 4.18 of (Khare and Zhou, 2009). Concretely, we have

$$\pi(x) = \frac{\prod_{i=1}^{d+1} \binom{l_i}{x^i}}{\binom{l_1+\ldots+l_{d+1}}{N}}, \quad x \in \mathcal{X},$$

for some fixed parameters $l_1 = \ldots = l_d = 1$ and $l_{d+1} = d$, which represents the total number of "particles" of type i. Under this setting, we let $x^{d+1} = N - \sum_{i=1}^{d} x^i$, and hence the state space is of product form with $\mathcal{X} = \{0, 1\}^d$.

Following the spectral decomposition for reversible Markov chains (see Section 2.1 of (Khare and Zhou, 2009) for background), the transition matrix P is written as:

$$P(x, y) = \sum_{n=0}^{N} \beta_n \phi_n(x) \phi_n(y) \pi(y),$$

where β_n are the eigenvalues and $\phi_n(x)$ is the eigenfunction.

From Definition 4.15 of (Khare and Zhou, 2009), in the Bernoulli-Laplace level model, s is the swap size parameter satisfying

$$0 \leq s \leq \min\left\{N, \sum_{i=1}^{d+1} l_i - N\right\},$$

where we consider $\sum_{i=1}^{d+1} l_i > N$. From Theorem 4.19 of (Khare and Zhou, 2009), the eigenvalues for the Bernoulli-Laplace level model are given by

$$\beta_n = \sum_{k=0}^{n} \binom{n}{k} \frac{(N-s)_{[n-k]} s_{[k]}}{N_{[n-k]} \left(\sum_{i=1}^{d+1} l_i - N\right)_{[k]}}, \quad 0 \leq n \leq N,$$

where $a_{[k]} = a(a-1)\cdots(a-k+1)$, and we apply the convention that $a_{[0]} = 1$.

In this case, we choose the eigenfunction as

$$\phi_n(x) = \left\{\mathbf{Q_n}\left(x; N, -\sum_{i=1}^{d+1} l_i\right)\right\}_{|\mathbf{n}|=n},$$

where $\mathbf{Q_n}$ are the multivariate Hahn polynomials for the hypergeometric distribution as defined in Proposition 2.3 of (Khare and Zhou, 2009).

Part I
Subset selection for a single multivariate Markov chain

3 Submodular maximization of the entropy rate $H(P^{(S)})$

Given $P \in \mathcal{L}(\mathcal{X})$ and $m \in \mathbb{N}$, we aim to investigate the following submodular maximization problem with cardinality constraint:

$$\max_{S \subseteq [\![d]\!];\ |S| \leq m} H(P^{(S)}). \tag{8}$$

From Theorem 2.10, the map $S \mapsto H(P^{(S)})$ is submodular but generally not monotonically non-decreasing. Since the widely-used heuristic greedy algorithm is near-optimal only when the objective submodular function is monotonically non-decreasing (see Section 4 of (Nemhauser et al., 1978)), in this regard our problem does not have a classical greedy-based approximation guarantee. On the other hand, since $H(P^{(S)}) \geq 0$ and $H(P^{(\emptyset)}) = 0$, if we consider the unconstrained maximization problem of (8), we can apply Algorithm 1 with $\left(\frac{1}{3} - \frac{\epsilon}{d}\right)$-approximation guarantee (see Theorem 2.12).

Instead, we consider

$$H(P) = H(\pi \boxtimes P) - H(\pi),$$

where we define the edge measure of P with respect to π as $(\pi \boxtimes P)(x, y) := \pi(x)P(x, y)$ and $\pi \boxtimes P \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$.

Then, the map

$$S \mapsto H(P^{(S)}) = H(\pi^{(S)} \boxtimes P^{(S)}) - H(\pi^{(S)}) \tag{9}$$

can be considered as a monotonically non-decreasing submodular function $H(\pi^{(S)} \boxtimes P^{(S)})$ minus a non-negative modular function $H(\pi^{(S)})$ if we assume π to be of product form. This fits into the setting of the distorted greedy as in Algorithm 2, and leads us to Corollary 3.1.

Corollary 3.1. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary where π is of product form. In Algorithm 2, we take $g(S) = H(\pi^{(S)} \boxtimes P^{(S)})$, $c(S) = H(\pi^{(S)})$, and $\mathrm{OPT} = \arg\max_{S \subseteq [\![d]\!];\ |S| \leq m} H(P^{(S)})$. Therefore, Theorem 2.13 gives*

$$H(P^{(S_m)}) \geq (1 - e^{-1})H(\pi^{(\mathrm{OPT})} \boxtimes P^{(\mathrm{OPT})}) - H(\pi^{(\mathrm{OPT})}),$$

where S_m is the output of Algorithm 2.

More generally for P with non-product-form π as stationary distribution, in view of Theorem 2.8, for any $\beta \in \mathbb{R}$ we have a monotonically non-decreasing submodular g given by

$$g(S) = H(P^{(S)}) - \beta + \sum_{e \in S}(H(P^{(-e)}) - H(P)), \tag{10}$$

and we also denote the following modular function

$$\begin{aligned} c(S) &= -\beta + \sum_{e \in S}(H(P^{(-e)}) - H(P)) \\ &= -\beta + \sum_{e \in S}(D(P \| P^{(e)} \otimes P^{(-e)}) - H(P^{(e)})). \end{aligned} \tag{11}$$

As $H(P^{(e)}) \leq \log |\mathcal{X}^{(e)}|$, c is ensured to be non-negative if $\beta \leq -\sum_{i=1}^{d} \log |\mathcal{X}^{(i)}|$. Since

$$H(P^{(S)}) = g(S) - c(S),$$

we can employ Algorithm 2 to perform distorted greedy maximization with a lower bound.

Corollary 3.2. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary. In Algorithm 2, we take g as in* (10)*, c as in* (11)*, $\beta \leq -\sum_{i=1}^{d} \log |\mathcal{X}^{(i)}|$, and* $\mathrm{OPT} = \arg\max_{S \subseteq [\![d]\!];\ |S| \leq m} H(P^{(S)})$*. Therefore, Theorem 2.13 gives*

$$H(P^{(S_m)}) \geq (1 - e^{-1}) g(\mathrm{OPT}) - c(\mathrm{OPT}),$$

where S_m is the output of Algorithm 2.

Note that the lower bound of Corollary (3.2) depends on β through g and c. If β is chosen to be too small, then the lower bound might be too loose as the right hand side might be negative.

3.1 k-submodular maximization of the entropy rate of the tensorized keep-S_i-in matrices $H(\otimes_{i=1}^{k} P^{(S_i)})$

In this subsection, we investigate the the following map

$$(k+1)^{[\![d]\!]} \ni \mathbf{S} = (S_1, \ldots, S_k) \mapsto f(\mathbf{S}) = H(\otimes_{i=1}^{k} P^{(S_i)}) = \sum_{i=1}^{k} H(P^{(S_i)}), \tag{12}$$

and consider maximization problems of the form, for given $\mathbf{V} \in (k+1)^{[\![d]\!]}$,

$$\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} H(\otimes_{i=1}^{k} P^{(S_i)}). \tag{13}$$

In the special case of $k = 1$ and $\mathbf{V} = [\![d]\!]$, we recover the problem (8).

First, we consider the special case where P is π-stationary with π taking on a product form. Similar to the map (9), we re-write the map (12) as

$$\mathbf{S} \mapsto f(\mathbf{S}) = \sum_{i=1}^{k} H(\pi^{(S_i)} \boxtimes P^{(S_i)}) - \sum_{i=1}^{k} H(\pi^{(S_i)}). \tag{14}$$

Since $H(\pi^{(S_i)} \boxtimes P^{(S_i)})$ is monotonically non-decreasing and submodular, then by Corollary 2.7, the following function g is monotonically non-decreasing and k-submodular

$$g(\mathbf{S}) = \sum_{i=1}^{k} H(\pi^{(S_i)} \boxtimes P^{(S_i)}). \tag{15}$$

Since π is of product form, we denote the non-negative modular function c as

$$c(\mathbf{S}) = \sum_{i=1}^{k} H(\pi^{(S_i)}). \tag{16}$$

Therefore, we have

$$f(\mathbf{S}) = g(\mathbf{S}) - c(\mathbf{S}),$$

and the distorted greedy algorithm yields an approximate maximizer with a lower bound as in Theorem 2.16.

Corollary 3.3. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary where π is of product form. In Algorithm 3, we take g as in* (15) *and c as in* (16)*, and* $\mathbf{OPT} = \arg\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} f(\mathbf{S})$*. Then by Theorem 2.16, we have the following lower bound*

$$f(\mathbf{S}_m) = H(\otimes_{i=1}^{k} P^{(S_{m,i})}) \geq (1 - e^{-1}) g(\mathbf{OPT}) - c(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1}, \ldots, S_{m,k})$ is the output of Algorithm 3.

In the special case where $k = 1$ and $\mathbf{V} = [\![d]\!]$, we recover Corollary 3.1.

Next, we investigate the case where P is π-stationary for general π which may not be of product form. We first prove an orthant submodularity result.

Lemma 3.4. *The map* (12) *is orthant submodular.*

Proof. We shall prove that $\Delta_{e,i}f(\mathbf{S}) \geq \Delta_{e,i}f(\mathbf{T})$, where we choose $\mathbf{S} \preceq \mathbf{T}$ and $e \notin \mathrm{supp}(\mathbf{T})$. Given the submodularity of $S \mapsto H(P^{(S)})$, we have

$$H(P^{(S_i\cup\{e\})}) - H(P^{(S_i)}) \geq H(P^{(T_i\cup\{e\})}) - H(P^{(T_i)}),$$

which is equivalent to $\Delta_{e,i}f(\mathbf{S}) \geq \Delta_{e,i}f(\mathbf{T})$. □

In view of Theorem 2.9, since the map (12) is orthant submodular, then for any $\beta \in \mathbb{R}$, if $\mathbf{S} \preceq \mathbf{V}$, we have a monotonically non-decreasing k-submodular function g given by

$$g(\mathbf{S}) = \sum_{i=1}^{k} H(P^{(S_i)}) - \beta + \sum_{i=1}^{k}\sum_{e\in S_i}(H(P^{(V_i\setminus\{e\})}) - H(P^{(V_i)})), \tag{17}$$

and we also denote the following modular function

$$\begin{aligned} c(\mathbf{S}) &= -\beta + \sum_{i=1}^{k}\sum_{e\in S_i}(H(P^{(V_i\setminus\{e\})}) - H(P^{(V_i)})) \\ &= -\beta + \sum_{i=1}^{k}\sum_{e\in S_i}(D(P^{(V_i)}\|P^{(e)}\otimes P^{(V_i\setminus\{e\})}) - H(P^{(e)})). \end{aligned} \tag{18}$$

As $H(P^{(e)}) \leq \log|\mathcal{X}^{(e)}|$, c is ensured to be non-negative if $\beta \leq -\sum_{i=1}^{k}\sum_{e\in V_i}\log|\mathcal{X}^{(e)}|$. Since

$$f(\mathbf{S}) = \sum_{i=1}^{k} H(P^{(S_i)}) = g(\mathbf{S}) - c(\mathbf{S}),$$

then we can apply Algorithm 3 to perform distorted greedy maximization with a guaranteed lower bound.

Corollary 3.5. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary and $\mathbf{V} \in (k+1)^{[\![d]\!]}$. In Algorithm 3, we take g as in* (17) *and c as in* (18)*, $\beta \leq -\sum_{i=1}^{k}\sum_{e\in V_i}\log|\mathcal{X}^{(e)}|$, and $\mathbf{OPT} = \arg\max_{\mathbf{S}\preceq\mathbf{V};\ |\mathrm{supp}(\mathbf{S})|\leq m} f(\mathbf{S})$. Therefore, Theorem 2.16 gives*

$$f(\mathbf{S}_m) = H(\otimes_{i=1}^{k}P^{(S_{m,i})}) \geq (1-e^{-1})g(\mathbf{OPT}) - c(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1}, S_{m,2}, \ldots, S_{m,k})$ is the output of Algorithm 3.

Note that the lower bound of Corollary 3.5 depends on β through g and c. If β is chosen to be too small, then the lower bound might be too loose as the right hand side might be negative.

4 Submodular optimization of distance to factorizability $D(P\|P^{(S)}\otimes P^{(-S)})$

4.1 Submodular minimization of the distance to factorizability

For

$$2^{[\![d]\!]} \ni S \mapsto D(P\|P^{(S)}\otimes P^{(-S)}),$$

we first recall that this map is submodular (see Lemma 2.10). Since $D(P\|P^{(S)}\otimes P^{(-S)}) = D(P\|P^{(-S)}\otimes P^{(S)})$, then this map is also symmetric. In this case, there exists an algorithm for minimizing non-negative symmetric submodular functions (see Theorem 14.25 of (Korte and Vygen, 2008)) that gives

$$S^* \in \operatorname*{arg\,min}_{\emptyset\neq S\subset[\![d]\!];\ |S|\leq m} D(P\|P^{(S)}\otimes P^{(-S)})$$

with time complexity $\mathcal{O}(d^3\theta)$. Here, θ denotes the worst case time needed to evaluate $D(P\|P^{(S)}\otimes P^{(-S)})$ for any given subset S.

4.2 Submodular maximization of the distance to factorizability

Given $P \in \mathcal{L}(\mathcal{X})$ and $m \in \mathbb{N}$, we aim to investigate the following submodular maximization problem subject to a cardinality constraint

$$\max_{S \subseteq [\![d]\!];\ |S| \leq m} D(P \| P^{(S)} \otimes P^{(-S)}). \tag{19}$$

Since $D(P\|P^{(S)} \otimes P^{(-S)}) \geq 0$ and $D(P\|P^{(\emptyset)} \otimes P^{([\![d]\!])}) = 0$, if we consider the unconstrained version of (19), we can apply Algorithm 1 with $\left(\frac{1}{2} - \frac{\epsilon}{d}\right)$-approximation guarantee (see Theorem 2.12) since $D(P\|P^{(S)} \otimes P^{(-S)})$ is symmetric.

In view of Theorem 2.8, we choose $\beta = 0$ and take

$$g(S) = D(P\|P^{(S)} \otimes P^{(-S)}) + \sum_{e \in S} D(P\|P^{(-e)} \otimes P^{(e)}), \tag{20}$$

which is submodular and monotonically non-decreasing. In this case, we also take the modular and non-negative function c to be

$$c(S) = \sum_{e \in S} D(P\|P^{(-e)} \otimes P^{(e)}). \tag{21}$$

Therefore,

$$D(P\|P^{(S)} \otimes P^{(-S)}) = g(S) - c(S)$$

can be expressed as the difference of a non-negative, submodular, monotonically non-decreasing function and a non-negative modular function, hence Algorithm 2 can be applied to approximately maximize $D(P\|P^{(S)} \otimes P^{(-S)})$.

Corollary 4.1. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary. In Algorithm 2, we take g as in* (20) *and c as in* (21)*, and* $\mathrm{OPT} = \arg\max_{S \subseteq [\![d]\!];\ |S| \leq m} D(P\|P^{(S)} \otimes P^{(-S)})$*. By Theorem 2.13, we have*

$$D(P\|P^{(S_m)} \otimes P^{(-S_m)}) \geq (1 - e^{-1}) g(\mathrm{OPT}) - c(\mathrm{OPT}),$$

where S_m is the final output set of Algorithm 2.

4.3 k-submodular maximization of distance to factorizability of the tensorized keep-S_i-in matrices $D(P\|P^{(S_1)} \otimes \ldots \otimes P^{(S_k)} \otimes P^{(-\cup_{i=1}^k S_i)})$

In this section, we investigate the following map

$$(k+1)^{[\![d]\!]} \ni \mathbf{S} \mapsto f(\mathbf{S}) = D(P\|P^{(S_1)} \otimes \ldots \otimes P^{(S_k)} \otimes P^{(-\cup_{i=1}^k S_i)}), \tag{22}$$

We consider the maximization problem of the form, for given $\mathbf{V} \in (k+1)^{[\![d]\!]}$,

$$\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} D(P\|P^{(S_1)} \otimes \ldots \otimes P^{(S_k)} \otimes P^{(-\cup_{i=1}^k S_i)}). \tag{23}$$

In the special case of $k = 1$ and $\mathbf{V} = [\![d]\!]$, we recover problem (19).

Lemma 4.2. *The map* (22) *is orthant submodular.*

Proof. We shall prove that $\Delta_{e,i} f(\mathbf{S}) \geq \Delta_{e,i} f(\mathbf{T})$, where we choose $\mathbf{S} \preceq \mathbf{T}$ and $e \notin \mathrm{supp}(\mathbf{T})$. We compute that

$$\begin{aligned}
\Delta_{e,i} f(\mathbf{S}) - \Delta_{e,i} f(\mathbf{T}) &= H(P^{(S_i \cup \{e\})}) - H(P^{(S_i)}) + H(P^{(-\mathrm{supp}(\mathbf{S}) \cup \{e\})}) - H(P^{(-\mathrm{supp}(\mathbf{S}))}) \\
&\quad - H(P^{(T_i \cup \{e\})}) + H(P^{(T_i)}) - H(P^{(-\mathrm{supp}(\mathbf{T}) \cup \{e\})}) + H(P^{(-\mathrm{supp}(\mathbf{T}))}) \\
&= \Big[\big(H(P^{(S_i \cup \{e\})}) - H(P^{(S_i)})\big) - \big(H(P^{(T_i \cup \{e\})}) - H(P^{(T_i)})\big)\Big] \\
&\quad + \Big[\big(H(P^{(-\mathrm{supp}(\mathbf{T}))}) - H(P^{(-\mathrm{supp}(\mathbf{T}) \cup \{e\})})\big) \\
&\quad - \big(H(P^{(-\mathrm{supp}(\mathbf{S}))}) - H(P^{(-\mathrm{supp}(\mathbf{S}) \cup \{e\})})\big)\Big],
\end{aligned}$$

where each of the two terms above are non-negative given the submodularity of $S \mapsto H(P^{(S)})$ (recall Theorem 2.10). □

In view of Theorem 2.9, since the map (22) is orthant submodular, for any $\beta \in \mathbb{R}$, if $\mathbf{S} \preceq \mathbf{V}$, we have a monotonically non-decreasing k-submodular function given by

$$\begin{aligned} g(\mathbf{S}) &= f(\mathbf{S}) - \beta + \sum_{i=1}^{k} \sum_{e \in S_i} \Big[D(P \| P^{(V_1)} \otimes \ldots \otimes P^{(V_i \setminus \{e\})} \otimes \ldots \otimes P^{(V_k)} \otimes P^{(-\mathrm{supp}(\mathbf{V}) \setminus \{e\})}) \\ &\quad - D(P \| P^{(V_1)} \otimes \ldots \otimes P^{(V_i)} \otimes \ldots \otimes P^{(V_k)} \otimes P^{(-\mathrm{supp}(\mathbf{V}))}) \Big] \\ &= f(\mathbf{S}) - \beta + \sum_{i=1}^{k} \sum_{e \in S_i} \Big[H(P^{(V_i \setminus \{e\})}) + H(P^{(-\mathrm{supp}(\mathbf{V}) \setminus \{e\})}) - H(P^{(V_i)}) - H(P^{(-\mathrm{supp}(\mathbf{V}))}) \Big] \\ &= f(\mathbf{S}) - \beta + \sum_{i=1}^{k} \sum_{e \in S_i} \Big[D(P^{(V_i)} \| P^{(V_i \setminus \{e\})} \otimes P^{(e)}) - D(P^{(-\mathrm{supp}(\mathbf{V}) \setminus \{e\})} \| P^{(-\mathrm{supp}(\mathbf{V}))} \otimes P^{(e)}) \Big], \quad (24) \end{aligned}$$

and we also obtain the following modular function

$$c(\mathbf{S}) = -\beta + \sum_{i=1}^{k} \sum_{e \in S_i} \Big[D(P^{(V_i)} \| P^{(V_i \setminus \{e\})} \otimes P^{(e)}) - D(P^{(-\mathrm{supp}(\mathbf{V}) \setminus \{e\})} \| P^{(-\mathrm{supp}(\mathbf{V}))} \otimes P^{(e)}) \Big]. \quad (25)$$

Thus, if we choose

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

then c is non-negative. With these choices, f can be written as

$$f(\mathbf{S}) = D(P \| P^{(S_1)} \otimes \ldots \otimes P^{(S_k)} \otimes P^{(-\cup_{i=1}^{k} S_i)}) = g(\mathbf{S}) - c(\mathbf{S}).$$

We can then apply Algorithm 3 to perform distorted greedy maximization with a lower bound.

Corollary 4.3. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary and $\mathbf{V} \in (k+1)^{[\![d]\!]}$. In Algorithm 3, we take g as in* (24) *and c as in* (25)*. We choose*

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

and let $\mathbf{OPT} = \arg\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} f(\mathbf{S})$*. Therefore, Theorem 2.16 gives*

$$f(\mathbf{S}_m) = D(P \| P^{(S_{m,1})} \otimes \ldots \otimes P^{(S_{m,k})} \otimes P^{(-\cup_{i=1}^{k} S_{m,i})}) \geq (1 - e^{-1}) g(\mathbf{OPT}) - c(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1}, \ldots, S_{m,k})$ is the output of Algorithm 3.

Note that the lower bound of Corollary 4.3 depends on β through g and c. If β is chosen to be too small, then the lower bound might be too loose as the right hand side might be negative.

5 Supermodular minimization of distance to independence $\mathbb{I}(P^{(S)})$

Given $P \in \mathcal{L}(\mathcal{X})$ and $d, m \geq 2$, we aim to investigate the following supermodular (recall Theorem 2.10) minimization problem

$$\min_{S \subseteq [\![d]\!];\ |S| = m} \mathbb{I}(P^{(S)}). \quad (26)$$

We shall be interested in the constraint $|S| = m$ rather than $|S| \leq m$ as in Section 3 and Section 4 because $S \mapsto \mathbb{I}(P^{(S)})$ is monotonically non-decreasing.

The supermodular minimization problem (26) is equivalent to the following submodular maximization problem

$$\max_{S \subseteq [\![d]\!];\ |S| = m} f(S) = -\mathbb{I}(P^{(S)}) = H(P^{(S)}) - \sum_{e \in S} H(P^{(e)}). \quad (27)$$

Note that we restrict m to be at least 2, since we have the trivial result that $\mathbb{I}(P^{(e)}) = \mathbb{I}(P^{(\emptyset)}) = 0$ if the constraint is $m = 0$ or $m = 1$. From Theorem 2.10, $f(S)$ is monotonically non-increasing and submodular. Therefore, the heuristic greedy algorithm (see Section 4 of (Nemhauser et al., 1978)) cannot provide a theoretical guarantee.

In view of Theorem 2.8, for any $\beta \in \mathbb{R}$, we have a monotonically non-decreasing submodular function g given by

$$\begin{aligned} g(S) &= f(S) - \beta + \sum_{e \in S} (H(P^{(-e)}) + H(P^{(e)}) - H(P)) \\ &= f(S) - \beta + \sum_{e \in S} D(P \| P^{(e)} \otimes P^{(-e)}). \end{aligned} \tag{28}$$

We choose $\beta = 0$ and let the following non-negative, modular function be

$$c(S) = \sum_{e \in S} D(P \| P^{(e)} \otimes P^{(-e)}) \tag{29}$$

so that $f(S) = g(S) - c(S)$. By Theorem 2.13, we can apply Algorithm 2 to obtain a lower bound.

Corollary 5.1. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary along with $d, m \geq 2$. In Algorithm 2, we take g as in* (28)*, c as in* (29)*, and* $\mathrm{OPT} = \max_{S \subseteq [\![d]\!];\ |S| = m} f(S)$*. By Theorem 2.13, we have the following lower bound*

$$f(S_m) = -\mathbb{I}(P^{(S_m)}) \geq (1 - e^{-1}) g(\mathrm{OPT}) - c(\mathrm{OPT}),$$

where S_m is the output of Algorithm 2.

5.1 Supermodular minimization of distance to independence of the complement set $\mathbb{I}(P^{(-S)})$

From Theorem 2.11, $\mathbb{I}(P^{(-S)})$ is monotonically non-increasing and supermodular. Given $P \in \mathcal{L}(\mathcal{X})$, $d \geq 2$, and $m \leq d - 2$, we shall investigate the following optimization problem

$$\max_{S \subseteq [\![d]\!];\ |S| \leq m} f(S) = -\mathbb{I}(P^{(-S)}).$$

Note that we restrict m to be at most $d - 2$, since we have the trivial result that $\mathbb{I}(P^{(e)}) = \mathbb{I}(P^{(\emptyset)}) = 0$ if the constraint is $m = d$ or $m = d - 1$.

Since $f(S) = -\mathbb{I}(P^{(-S)})$ is monotonically non-decreasing and submodular, then we can apply the heuristic greedy algorithm (see Section 4 of (Nemhauser et al., 1978)) that comes along with a $(1 - e^{-1})$-approximation guarantee.

5.2 k-supermodular minimization of distance to independence of the tensorized keep-S_i-in matrices $\mathbb{I}(\otimes_{i=1}^k P^{(S_i)})$

In this section, we investigate the following map

$$(k+1)^{[\![d]\!]} \ni \mathbf{S} = (S_1, \ldots, S_k) \mapsto \mathbb{I}(\otimes_{i=1}^k P^{(S_i)}). \tag{30}$$

Lemma 5.2. *For $k \in \mathbb{N}$ and $\mathbf{S} \in (k+1)^{[\![d]\!]}$, we have*

$$\mathbb{I}(\otimes_{i=1}^k P^{(S_i)}) = \sum_{i=1}^k \mathbb{I}(P^{(S_i)}).$$

Proof. We shall prove by induction on k. When $k = 1$, the equality trivially holds. When $k = 2$, according to the chain rule of KL divergence (see Theorem 2.15 of (Polyanskiy and Wu, 2025)),

$$\begin{aligned} \mathbb{I}(P^{(S_1)} \otimes P^{(S_2)}) &= D(P^{(S_1)} \otimes P^{(S_2)} \| \otimes_{i \in S_1 \cup S_2} P^{(i)}) \\ &= D(P^{(S_1)} \| \otimes_{i \in S_1} P^{(i)}) + D(P^{(S_2)} \| \otimes_{i \in S_2} P^{(i)}) \\ &= \mathbb{I}(P^{(S_1)}) + \mathbb{I}(P^{(S_2)}). \end{aligned}$$

Suppose $\mathbb{I}(\otimes_{i=1}^m P^{(S_i)}) = \sum_{i=1}^m \mathbb{I}(P^{(S_i)})$ holds ($k=m$), then using the chain rule of KL divergence again (Theorem 2.15 of (Polyanskiy and Wu, 2025)), we have

$$\begin{aligned}\mathbb{I}(\otimes_{i=1}^{m+1} P^{(S_i)}) &= D(\otimes_{i=1}^m P^{(S_i)} \otimes P^{(S_{m+1})} \| \otimes_{i\in(\cup_{i=1}^m S_i)\cup S_{m+1}} P^{(i)}) \\ &= D(\otimes_{i=1}^m P^{(S_i)} \| \otimes_{i\in\cup_{i=1}^m S_i} P^{(i)}) + D(P^{(S_{m+1})} \| \otimes_{i\in S_{m+1}} P^{(i)}) \\ &= \sum_{i=1}^{m+1} \mathbb{I}(P^{(S_i)}).\end{aligned}$$

□

Lemma 5.3. *The map* (30) *is pairwise monotonically non-decreasing. In particular, when P is non-factorizable and π-stationary, the map* (30) *is pairwise monotonically strictly increasing for all pairs.*

Proof. Let $f(\mathbf{S}) = \mathbb{I}(\otimes_{i=1}^k P^{(S_i)})$. We shall prove that $\Delta_{e,i} f(\mathbf{S}) + \Delta_{e,j} f(\mathbf{S}) \geq 0$, where $i \neq j \in [\![d]\!]$ and $e \notin \mathrm{supp}(\mathbf{T})$. Since $\mathbb{I}(P^{(S)}) = \sum_{i\in S} H(P^{(i)}) - H(P^{(S)})$, we note that

$$\begin{aligned}\Delta_{e,i} f(\mathbf{S}) + \Delta_{e,j} f(\mathbf{S}) &= \mathbb{I}(P^{(S_i\cup\{e\})}) - \mathbb{I}(P^{(S_i)}) + \mathbb{I}(P^{(S_j\cup\{e\})}) - \mathbb{I}(P^{(S_i)}) \\ &= \big[H(P^{(e)}) + H(P^{(S_i)}) - H(P^{(S_i\cup\{e\})})\big] \\ &\quad + \big[H(P^{(e)}) + H(P^{(S_j)}) - H(P^{(S_j\cup\{e\})})\big] \\ &= D(P^{(S_i\cup\{e\})} \| P^{(S_i)} \otimes P^{(e)}) + D(P^{(S_j\cup\{e\})} \| P^{(S_j)} \otimes P^{(e)}),\end{aligned}$$

which is non-negative. In particular, when P is non-factorizable, it is strictly positive. □

Lemma 5.4. *The map* (30) *is orthant supermodular.*

Proof. Let $f(\mathbf{S}) = \mathbb{I}(\otimes_{i=1}^k P^{(S_i)})$. For any $\mathbf{S} \preceq \mathbf{T}$, we shall prove that $\Delta_{e,i} f(\mathbf{S}) \leq \Delta_{e,i} f(\mathbf{T})$, where $i \in [\![d]\!]$ and $e \in [\![d]\!] \backslash \mathrm{supp}(\mathbf{T})$.

$$\begin{aligned}\Delta_{e,i} f(\mathbf{S}) - \Delta_{e,i} f(\mathbf{T}) &= \big[H(P^{(e)}) + H(P^{(S_i)}) - H(P^{(S_i\cup\{e\})})\big] \\ &\quad - \big[H(P^{(e)}) + H(P^{(T_i)}) - H(P^{(T_i\cup\{e\})})\big] \\ &= \big[H(P^{(T_i\cup\{e\})}) - H(P^{(T_i)})\big] - \big[H(P^{(S_i\cup\{e\})}) - H(P^{(S_i)})\big] \leq 0,\end{aligned}$$

where the inequality holds owing to the submodularity of $S \mapsto H(P^{(S)})$ in view of Theorem 2.10. □

Collecting the previous two results, we see that, for non-factorizable P, the map (30) is not k-supermodular as k-supermodularity requires both the pairwise monotonically non-increasing property and orthant supermodularity (see Theorem 2.5).

Given $P \in \mathcal{L}(\mathcal{X})$, $d, m \geq k+1$ and $\mathbf{V} \in (k+1)^{[\![d]\!]}$, since the map (30) is orthant supermodular, we are interested in the following orthant submodular maximization problem

$$\max_{\mathbf{S}\preceq\mathbf{V};\, |\mathrm{supp}(\mathbf{S})|=m} f(\mathbf{S}) = -\mathbb{I}(\otimes_{i=1}^k P^{(S_i)}) = -\sum_{i=1}^k \mathbb{I}(P^{(S_i)}).$$

We are restricting m to be at least $k+1$ following the pigeonhole principle, as we need at least one S_i with $|S_i| > 1$. If $m \leq k$, we can take either $S_i = \{e\}$ or $S_i = \emptyset$ for all $i \in [\![k]\!]$ so that the optimization problem becomes trivial.

In view of Theorem 2.9, we have a monotonically non-decreasing and k-submodular function g given by

$$\begin{aligned}g(\mathbf{S}) &= f(\mathbf{S}) - \beta + \sum_{i=1}^k \sum_{e\in S_i} [H(P^{(V_i\backslash\{e\})}) + H(P^{(e)}) - H(P^{(V_i)})] \\ &= f(\mathbf{S}) - \beta + \sum_{i=1}^k \sum_{e\in S_i} D(P^{(V_i)} \| P^{(V_i\backslash\{e\})} \otimes P^{(e)}). \qquad (31)\end{aligned}$$

We take $\beta = 0$, and denote the following non-negative modular function as

$$c(\mathbf{S}) = \sum_{i=1}^k \sum_{e\in S_i} D(P^{(V_i)} \| P^{(V_i\backslash\{e\})} \otimes P^{(e)}) \qquad (32)$$

so that $f(\mathbf{S}) = g(\mathbf{S}) - c(\mathbf{S})$. By applying Algorithm 3, we can obtain a result with the following lower bound by Theorem 2.16.

Corollary 5.5. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary along with $d, m \geq k+1$ and $\mathbf{V} \in (k+1)^{[\![d]\!]}$. In Algorithm 3, we take g as in (31), c as in (32), and $\mathbf{OPT} = \arg\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})|=m} f(\mathbf{S})$, then by Theorem 2.16, we have the following lower bound*

$$f(\mathbf{S}_m) = -\mathbb{I}(\otimes_{i=1}^k P^{(S_{m,i})}) \geq (1 - e^{-1}) g(\mathbf{OPT}) - c(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1}, \ldots, S_{m,k})$ is the output of Algorithm 3.

In the special case where $k = 1$ and $\mathbf{V} = [\![d]\!]$, we recover Corollary 5.1.

5.3 k-supermodular minimization of distance to independence of the tensorized keep-$V_i \backslash S_i$-in matrices $\mathbb{I}(\otimes_{i=1}^k P^{(V_i \backslash S_i)})$

For given $\mathbf{V} \in (k+1)^{[\![d]\!]}$, we consider the following map in view of Lemma 5.2,

$$\{\mathbf{S} \in (k+1)^{[\![d]\!]};\ \mathbf{S} \preceq \mathbf{V}\} \ni \mathbf{S} = (S_1, \ldots, S_k) \mapsto \mathbb{I}(\otimes_{i=1}^k P^{(V_i \backslash S_i)}) = \sum_{i=1}^k \mathbb{I}(P^{(V_i \backslash S_i)}). \tag{33}$$

We first prove a result concerning monotonicity and k-supermodularity of the map above.

Theorem 5.6. *The map (33) is monotonically non-increasing and k-supermodular.*

Proof. In view of Theorem 2.11, for each component S_i, we take V_i as the ground set, hence $\mathbb{I}(P^{(V_i \backslash S_i)})$ is monotonically non-increasing and supermodular. From Lemma 5.2, this function is the sum of k monotonically non-increasing and supermodular functions. From Lemma 2.6, we conclude that this map is k-supermodular and monotonically non-increasing. □

Therefore, we denote the following monotonically non-decreasing, k-submodular function g as

$$g(\mathbf{S}) = -\mathbb{I}(\otimes_{i=1}^k P^{(V_i \backslash S_i)}) = -\sum_{i=1}^k \mathbb{I}(P^{(V_i \backslash S_i)}). \tag{34}$$

Given $d \geq k+1$, $m \leq d-k-1$, we are interested in the following maximization problem given by

$$\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} g(\mathbf{S}).$$

We are restricting m by $m \leq d-k-1$ following the pigeonhole principle, as we want $|V_i \backslash S_i| \geq 2$ for at least one i. If $m \geq d-k$, we can choose either $V_i \backslash S_i = \{e\}$ or $V_i \backslash S_i = \emptyset$ so that the optimization problem is trivial.

By taking $c = 0$ as a non-negative modular function, we can apply Algorithm 3 to obtain an optimization result with $(1 - e^{-1})$-approximation guarantee.

Corollary 5.7. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary along with $d \geq k+1$, $m \leq d-k-1$ and $\mathbf{V} \in (k+1)^{[\![d]\!]}$. In Algorithm 3, we take g as in (34), $c = 0$ and denote*

$$\mathbf{OPT} = \operatorname*{arg\,max}_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq m} g(\mathbf{S}).$$

From Theorem 2.16, we can obtain the following lower bound

$$g(\mathbf{S}_m) \geq (1 - e^{-1}) g(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1}, \ldots, S_{m,k})$ is the output of Algorithm 3.

6 Supermodular minimization of distance to stationarity $D(P^{(S)}\|\Pi^{(S)})$

In this section, we investigate the following map:

$$2^{[\![d]\!]} \ni S \mapsto D(P^{(S)}\|\Pi^{(S)}), \tag{35}$$

where Π is the matrix of stationary distribution with each row of Π being π. We first show that this map is monotonically non-decreasing.

Lemma 6.1. *The map* (35) *is monotonically non-decreasing.*

Proof. We choose $S \subseteq T \subseteq [\![d]\!]$. By the partition lemma (Theorem 2.1), we have

$$D(P^{(S)}\|\Pi^{(S)}) \leq D(P^{(T)}\|\Pi^{(T)}),$$

and hence this map is monotonically non-decreasing. □

We are interested in the following optimization problem

$$\max_{S\subseteq[\![d]\!];\ |S|=m} D(P^{(S)}\|\Pi^{(S)}),$$

as solving the above can help to identify coordinates which are furthest away from the equilibrium in one step.

To solve this optimization problem with a theoretical guarantee, we recall the batch greedy algorithm (Algorithm 4, see Theorem 7 of (Jagalur-Mohan and Marzouk, 2021)).

Algorithm 4 Batch greedy algorithm

Require: monotonically non-decreasing set function f; ground set U; total cardinality constraint m; number of steps l and cardinality constraints q_i such that $\sum_{i=1}^{l} q_i = m$

1: Initialize $S_0 = \emptyset$
2: **for** $i = 1$ to l **do**
3: Determine incremental gains $f(S_{i-1} \cup \{e\}) - f(S_{i-1})$, $\forall e \in U \backslash S_{i-1}$
4: Find Q, comprising the elements with top-q_i incremental gains
5: $S_i \leftarrow S_{i-1} \cup Q$
6: **end for**
7: **Output**: S_l

It turns out that the theoretical guarantee depends on the supermodularity ratio and submodularity ratio of a set function f, that we shall now briefly recall. The supermodularity ratio of a non-negative set function f (Definition 6 of (Jagalur-Mohan and Marzouk, 2021)) with respect to the set U and a cardinality constraint $m \geq 1$ is

$$\eta_{U,m} := \min_{S\subseteq U;\ T:|T|\leq m, S\cap T=\emptyset} \frac{f(S\cup T) - f(S)}{\sum_{e\in T}[f(S\cup\{e\}) - f(S)]},$$

while the submodularity ratio of f (Definition 32 of (Jagalur-Mohan and Marzouk, 2021)) with respect to the set U and a cardinality constraint $k \geq 1$ is

$$\gamma_{U,m} := \min_{S\subseteq U;\ T:|T|\leq m, S\cap T=\emptyset} \frac{\sum_{e\in T}[f(S\cup\{e\}) - f(S)]}{f(S\cup T) - f(S)}.$$

We then state the lower bound pertaining to Algorithm 4 (see Theorem 7 of (Jagalur-Mohan and Marzouk, 2021)).

Theorem 6.2 (Lower bound for batch greedy algorithm)**.** *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary and U be the ground set. Let f be a monotonically non-decreasing set function with $f(\emptyset) = 0$. Algorithm 4 yields the following lower bound*

$$f(S_l) \geq \left(1 - \prod_{i=1}^{l}\left(1 - \frac{q_i \cdot \eta_{U,q_i} \cdot \gamma_{U,m}}{m}\right)\right) \max_{S\subseteq U;\ |S|=m} f(S),$$

where S_l is the output set of Algorithm 4.

Since we have a monotonically mon-decreasing map (35) with $D(P^{(\emptyset)}\|\Pi^{(\emptyset)}) = 0$, we can apply the Algorithm 4 (see Theorem 7 of (Jagalur-Mohan and Marzouk, 2021)) with the following lower bound.

Corollary 6.3. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary and $U = [\![d]\!]$ be the ground set. Let f be (35) which is a monotonically non-decreasing set function with $f(\emptyset) = 0$. Algorithm 4 yields the following lower bound*

$$f(S_l) \geq \left(1 - \prod_{i=1}^{l}\left(1 - \frac{q_i \cdot \eta_{U,q_i} \cdot \gamma_{U,m}}{m}\right)\right) \max_{S\subseteq[\![d]\!];\ |S|=m} f(S),$$

where S_l is the output set of Algorithm 4.

We now consider the special case where the stationary distribution π is of product form. In this case, we can show the supermodularity of the map (35).

Lemma 6.4. *The map (35) is supermodular if P is π-stationary where π is of product form.*

Proof.

$$\begin{aligned}
D(P^{(S)}\|\Pi^{(S)}) &= \sum_{x^{(S)}}\sum_{y^{(S)}} \pi^{(S)}(x^{(S)})P^{(S)}(x^{(S)},y^{(S)}) \ln \frac{P^{(S)}(x^{(S)},y^{(S)})}{\pi^{(S)}(y^{(S)})} \\
&= -H(P^{(S)}) - \sum_{x^{(S)}}\sum_{y^{(S)}} \pi^{(S)}(x^{(S)})P^{(S)}(x^{(S)},y^{(S)}) \ln \pi^{(S)}(y^{(S)}) \\
&= -H(P^{(S)}) - \sum_{y^{(S)}} \ln \pi^{(S)}(y^{(S)}) \sum_{x^{(S)}} \pi^{(S)}(x^{(S)})P^{(S)}(x^{(S)},y^{(S)}) \\
&= -H(P^{(S)}) + H(\pi^{(S)}).
\end{aligned}$$

The last equation holds since P is π-stationary and hence

$$\pi^{(S)}(y^{(S)}) = \sum_{x^{(S)}} \pi^{(S)}(x^{(S)})P^{(S)}(x^{(S)},y^{(S)}).$$

Since the stationary distribution π is of product form, then $\pi = \otimes_{i=1}^{d}\pi^{(i)}$, hence $H(\pi^{(S)}) = \sum_{i\in S} H(\pi^{(i)})$, which is a modular function. Also, since $H(P^{(S)})$ is submodular, then $-H(P^{(S)})$ is supermodular. Therefore, $D(P^{(S)}\|\Pi^{(S)})$ is supermodular because it is a sum of a supermodular function and a modular function. □

We proceed to investigate the following optimization problem when P is π-stationary with product form π,

$$\max_{S\subseteq[\![d]\!];\ |S|\leq m} f(S) = -D(P^{(S)}\|\Pi^{(S)}).$$

In view of Theorem 2.8, the following function g is monotonically non-decreasing and submodular since f is submodular:

$$\begin{aligned}
g(S) &= f(S) - \beta + \sum_{e\in S}(H(P^{(-e)}) - H(\pi^{(-e)}) - H(P) + H(\pi)) \\
&= f(S) - \beta + \sum_{e\in S}(D(P\|P^{(e)}\otimes P^{(-e)}) + D(P^{(e)}\|\Pi^{(e)})). \qquad (36)
\end{aligned}$$

Choosing $\beta = 0$, we denote the following non-negative modular function as

$$c(S) = \sum_{e\in S}(D(P\|P^{(e)}\otimes P^{(-e)}) + D(P^{(e)}\|\Pi^{(e)})). \qquad (37)$$

Since $f(S) = g(S) - c(S)$, we apply Algorithm 2 to obtain a result with the following lower bound:

Corollary 6.5. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary with π to be product form. In Algorithm 2, we take g as in (36), c as in (37), and $\mathrm{OPT} = \arg\max_{S\subseteq[\![d]\!];\ |S|\leq m} f(S)$. By Theorem 2.13, we have the following lower bound*

$$f(S_m) = -D(P^{(S_m)}\|\Pi^{(S_m)}) \geq (1 - e^{-1})g(\mathrm{OPT}) - c(\mathrm{OPT}),$$

where S_m is the output set of Algorithm 2.

6.1 Supermodular minimization of distance to stationarity of the complement set $D(P^{(-S)}\|\Pi^{(-S)})$

In this section, we shall investigate the following map:

$$2^{[\![d]\!]} \ni S \mapsto D(P^{(-S)}\|\Pi^{(-S)}). \tag{38}$$

Owing to Lemma 6.1, we first see that the map (38) is monotonically non-increasing. In addition, the map (38) is supermodular if P is π-stationary with product form π in view of Lemma 2.4 and Lemma 6.4.

We are interested in the following optimization problem

$$\max_{S\subseteq[\![d]\!];\ |S|\le m} f(S) = -D(P^{(-S)}\|\Pi^{(-S)}),$$

as solving the above allows us to identify coordinates whose complement set is the closest to equilibrium in one step.

Under the assumption of product form π, as the map (38) is monotonically non-increasing and supermodular, f is monotonically non-decreasing and submodular. We apply the heuristic greedy algorithm (Section 4 of (Nemhauser et al., 1978)) to obtain an approximate maximizer along with a $(1-e^{-1})$-approximation guarantee.

6.2 k-supermodular minimization of distance to stationarity of tensorized keep-S_i-in matrices $D(\otimes_{i=1}^k P^{(S_i)}\|\otimes_{i=1}^k \Pi^{(S_i)})$

In this section, for given $\mathbf{V}\in(k+1)^{[\![d]\!]}$, we investigate the following map:

$$(k+1)^{[\![d]\!]} \ni \mathbf{S}=(S_1,\dots,S_k) \mapsto f(\mathbf{S}) = D(\otimes_{i=1}^k P^{(S_i)}\|\otimes_{i=1}^k \Pi^{(S_i)}). \tag{39}$$

We first give an orthant supermodularity result.

Lemma 6.6. *The map* (39) *is orthant supermodular if P is π-stationary where π is of product form.*

Proof. By the chain rule or tensorization property of KL divergence (see Theorem 2.15 and 2.16 of (Polyanskiy and Wu, 2025)), we see that

$$D(\otimes_{i=1}^k P^{(S_i)}\|\otimes_{i=1}^k \Pi^{(S_i)}) = \sum_{i=1}^k D(P^{(S_i)}\|\Pi^{(S_i)}).$$

We now take $\mathbf{S}\preceq\mathbf{T}$ and $e\in[\![d]\!]\backslash T_i$. By (3), we aim to show that $\Delta_{e,i}f(\mathbf{S})\le\Delta_{e,i}f(\mathbf{T})$, which indeed holds since

$$\begin{aligned}\Delta_{e,i}f(\mathbf{S}) &= D(P^{(S_i\cup\{e\})}\|\Pi^{(S_i\cup\{e\})}) - D(P^{(S_i)}\|\Pi^{(S_i)})\\ &\le D(P^{(T_i\cup\{e\})}\|\Pi^{(T_i\cup\{e\})}) - D(P^{(T_i)}\|\Pi^{(T_i)}) = \Delta_{e,i}f(\mathbf{T}),\end{aligned}$$

because $S\mapsto D(P^{(S)}\|\Pi^{(S)})$ is supermodular (see Lemma 6.4). □

We are interested in the following optimization problem

$$\max_{\mathbf{S}\preceq\mathbf{V};\ |\mathrm{supp}(\mathbf{S}_m)|\le m} -f(\mathbf{S}),$$

where f is orthant supermodular.

In view of Theorem 2.9, we have the following monotonically non-decreasing, k-submodular function g:

$$g(\mathbf{S}) = -f(\mathbf{S}) - \beta + \sum_{i=1}^k\sum_{e\in S_i}(D(P^{(V_i)}\|P^{(e)}\otimes P^{(V_i\backslash e)}) + D(P^{(e)}\|\Pi^{(e)})). \tag{40}$$

We take $\beta=0$, and denote the non-negative modular function as

$$c(\mathbf{S}) = \sum_{i=1}^k\sum_{e\in S_i}(D(P^{(V_i)}\|P^{(e)}\otimes P^{(V_i\backslash e)}) + D(P^{(e)}\|\Pi^{(e)})). \tag{41}$$

Since $-f(\mathbf{S}) = g(\mathbf{S}) - c(\mathbf{S})$, we apply Algorithm 3 to obtain an approximate maximizer along with a lower bound.

Corollary 6.7. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary with π be of product form and $\mathbf{V} \in (k+1)^{[\![d]\!]}$. In Algorithm 3, we take g as in (40), c as in (41), and $\mathbf{OPT} = \arg\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S}_m)| \le m} -f(\mathbf{S})$. Then Theorem 2.16 gives the following lower bound*

$$-f(\mathbf{S}_m) \ge (1-e^{-1})g(\mathbf{OPT}) - c(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1}, \ldots, S_{m,k})$ is the output of Algorithm 3.

6.3 k-supermodular minimization of distance to stationarity of tensorized keep-$V_i \backslash S_i$-in matrices $D(\otimes_{i=1}^k P^{(V_i \backslash S_i)} \| \otimes_{i=1}^k \Pi^{(V_i \backslash S_i)})$

For given $\mathbf{V} \in (k+1)^{[\![d]\!]}$, we investigate the following map:

$$\{\mathbf{S} \in (k+1)^{[\![d]\!]};\ \mathbf{S} \preceq \mathbf{V}\} \ni \mathbf{S} = (S_1, \ldots, S_k) \mapsto D(\otimes_{i=1}^k P^{(V_i \backslash S_i)} \| \otimes_{i=1}^k \Pi^{(V_i \backslash S_i)}). \tag{42}$$

Theorem 6.8. *The map (42) is monotonically non-increasing and k-supermodular if P is π-stationary where π is of product form.*

Proof. By the chain rule or tensorization property of KL divergence (see Theorem 2.15 and 2.16 of (Polyanskiy and Wu, 2025)), we see that

$$D(\otimes_{i=1}^k P^{(V_i \backslash S_i)} \| \otimes_{i=1}^k \Pi^{(V_i \backslash S_i)}) = \sum_{i=1}^k D(P^{(V_i \backslash S_i)} \| \Pi^{(V_i \backslash S_i)}),$$

which is a sum of k monotonically non-increasing and supermodular functions in view of Lemma 2.6. □

We are interested in the following optimization problem

$$\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \le m} g(\mathbf{S}) = -D(\otimes_{i=1}^k P^{(V_i \backslash S_i)} \| \otimes_{i=1}^k \Pi^{(V_i \backslash S_i)}). \tag{43}$$

Since the map (42) is monotonically non-increasing and k-supermodular, then g is monotonically non-decreasing and k-submodular. We apply Algorithm 3 to obtain a $(1-e^{-1})$-approximation guarantee.

Corollary 6.9. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary with product form π and $\mathbf{V} \in (k+1)^{[\![d]\!]}$. We take g as in (43), $c = 0$ and $\mathbf{OPT} = \arg\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \le m} g(\mathbf{S})$. According to Theorem 2.16, we have the following lower bound for Algorithm 3*

$$g(\mathbf{S}_m) \ge (1-e^{-1})g(\mathbf{OPT}),$$

where $\mathbf{S}_m = (S_{m,1}, \ldots, S_{m,k})$ is the output of Algorithm 3.

In the special case where $k = 1$ and $\mathbf{V} = [\![d]\!]$, the above Corollary reduces to the $(1-e^{-1})$-approximation guarantee as in Section 6.1.

7 Distance to factorizability over a fixed set $D(P^{(W \cup S)} \| P^{(W)} \otimes P^{(S)})$

We fix a set $W \subseteq [\![d]\!]$ and investigate the following function:

$$\{S \subseteq [\![d]\!];\ S \cap W = \emptyset\} \ni S \mapsto f(S) = D(P^{(W \cup S)} \| P^{(W)} \otimes P^{(S)}). \tag{44}$$

We shall investigate the following optimization problem with cardinality constraint

$$\max_{S \subseteq [\![d]\!];\ S \cap W = \emptyset;\ |S| = m} f(S).$$

We pick $S, T \subseteq \{S \subseteq [\![d]\!];\ S \cap W = \emptyset\}$ with $S \subseteq T$ and compute that

$$f(S) - f(T) = [H(P^{(T \cup W)}) - H(P^{(T)})] - [H(P^{(S \cup W)}) - H(P^{(S)})] \le 0,$$

where the inequality follows from the property that $S \mapsto H(P^{(S)})$ is submodular (see Theorem 2.10). Therefore f is monotonically non-decreasing. Also, $f(\emptyset) = D(P^{(W)} \| P^{(W)} \otimes P^{(\emptyset)}) = 0$. As such, we can apply Algorithm 4 (see Theorem 6.2) with a lower bound.

Corollary 7.1. *Let $P \in \mathcal{L}(\mathcal{X})$ be π-stationary, $W \subseteq [\![d]\!]$, and $U = [\![d]\!] \backslash W$ be the ground set. Let f be (44) which is a monotonically non-decreasing set function with $f(\emptyset) = 0$. Algorithm 4 yields the following lower bound*

$$f(S_l) \geq \left(1 - \prod_{i=1}^{l} \left(1 - \frac{q_i \cdot \eta_{U,q_i} \cdot \gamma_{U,m}}{m}\right)\right) \max_{S \subseteq [\![d]\!];\ S \cap W = \emptyset;\ |S| = m} f(S),$$

where S_l is the output set of Algorithm 4.

8 Numerical experiments of Part I[1]

We conduct a case study to evaluate the numerical performance of the submodular optimization algorithms on the information-theoretic properties of multivariate Markov chains. We conduct numerical experiments on the 10-dimensional Markov chains ($d = 10$) associated with the Curie-Weiss model and the Bernoulli-Laplace level model (see Section 2.4 for details). For the Curie-Weiss model, we choose $T = 10$ as the temperature, $h = 1$ as the external magnetic field. For the Bernoulli-Laplace level model, we choose the swapping size $s = 1$. For the numerical experiments of the generalized distorted greedy algorithm (Algorithm 3), we choose $k = 3$ and the ground set $\mathbf{V} = (V_1, V_2, V_3)$, where $V_1 = \{1, 2, 3, 4\}$, $V_2 = \{5, 6, 7\}$ and $V_3 = \{8, 9, 10\}$.

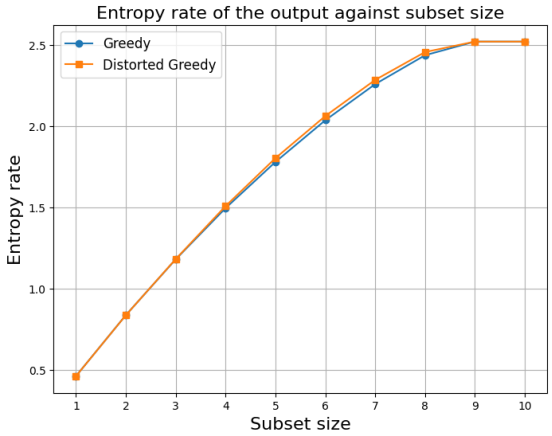
8.1 Experiment results of Section 3

In this section, we report the numerical experiment results related to Section 3, which contains the performance of the heuristic greedy algorithm (see Section 4 of (Nemhauser et al., 1978)), the distorted greedy algorithm (see Corollary 3.2), and the generalized distorted greedy algorithm (see Corollary 3.5) on the Bernoulli-Laplace level model and the Curie-Weiss model. For each experiment, we conduct submodular optimization with cardinality constraint m, with m ranging from 1 to 10.

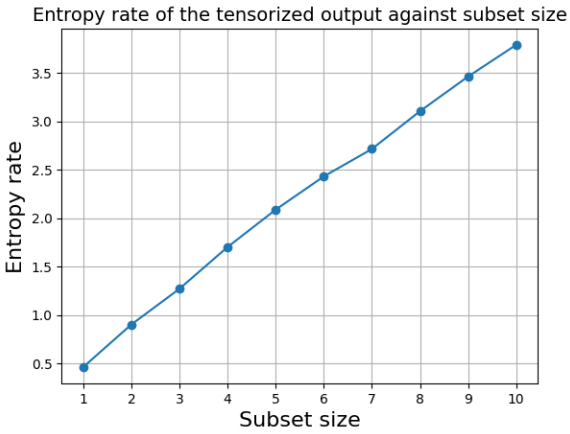
	Greedy		**Distorted Greedy**	
m	Subset S_m	$H\big(P^{(S_m)}\big)$	Subset S_m	$H\big(P^{(S_m)}\big)$
1	{10}	0.46094	{10}	0.46094
2	{3, 10}	0.83616	{1, 10}	0.83573
3	{1, 3, 10}	1.17940	{1, 2, 5}	1.18116
4	{1, 2, 3, 10}	1.49461	{1, 2, 3, 5}	1.50706
5	{1, 2, 3, 4, 10}	1.77855	{1, 2, 3, 4, 5}	1.80193
6	{1, 2, 3, 4, 5, 10}	2.03516	{1, 2, 3, 4, 5, 6}	2.06105
7	{1, 2, 3, 4, 5, 6, 10}	2.25729	{1, 2, 3, 4, 5, 6, 7}	2.28328
8	{1, 2, 3, 4, 5, 6, 7, 10}	2.43498	{1, 2, 3, 4, 5, 6, 7, 8}	2.45453
9	{1, 2, 3, 4, 5, 6, 7, 8, 10}	2.51897	{1, 2, 3, 4, 5, 6, 7, 8, 10}	2.51897
10	{1, 2, 3, 4, 5, 6, 7, 8, 10}	2.51897	{1, 2, 3, 4, 5, 6, 7, 8, 10}	2.51897

Table 1: Comparison of the greedy algorithm and the distorted greedy algorithm. Entropy rate of the full chain of the Bernoulli-Laplace level model is $H(P) = 1.96068$.

[1]The code is available at: `https://github.com/zheyuanlai/SubmodOptMC`.



(a) Greedy and Algorithm 2



(b) Algorithm 3

Figure 1: Entropy rate against subset size for the three algorithms (B-L model).

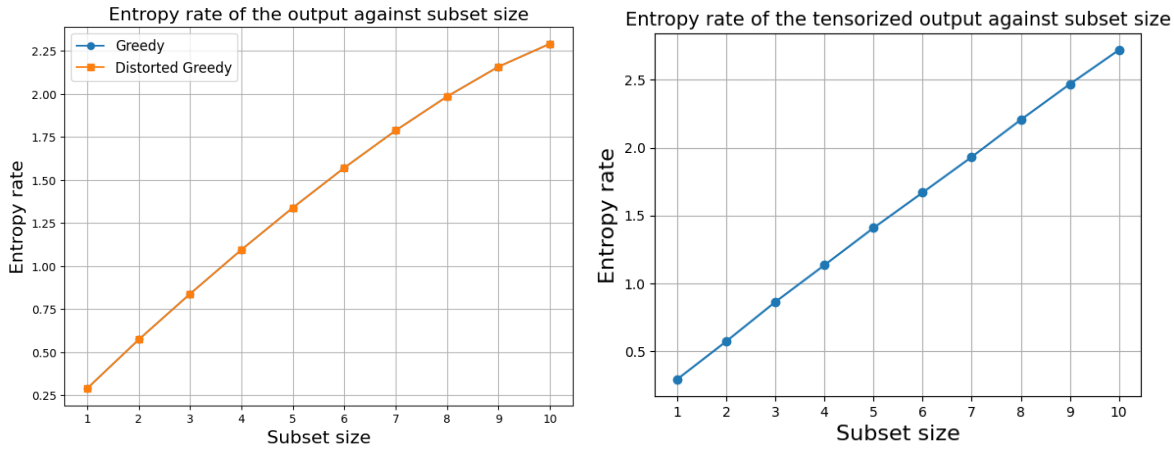
Cardinality m	**Subset** $S_{m,1}$	**Subset** $S_{m,2}$	**Subset** $S_{m,3}$	$H(\otimes_{i=1}^{3} P^{(S_{m,i})})$
1	$\emptyset$	$\emptyset$	{10}	0.46094
2	$\emptyset$	{7}	{10}	0.90046
3	$\emptyset$	{7}	{8, 9}	1.26966
4	{4}	{7}	{8, 9}	1.70072
5	{4}	{5, 7}	{8, 9}	2.08692
6	{4}	{5, 6, 7}	{8, 9}	2.43035
7	{4}	{5, 6, 7}	{8, 9, 10}	2.71405
8	{3, 4}	{5, 6, 7}	{8, 9, 10}	3.10451
9	{1, 2, 4}	{5, 6, 7}	{8, 9, 10}	3.46267
10	{1, 2, 3, 4}	{5, 6, 7}	{8, 9, 10}	3.78968

Table 2: Performance evaluation of the generalized distorted greedy algorithm. Entropy rate of the full chain of the Bernoulli-Laplace level model is $H(P) = 1.96068$.

For the Bernoulli-Laplace level model, Table 1 and Figure 1a show the entropy rates of the output of the greedy algorithm and the distorted greedy algorithm (Algorithm 2); Table 2 and Figure 1b show the entropy rates of the tensorized output of the generalized distorted greedy algorithm (Algorithm 3).

	Greedy		**Distorted Greedy**	
m	Subset S_m	$H\big(P^{(S_m)}\big)$	Subset S_m	$H\big(P^{(S_m)}\big)$
1	{1}	0.29085	{1}	0.29085
2	{1, 10}	0.57371	{1, 10}	0.57371
3	{1, 9, 10}	0.83933	{1, 9, 10}	0.83933
4	{1, 2, 9, 10}	1.09570	{1, 2, 9, 10}	1.09570
5	{1, 2, 6, 9, 10}	1.33953	{1, 2, 6, 9, 10}	1.33953
6	{1, 2, 4, 6, 9, 10}	1.57098	{1, 2, 4, 6, 9, 10}	1.57098
7	{1, 2, 4, 6, 8, 9, 10}	1.78757	{1, 2, 4, 6, 8, 9, 10}	1.78757
8	{1, 2, 3, 4, 6, 8, 9, 10}	1.98500	{1, 2, 3, 4, 6, 7, 9, 10}	1.98458
9	{1, 2, 3, 4, 6, 7, 8, 9, 10}	2.15793	{1, 2, 3, 4, 6, 7, 8, 9, 10}	2.15793
10	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	2.29109	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	2.29109

Table 3: Comparison of the greedy algorithm and the distorted greedy algorithm. Entropy rate of the full chain of the Curie-Weiss model is $H(P) = 2.29109$.



(a) Greedy and Algorithm 2

(b) Algorithm 3

Figure 2: Entropy rate against subset size for the three algorithms (C-W model).

Cardinality m	**Subset** $S_{m,1}$	**Subset** $S_{m,2}$	**Subset** $S_{m,3}$	$H(\otimes_{i=1}^{3} P^{(S_{m,i})})$
1	$\{1\}$	$\emptyset$	$\emptyset$	0.29085
2	$\{1\}$	$\{7\}$	$\emptyset$	0.57067
3	$\{1\}$	$\{7\}$	$\{10\}$	0.86152
4	$\{1\}$	$\{5,7\}$	$\{10\}$	1.13316
5	$\{1\}$	$\{5,7\}$	$\{9,10\}$	1.40732
6	$\{1\}$	$\{5,6,7\}$	$\{9,10\}$	1.66816
7	$\{1\}$	$\{5,6,7\}$	$\{8,9,10\}$	1.93090
8	$\{1,2\}$	$\{5,6,7\}$	$\{8,9,10\}$	2.20505
9	$\{1,2,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	2.46832
10	$\{1,2,3,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	2.72011

Table 4: Performance evaluation of the generalized distorted greedy algorithm. Entropy rate of the full chain of the Curie-Weiss model is $H(P) = 2.29109$.

For the Curie-Weiss model, Table 3 and Figure 2a show the entropy rates of the output of the greedy algorithm and the distorted greedy algorithm (Algorithm 2); Table 4 and Figure 2b show the entropy rates of the tensorized output of the generalized distorted greedy algorithm (Algorithm 3).

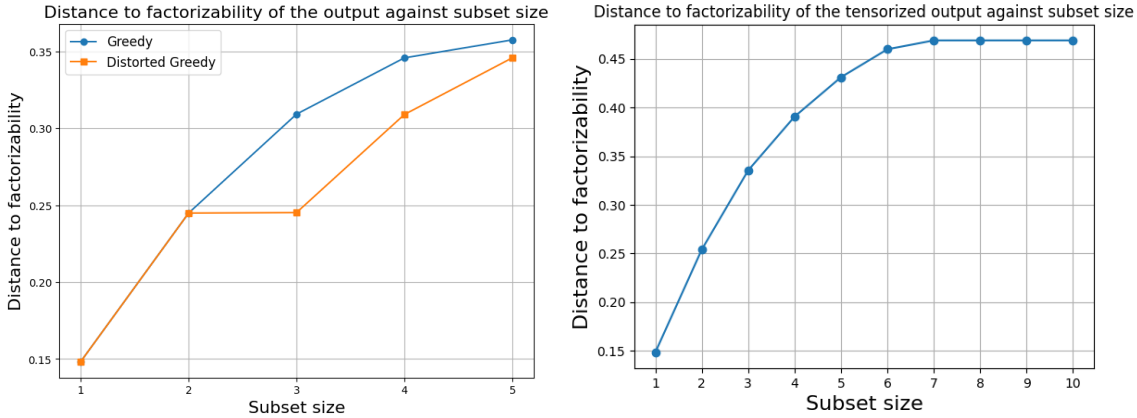
Notably, in Table 1 and Figure 1a, the distorted greedy algorithm outperforms the heuristic greedy algorithm when the cardinality constraint equals to $m = 3, 4, 5, 6, 7, 8$. This is because, in the distorted greedy algorithm, the distortion term $(1 - \frac{1}{m})^{m-(i+1)}$ at each step is different with different cardinality constraint m, which results in possibly better or different results than the heuristic greedy algorithm. However, the distorted greedy algorithm does not necessarily select better subset than the heuristic greedy algorithm, see the example of $m = 2$ in Table 1 and $m = 8$ in Table 3.

8.2 Experiment results of Section 4

We report the numerical experiment results related to Section 4, which contains the performance of the heuristic greedy algorithm (Section 4 of (Nemhauser et al., 1978)), the distorted greedy algorithm (Algorithm 2), and the generalized distorted greedy algorithm (Algorithm 3) on the Curie-Weiss model.

	Greedy		Distorted Greedy	
m	Subset S_m	$D\left(P\|P^{(S_m)} \otimes P^{(-S_m)}\right)$	Subset S_m	$D\left(P\|P^{(S_m)} \otimes P^{(-S_m)}\right)$
1	$\{6\}$	0.14837	$\{6\}$	0.14837
2	$\{2, 6\}$	0.24497	$\{3, 10\}$	0.24496
3	$\{2, 6, 9\}$	0.30927	$\{3, 7\}$	0.24525
4	$\{2, 5, 6, 9\}$	0.34590	$\{2, 7, 10\}$	0.30905
5	$\{2, 3, 5, 6, 9\}$	0.35758	$\{2, 3, 6, 10\}$	0.34590

Table 5: Comparison of the greedy algorithm and the distorted greedy algorithm.



(a) Greedy and Algorithm 2

(b) Algorithm 3

Figure 3: Distance to factorizability against subset size for the three algorithms.

For the experiments related to heuristic greedy and distorted greedy algorithms, since the map $S \mapsto D(P\|P^{(S)} \otimes P^{(-S)})$ is symmetric, we conduct submodular maximization with cardinality constraint m, with m only ranging from 1 to 5. The results are shown on Table 5 and Figure 3a. These results show that although the distorted greedy algorithm has a lower bound as detailed in Corollary 4.1, the performance is not guaranteed to be better than the heuristic greedy algorithm. We also conduct the generalized distorted greedy algorithm as detailed in Corollary 4.3 with cardinality constraint m ranging from 1 to 10, and the results are shown on Table 6 and Figure 3b.

m	**Subset** $S_{m,1}$	**Subset** $S_{m,2}$	**Subset** $S_{m,3}$	$D\left(P\|\left(\otimes_{i=1}^{3} P^{(S_{m,i})}\right) \otimes P^{(-\cup_{i=1}^{3} S_{m,i})}\right)$
1	$\emptyset$	$\{6\}$	$\emptyset$	0.14836
2	$\emptyset$	$\{7\}$	$\{8\}$	0.25388
3	$\{4\}$	$\{7\}$	$\{8\}$	0.33529
4	$\{4\}$	$\{5, 7\}$	$\{8\}$	0.39056
5	$\{2, 4\}$	$\{5, 7\}$	$\{8\}$	0.43104
6	$\{2, 4\}$	$\{5, 7\}$	$\{8, 10\}$	0.45978
7	$\{2, 4\}$	$\{5, 6, 7\}$	$\{8, 10\}$	0.46887
8	$\{2, 4\}$	$\{5, 6, 7\}$	$\{8, 10\}$	0.46887
9	$\{2, 4\}$	$\{5, 6, 7\}$	$\{8, 10\}$	0.46887
10	$\{2, 4\}$	$\{5, 6, 7\}$	$\{8, 10\}$	0.46887

Table 6: Performance evaluation of the generalized distorted greedy algorithm.

We conduct similar numerical experiments on the Bernoulli-Laplace level model. Among all cardinality constraints, the greedy algorithm and the distorted greedy algorithm output $S_m = \{10\}$, and the generalized distorted greedy algorithm outputs $S_{m,1} = S_{m,2} = \emptyset$, $S_{m,3} = \{10\}$. The reason behind it is that for a 10-dimensional Markov chain, the coordinate 10 is "far" from other coordinates.

8.3 Experiment results of Section 5

We report the numerical experiment results related to Section 5, which contains the performance of the heuristic greedy algorithm (see Section 4 of (Nemhauser et al., 1978)), the distorted greedy algorithm (see Corollary 5.1), and the generalized distorted greedy algorithm (see Corollary 5.5) on the Bernoulli-Laplace level model and the Curie-Weiss model. For each experiment, we conduct supermodular minimization with different cardinality constraint m's.

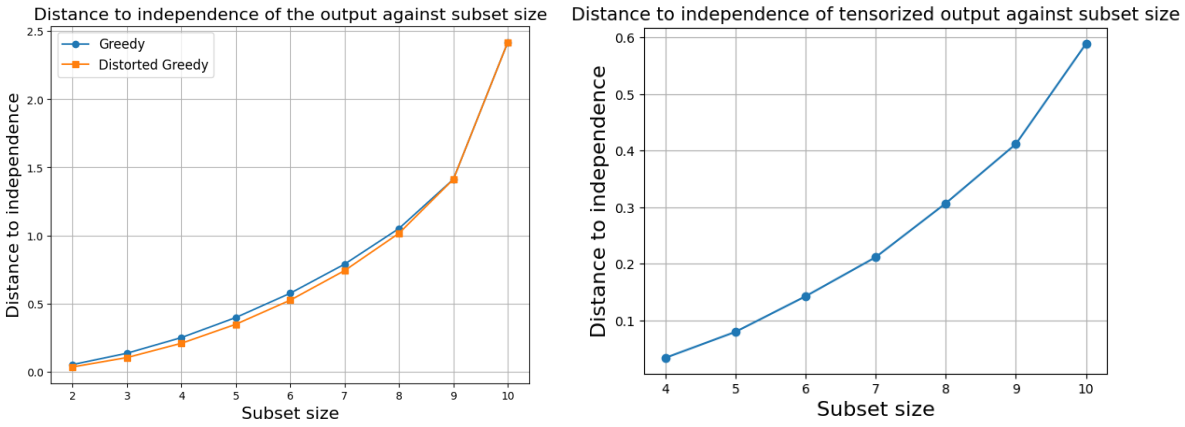
For the Bernoulli-Laplace level model, Table 7 and Figure 4a show the distance to independence of the outputs of the greedy algorithm and the distorted greedy algorithm (Algorithm 2). We note that the distorted greedy algorithm often outperforms the greedy algorithm. Table 8 and Figure 4b show the distance to independence of the tensorized outputs of the generalized distorted greedy algorithm (Algorithm 3).

	Greedy		**Distorted Greedy**	
m	Subset S_m	$\mathbb{I}(P^{(S_m)})$	Subset S_m	$\mathbb{I}(P^{(S_m)})$
2	{1, 10}	0.05140	{1, 2}	0.03406
3	{1, 2, 10}	0.13505	{1, 2, 3}	0.10318
4	{1, 2, 3, 10}	0.24989	{1, 2, 3, 4}	0.20793
5	{1, 2, 3, 4, 10}	0.39701	{1, 2, 3, 4, 5}	0.34753
6	{1, 2, 3, 4, 5, 10}	0.57523	{1, 2, 3, 4, 5, 6}	0.52441
7	{1, 2, 3, 4, 5, 6, 10}	0.78911	{1, 2, 3, 4, 5, 6, 7}	0.74171
8	{1, 2, 3, 4, 5, 6, 7, 10}	1.05094	{1, 2, 3, 4, 5, 6, 7, 8}	1.01576
9	{1, 2, 3, 4, 5, 6, 7, 8, 10}	1.41226	{1, 2, 3, 4, 5, 6, 7, 8, 10}	1.41226
10	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	2.41825	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	2.41825

Table 7: Comparison of the greedy algorithm and the distorted greedy algorithm (B-L model).

m	**Subset** $S_{m,1}$	**Subset** $S_{m,2}$	**Subset** $S_{m,3}$	$\mathbb{I}\left(\otimes_{i=1}^{3} P^{(S_{m,i})}\right)$
4	{1, 2}	{5}	{8}	0.03406
5	{1, 2}	{5, 6}	{8}	0.07999
6	{1, 2}	{5, 6}	{8, 9}	0.14286
7	{1, 2, 3}	{5, 6}	{8, 9}	0.21199
8	{1, 2, 3}	{5, 6, 7}	{8, 9}	0.30727
9	{1, 2, 3, 4}	{5, 6, 7}	{8, 9}	0.41202
10	{1, 2, 3, 4}	{5, 6, 7}	{8, 9, 10}	0.58925

Table 8: Performance evaluation of the generalized distorted greedy algorithm (B-L model).



(a) Greedy and Algorithm 2

(b) Algorithm 3

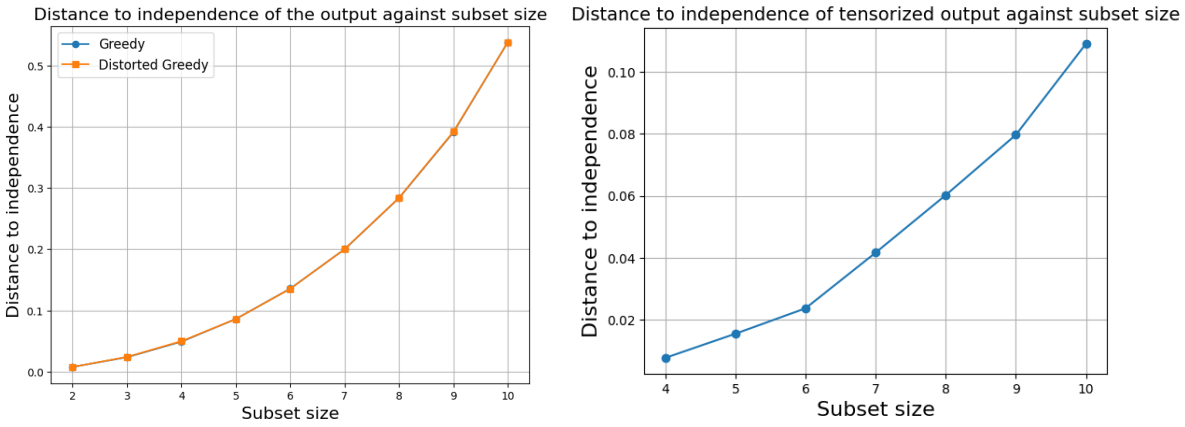
Figure 4: Distance to independence against subset size for the three algorithms (B-L model).

	Greedy		Distorted Greedy	
m	Subset S_m	$\mathbb{I}(P^{(S_m)})$	Subset S_m	$\mathbb{I}(P^{(S_m)})$
2	{4, 10}	0.00757	{1, 7}	0.00757
3	{4, 7, 10}	0.02350	{1, 6, 10}	0.02398
4	{2, 4, 7, 10}	0.04889	{1, 5, 7, 10}	0.04961
5	{2, 4, 6, 7, 10}	0.08592	{1, 3, 5, 7, 10}	0.08591
6	{2, 4, 6, 7, 8, 10}	0.13555	{1, 3, 5, 7, 8, 10}	0.13533
7	{2, 3, 4, 6, 7, 8, 10}	0.19989	{1, 3, 4, 5, 7, 8, 10}	0.20017
8	{2, 3, 4, 5, 6, 7, 8, 10}	0.28356	{1, 3, 4, 5, 6, 7, 8, 10}	0.28399
9	{2, 3, 4, 5, 6, 7, 8, 9, 10}	0.39102	{1, 3, 4, 5, 6, 7, 8, 9, 10}	0.39191
10	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	0.53813	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	0.53813

Table 9: Comparison of the greedy algorithm and the distorted greedy algorithm (C-W model).

m	**Subset** $S_{m,1}$	**Subset** $S_{m,2}$	**Subset** $S_{m,3}$	$\mathbb{I}\left(\otimes_{i=1}^{3} P^{(S_{m,i})}\right)$
4	{1}	{5, 7}	{8}	0.00778
5	{1, 4}	{5, 7}	{8}	0.01556
6	{1, 4}	{5, 7}	{8, 10}	0.02376
7	{1, 3, 4}	{5, 7}	{8, 10}	0.04172
8	{1, 3, 4}	{5, 6, 7}	{8, 10}	0.06029
9	{1, 3, 4}	{5, 6, 7}	{8, 9, 10}	0.07972
10	{1, 2, 3, 4}	{5, 6, 7}	{8, 9, 10}	0.10911

Table 10: Performance evaluation of the generalized distorted greedy algorithm (C-W model).



(a) Greedy and Algorithm 2

(b) Algorithm 3

Figure 5: Distance to independence against subset size for the three algorithms (C-W model).

For the Curie-Weiss model, Table 9 and Figure 5a show the distance of independence of the outputs of the greedy algorithm and the distorted greedy algorithm (Algorithm 2), in which these two algorithms output similar results. Table 10 and Figure 5b show the distance of independence of the tensorized outputs of the generalized distorted greedy algorithm (Algorithm 3).

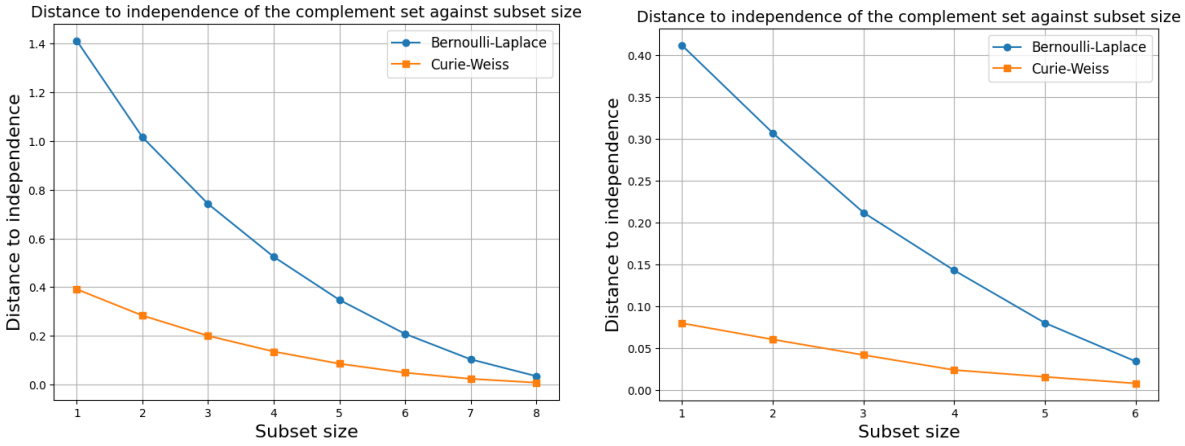
In addition, we report the numerical experiment results related to the distance to independence of the complement set, as detailed in Section 5.1 and Section 5.3. The performance of the greedy algorithm on the two models is shown in Table 11 and Figure 6a, while the performance of the generalized distorted greedy algorithm can be seen from Table 12 and Figure 6b.

	Bernoulli-Laplace		Curie-Weiss	
m	**Subset** S_m	$\mathbb{I}(P^{(-S_m)})$	**Subset** S_m	$\mathbb{I}(P^{(-S_m)})$
1	{9}	1.41226	{1}	0.39102
2	{9, 10}	1.01576	{1, 10}	0.28314
3	{8, 9, 10}	0.74171	{1, 5, 10}	0.19981
4	{7, 8, 9, 10}	0.52441	{1, 5, 7, 10}	0.13517
5	{6, 7, 8, 9, 10}	0.34753	{1, 3, 5, 7, 10}	0.08523
6	{5, 6, 7, 8, 9, 10}	0.20793	{1, 3, 5, 7, 8, 10}	0.04845
7	{4, 5, 6, 7, 8, 9, 10}	0.10318	{1, 3, 4, 5, 7, 8, 10}	0.02304
8	{3, 4, 5, 6, 7, 8, 9, 10}	0.03406	{1, 3, 4, 5, 7, 8, 9, 10}	0.00736

Table 11: Performance evaluation of greedy algorithm.

	Bernoulli-Laplace				Curie-Weiss			
m	$S_{m,1}$	$S_{m,2}$	$S_{m,3}$	$\mathbb{I}\left(\otimes_{i=1}^{3} P^{(-S_{m,i})}\right)$	$S_{m,1}$	$S_{m,2}$	$S_{m,3}$	$\mathbb{I}\left(\otimes_{i=1}^{3} P^{(-S_{m,i})}\right)$
1	$\emptyset$	$\emptyset$	{10}	0.41202	{2}	$\emptyset$	$\emptyset$	0.07972
2	{4}	$\emptyset$	{10}	0.30727	{2}	$\emptyset$	{9}	0.06029
3	{4}	{7}	{10}	0.21198	{2}	{6}	{9}	0.04172
4	{3, 4}	{7}	{10}	0.14286	{2, 3}	{6}	{9}	0.02376
5	{3, 4}	{7}	{9, 10}	0.07999	{2, 3}	{6}	{9, 10}	0.01556
6	{3, 4}	{5, 7}	{9, 10}	0.03406	{1, 2, 3}	{6}	{9, 10}	0.00778

Table 12: Performance evaluation of the generalized distorted greedy algorithm.



(a) Greedy

(b) Algorithm 3

Figure 6: Distance to independence of the complement set against subset size.

8.4 Experiment results of Section 6

We first report the numerical experiment results related to Algorithm 4. For both the Bernoulli-Laplace level model and the Curie-Weiss model, we consider the following two configurations of the batch greedy algorithm to maximize $D(P^{(S)}\|\Pi^{(S)})$ subject to the cardinality constraint m:

- Approach 1: $l = m$ and $q_i = 1$ for $i \in [\![l]\!]$;
- Approach 2: $l = \lceil \frac{m}{2} \rceil$, $q_i = 2$ for $i \in [\![l-1]\!]$; $q_l = 2$ if m is even, $q_l = 1$ if m is odd.

In Approach 1, we recover the heuristic greedy algorithm since we are adding one element per iteration. We compare the performance of Approach 1 and Approach 2 for both models, and the results are shown in Table 13 and Table 14. Although the stationary distribution π of the Bernoulli-Laplace level model and the Curie-Weiss model are not of product form, we still apply the heuristic distorted greedy algorithm

as in Corollary 6.5, and the results are summarized in Table 15. The comparison of these algorithms on the two models is shown in Figure 7.

From these results, one can conclude that the performance of Approach 1 is slightly better than Approach 2, and the performance of the distorted greedy algorithm is the worst among the three approaches.

	Approach 1		**Approach 2**	
m	**Subset** S_l	$D(P^{(S_l)}\|\|\Pi^{(S_l)})$	**Subset** S_l	$D(P^{(S_l)}\|\|\Pi^{(S_l)})$
1	$\{1\}$	0.26693	$\{1\}$	0.26693
2	$\{1,2\}$	0.59421	$\{1,2\}$	0.59421
3	$\{1,2,7\}$	0.98856	$\{1,2,7\}$	0.98856
4	$\{1,2,7,10\}$	1.47330	$\{1,2,4,7\}$	1.46082
5	$\{1,2,7,9,10\}$	2.07889	$\{1,2,4,7,10\}$	2.03226
6	$\{1,2,7,8,9,10\}$	2.85834	$\{1,2,4,7,9,10\}$	2.73225
7	$\{1,2,6,7,8,9,10\}$	3.70196	$\{1,2,4,7,8,9,10\}$	3.64286
8	$\{1,2,5,6,7,8,9,10\}$	4.69790	$\{1,2,4,6,7,8,9,10\}$	4.65621
9	$\{1,2,4,5,6,7,8,9,10\}$	5.91911	$\{1,2,4,5,6,7,8,9,10\}$	5.91911
10	$\{1,2,3,4,5,6,7,8,9,10\}$	7.56130	$\{1,2,3,4,5,6,7,8,9,10\}$	7.56130

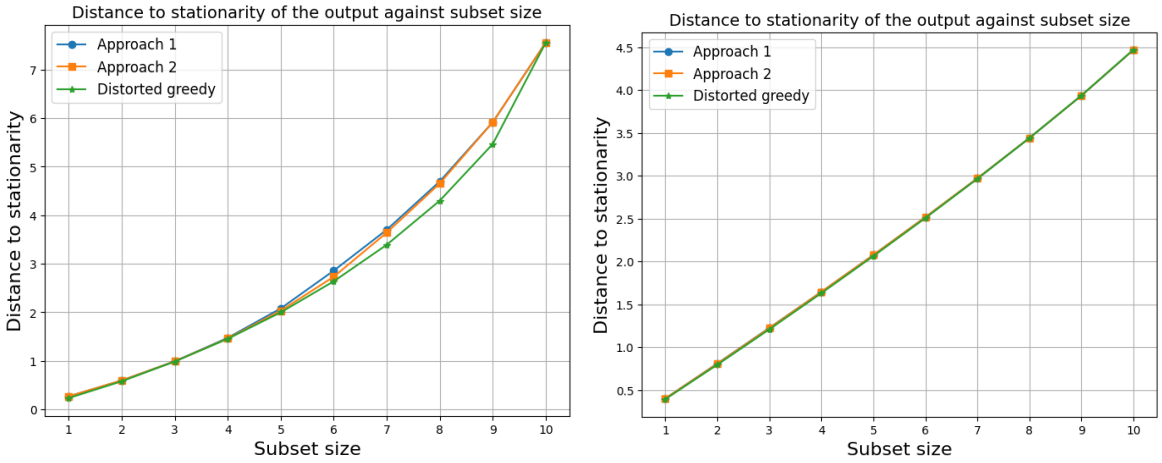
Table 13: Comparison of different configurations of the batch greedy algorithm (B-L model).

	Approach 1		**Approach 2**	
m	**Subset** S_l	$D(P^{(S_l)}\|\|\Pi^{(S_l)})$	**Subset** S_l	$D(P^{(S_l)}\|\|\Pi^{(S_l)})$
1	{6}	0.40245	{6}	0.40245
2	{3, 6}	0.81082	{5, 6}	0.80739
3	{3, 6, 8}	1.22606	{5, 6, 8}	1.22234
4	{3, 4, 6, 8}	1.64626	{3, 5, 6, 8}	1.64615
5	{3, 4, 6, 8, 9}	2.07613	{2, 3, 5, 6, 8}	2.07601
6	{2, 3, 4, 6, 8, 9}	2.51741	{2, 3, 5, 6, 8, 9}	2.51771
7	{2, 3, 4, 5, 6, 8, 9}	2.97051	{2, 3, 4, 5, 6, 8, 9}	2.97051
8	{1, 2, 3, 4, 6, 8, 9}	3.44141	{2, 3, 4, 5, 6, 7, 8, 9}	3.44085
9	{1, 2, 3, 4, 6, 8, 9, 10}	3.93647	{1, 2, 3, 4, 5, 6, 7, 8, 9}	3.93568
10	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	4.46975	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	4.46975

Table 14: Comparison of different configurations of the batch greedy algorithm (C-W model).

	Bernoulli-Laplace level model		**Curie-Weiss model**	
m	**Subset** S_m	$D(P^{(S_m)}\|\|\Pi^{(S_m)})$	**Subset** S_m	$D(P^{(S_m)}\|\|\Pi^{(S_m)})$
1	{10}	0.23219	{1}	0.39435
2	{1, 10}	0.57719	{1, 10}	0.79669
3	{1, 2, 10}	0.98552	{1, 2, 10}	1.20915
4	{1, 2, 3, 5}	1.45314	{1, 2, 9, 10}	1.63086
5	{1, 2, 3, 4, 5}	1.99871	{1, 2, 3, 9, 10}	2.06307
6	{1, 2, 3, 4, 5, 6}	2.63821	{1, 2, 3, 8, 9, 10}	2.50704
7	{1, 2, 3, 4, 5, 6, 7}	3.39168	{1, 2, 3, 4, 8, 9, 10}	2.96498
8	{1, 2, 3, 4, 5, 6, 7, 8}	4.30094	{1, 2, 3, 4, 5, 8, 9, 10}	3.43971
9	{1, 2, 3, 4, 5, 6, 7, 8, 10}	5.46950	{1, 2, 3, 4, 5, 6, 8, 9, 10}	3.93647
10	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	7.56130	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	4.46975

Table 15: Performance evaluation of the distorted greedy algorithm.



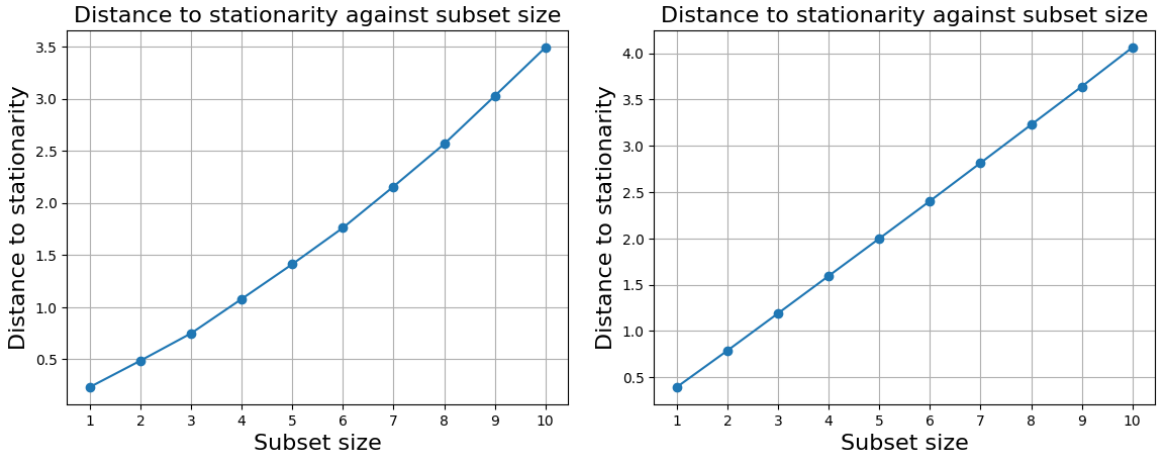
(a) Bernoulli-Laplace level model (b) Curie-Weiss model

Figure 7: Distance to stationarity of the output against subset size.

We then report the numerical experiment results in Section 6.2, see Table 16 and Figure 8. Note that since the stationary distributions of the Bernoulli-Laplace level model and the Curie-Weiss model are not of product form, these simulations are heuristic in nature, as Corollary 6.7 does not provide a theoretical guarantee in this setting.

	Bernoulli-Laplace level model				**Curie-Weiss model**			
m	$S_{m,1}$	$S_{m,2}$	$S_{m,3}$	Value	$S_{m,1}$	$S_{m,2}$	$S_{m,3}$	Value
1	$\emptyset$	$\emptyset$	$\{10\}$	0.23191	$\{1\}$	$\emptyset$	$\emptyset$	0.39436
2	$\emptyset$	$\{7\}$	$\{10\}$	0.48566	$\{1\}$	$\emptyset$	$\{10\}$	0.78871
3	$\{4\}$	$\{7\}$	$\{10\}$	0.74787	$\{1\}$	$\{7\}$	$\{10\}$	1.19100
4	$\{3,4\}$	$\{7\}$	$\{10\}$	1.07820	$\{1\}$	$\{7\}$	$\{9,10\}$	1.59492
5	$\{3,4\}$	$\{5,7\}$	$\{10\}$	1.41218	$\{1,2\}$	$\{7\}$	$\{9,10\}$	1.99886
6	$\{3,4\}$	$\{5,7\}$	$\{8,10\}$	1.76157	$\{1,2\}$	$\{6,7\}$	$\{9,10\}$	2.40381
7	$\{1,3,4\}$	$\{5,7\}$	$\{8,10\}$	2.15778	$\{1,2\}$	$\{5,6,7\}$	$\{9,10\}$	2.81582
8	$\{1,3,4\}$	$\{5,6,7\}$	$\{8,10\}$	2.56632	$\{1,2,3\}$	$\{5,6,7\}$	$\{9,10\}$	3.22828
9	$\{1,3,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	3.02745	$\{1,2,3\}$	$\{5,6,7\}$	$\{8,9,10\}$	3.64075
10	$\{1,2,3,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	3.49326	$\{1,2,3,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	4.06242

Table 16: Performance evaluation of Algorithm 3. "Value" refers to $D(\otimes_{i=1}^{3} P^{(S_{m,i})} \| \otimes_{i=1}^{3} \Pi^{(S_{m,i})})$.



(a) Bernoulli-Laplace level model (b) Curie-Weiss model

Figure 8: Performance evaluation of the generalized distorted greedy algorithm.

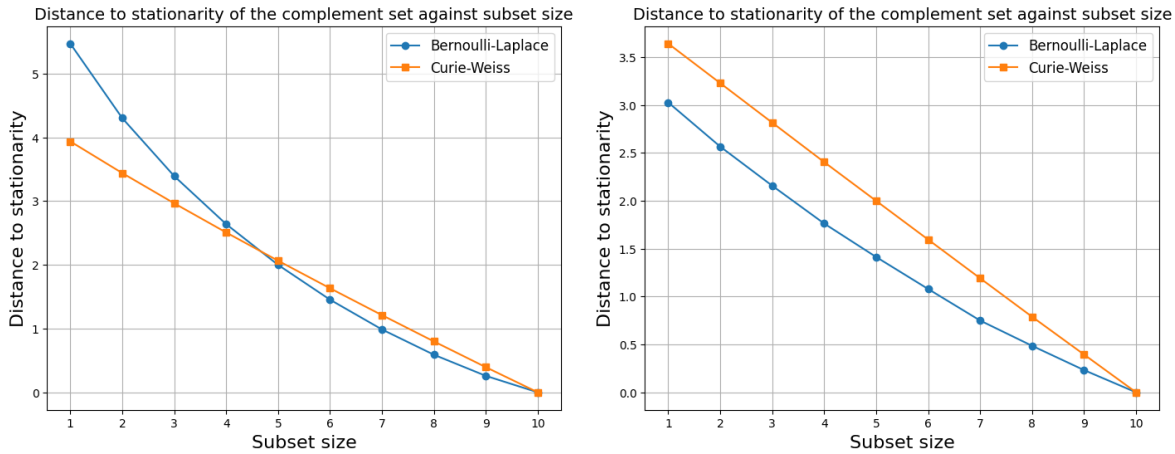
We proceed to present the numerical experiment results in Section 6.1 and Section 6.3 (see Table 17, Table 18, and Figure 9). Note that since the stationary distribution π of both models is not of product form, we do not have the $(1-e^{-1})$-approximation guarantee.

	Bernoulli-Laplace level model		**Curie-Weiss model**	
m	**Subset** S_m	$D(P^{(-S_m)}\|\Pi^{(-S_m)})$	**Subset** S_m	$D(P^{(-S_m)}\|\Pi^{(-S_m)})$
1	{9}	5.46950	{10}	3.93568
2	{9, 10}	4.30094	{9, 10}	3.43908
3	{8, 9, 10}	3.39168	{8, 9, 10}	2.96487
4	{7, 8, 9, 10}	2.63821	{7, 8, 9, 10}	2.507645
5	{6, 7, 8, 9, 10}	1.99871	{6, 7, 8, 9, 10}	2.06420
6	{4, 6, 7, 8, 9, 10}	1.45314	{5, 6, 7, 8, 9, 10}	1.63242
7	{3, 4, 6, 7, 8, 9, 10}	0.98630	{4, 5, 6, 7, 8, 9, 10}	1.21075
8	{1, 3, 4, 6, 7, 8, 9, 10}	0.58961	{3, 4, 5, 6, 7, 8, 9, 10}	0.79828
9	{1, 2, 3, 4, 6, 7, 8, 9, 10}	0.25830	{2, 3, 4, 5, 6, 7, 8, 9, 10}	0.39435
10	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	0.00000	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	0.00000

Table 17: Performance evaluation of the greedy algorithm.

	Bernoulli-Laplace level model				**Curie-Weiss model**			
m	$S_{m,1}$	$S_{m,2}$	$S_{m,3}$	Value	$S_{m,1}$	$S_{m,2}$	$S_{m,3}$	Value
1	$\{4\}$	$\emptyset$	$\emptyset$	3.02668	$\{4\}$	$\emptyset$	$\emptyset$	3.64075
2	$\{4\}$	$\emptyset$	$\{9\}$	2.56554	$\{4\}$	$\emptyset$	$\{8\}$	3.22828
3	$\{4\}$	$\{6\}$	$\{9\}$	2.15700	$\{3,4\}$	$\emptyset$	$\{8\}$	2.81582
4	$\{1,4\}$	$\{6\}$	$\{9\}$	1.76235	$\{3,4\}$	$\{5\}$	$\{8\}$	2.40381
5	$\{1,4\}$	$\{6\}$	$\{8,9\}$	1.41297	$\{3,4\}$	$\{5,6\}$	$\{8\}$	1.99886
6	$\{1,4\}$	$\{5,6\}$	$\{8,9\}$	1.07899	$\{2,3,4\}$	$\{5,6\}$	$\{8\}$	1.59492
7	$\{1,2,4\}$	$\{5,6\}$	$\{8,9\}$	0.74955	$\{2,3,4\}$	$\{5,6\}$	$\{8,9\}$	1.19099
8	$\{1,2,3,4\}$	$\{5,6\}$	$\{8,9\}$	0.48566	$\{2,3,4\}$	$\{5,6,7\}$	$\{8,9\}$	0.78871
9	$\{1,2,3,4\}$	$\{5,6,7\}$	$\{8,9\}$	0.23191	$\{2,3,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	0.39436
10	$\{1,2,3,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	0.00000	$\{1,2,3,4\}$	$\{5,6,7\}$	$\{8,9,10\}$	0.00000

Table 18: Performance evaluation of Algorithm 3. "Value" refers to $D(\otimes_{i=1}^{3} P^{(V_i \setminus S_{m,i})} \| \otimes_{i=1}^{3} \Pi^{(V_i \setminus S_{m,i})})$.



(a) Greedy

(b) Algorithm 3

Figure 9: Distance to stationarity of the complement set against subset size.

8.5 Experiment results of Section 7

We perform Algorithm 4 with the following configuration: $l = \lceil \frac{m}{2} \rceil$, $q_i = 2$ for $i \in [\![l-1]\!]$; $q_l = 2$ if m is even, $q_l = 1$ if m is odd. We choose the fixed subset as $W = \{1, 2, 3\}$. The performance of the batch

greedy algorithm on the two models is shown in Table 19 and Figure 10.

	Bernoulli-Laplace level model		**Curie-Weiss model**	
m	Subset S_l	$D(P^{(W\cup S_l)}\|P^{(W)}\otimes P^{(S_l)})$	Subset S_l	$D(P^{(W\cup S_l)}\|P^{(W)}\otimes P^{(S_l)})$
1	$\{10\}$	0.14671	$\{4\}$	0.02751
2	$\{9,10\}$	0.26354	$\{4,10\}$	0.05651
3	$\{8,9,10\}$	0.37787	$\{4,5,10\}$	0.08919
4	$\{7,8,9,10\}$	0.49198	$\{4,5,9,10\}$	0.12616
5	$\{6,7,8,9,10\}$	0.61908	$\{4,5,6,9,10\}$	0.17028
6	$\{5,6,7,8,9,10\}$	0.79889	$\{4,5,6,8,9,10\}$	0.22527
7	$\{4,5,6,7,8,9,10\}$	1.06993	$\{4,5,6,7,8,9,10\}$	0.30491

Table 19: Performance evaluation of the batch greedy algorithm.

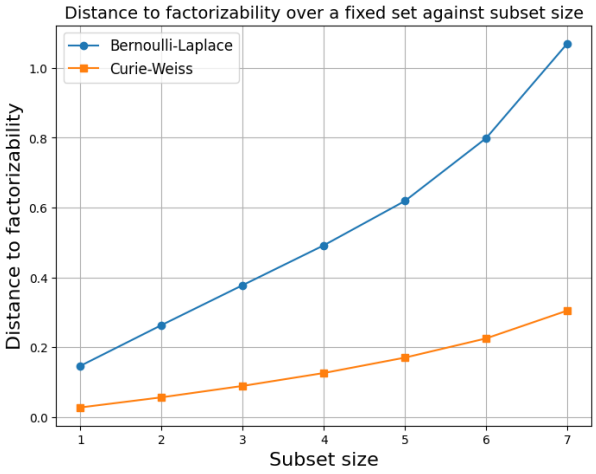


Figure 10: Performance evaluation of the batch greedy algorithm.

Part II
Minimax factorization for a family of multivariate Markov chains

9 The minimax optimization problem

We denote a feasible set $\mathcal{F}$ for the choice of factorizable transition matrix Q:

$$\mathcal{F} = \mathcal{F}(\mathbf{S}) := \{Q \in \mathcal{L}(\mathcal{X});\ \mathbf{S} = (S_1, \ldots, S_m) \in (m+1)^{[\![d]\!]},\ Q = Q^{(S_1)} \otimes \ldots \otimes Q^{(S_m)}\}.$$

We are interested in the following minimax optimization problem

$$\min_{Q \in \mathcal{F}} \max_{P \in \mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q), \tag{45}$$

in words, we seek to find an optimal factorizable $Q \in \mathcal{F}$ that minimize the worst-case information loss in approximating members of $\mathcal{B}$.

Since $\mathcal{F}$ is not a convex set, we denote

$$\mathcal{M} := \{M \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}\}$$

as the set of matrices on the state space $\mathcal{X}$ and study the weighted geometric mean and the following set:

$$\mathcal{A} := \left\{ A \in \mathcal{M};\ \exists l \in \mathbb{N}, \mathbf{c} \in \mathcal{S}_l \text{ s.t. } A(x,y) = \sum_{i=1}^{l} c_i \log Q_i(x,y),\ \forall x, y;\ Q_i \in \mathcal{F},\ \forall i \in [\![l]\!] \right\}.$$

Lemma 9.1. *The set $\mathcal{A}$ is convex.*

Proof. We choose $A, B \in \mathcal{A}$ such that there exists $\mathbf{c} \in \mathcal{S}_l$, $\mathbf{d} \in \mathcal{S}_k$, $Q_i, R_j \in \mathcal{F}$ for $i \in [\![l]\!], j \in [\![k]\!]$ and for all x, y we have

$$A(x,y) = \sum_{i=1}^{l} c_i \log Q_i(x,y),\ B(x,y) = \sum_{i=1}^{k} d_i \log R_i(x,y).$$

We choose $\alpha \in [0,1]$ and calculate that

$$\alpha A(x,y) + (1-\alpha)B(x,y) = \sum_{i=1}^{l} \alpha c_i \log Q_i(x,y) + \sum_{i=1}^{k} (1-\alpha) d_i \log R_i(x,y).$$

We thus conclude that $\alpha A + (1-\alpha)B \in \mathcal{A}$, and hence $\mathcal{A}$ is convex. □

We define the **elementwise exponential** of a matrix $M \in \mathcal{M}$ to be $\exp M$, that is, for all $x, y \in \mathcal{X}$,

$$\exp M(x,y) := e^{M(x,y)}.$$

We then define the **generalized KL divergence** from the non-negative and not necessarily stochastic matrix $\exp A$ to P to be

$$\begin{aligned} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A) &:= \sum_{x,y} \pi(x) P(x,y) \log \frac{P(x,y)}{\exp A(x,y)} \\ &= \sum_{x,y} \pi(x) P(x,y) \log P(x,y) - \sum_{x,y} \pi(x) P(x,y) A(x,y), \end{aligned}$$

which is linear in A, hence the map $\mathcal{A} \ni A \mapsto \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A)$ is convex.

We study the following minimax optimization problem

$$\min_{A \in \mathcal{A}} \max_{P \in \mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A), \tag{46}$$

and we can reformulate it as

$$\begin{aligned} &\min_{A\in\mathcal{A},\ r} \ r \\ &\text{s.t.} \quad \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) \le r,\ \forall i \in [\![n]\!], \end{aligned} \tag{47}$$

which is a constrained convex minimization problem.

Comparing problem (45) with problem (46), we note that for every $Q \in \mathcal{F}$, we can define an associated $A \in \mathcal{A}$ such that $A(x,y) = \log Q(x,y)$, and hence we have the following inequality:

$$\min_{Q\in\mathcal{F}} \max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q) \ge \min_{A\in\mathcal{A}} \max_{P\in\mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A). \tag{48}$$

Suppose $A \in \mathcal{A}$ such that $\exp A(x,y) = \prod_{i=1}^{l} Q_i(x,y)^{c_i}$ for any x, y, we then show a Pythagorean identity based on the proof of Theorem 2.22 of (Choi et al., 2024):

$$\begin{aligned} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A) &= \sum_{x,y} \pi(x)P(x,y)\log\frac{P(x,y)}{\prod_{i=1}^{l} Q_i(x,y)^{c_i}} \\ &= D^{\pi}_{\mathrm{KL}}(P\|\otimes_{i=1}^{m} P^{(S_i)}) + \sum_{x,y}\pi(x)P(x,y)\log\frac{\otimes_{i=1}^{m}P^{(S_i)}(x,y)}{\prod_{j=1}^{l}Q_j(x,y)^{c_j}} \\ &= D^{\pi}_{\mathrm{KL}}(P\|\otimes_{i=1}^{m} P^{(S_i)}) + \sum_{i=1}^{m}\sum_{j=1}^{l} c_j D^{\pi}_{\mathrm{KL}}(P^{(S_i)}\|Q_j^{(S_i)}) \ge \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A^*), \end{aligned} \tag{49}$$

where $A^* = A^*(S_1,\dots,S_m,P) \in \mathcal{A}$ is defined to be

$$A^*(x,y) := \log(\otimes_{i=1}^{m} P^{(S_i)}(x,y)).$$

Inspired by (49) and Lemma 2.2, for given $\mathbf{w} \in \mathcal{S}_n$, we show a weighted version of Pythagorean identity for generalized KL divergence:

$$\begin{aligned} \sum_{i=1}^{n} w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) &= \sum_{i=1}^{n} w_i \sum_{x,y}\pi(x)P_i(x,y)\log\frac{P_i(x,y)}{\prod_{k=1}^{l}Q_k(x,y)^{c_k}} \\ &= \sum_{i=1}^{n} w_i D^{\pi}_{\mathrm{KL}}(P_i\|\otimes_{j=1}^{m}\overline{P}^{(S_j)}) + \sum_{i=1}^{n} w_i \sum_{x,y}\pi(x)P_i(x,y)\log\frac{\otimes_{j=1}^{m}\overline{P}^{(S_j)}(x,y)}{\prod_{k=1}^{l}Q_k(x,y)^{c_k}} \\ &= \sum_{i=1}^{n} w_i D^{\pi}_{\mathrm{KL}}(P_i\|\otimes_{j=1}^{m}\overline{P}^{(S_j)}) + \sum_{j=1}^{m}\sum_{k=1}^{l} c_k D^{\pi^{(S_j)}}_{\mathrm{KL}}(\overline{P}^{(S_j)}\|Q_k^{(S_j)}) \\ &\ge \sum_{i=1}^{n} w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A_n^*(\mathbf{w})), \end{aligned} \tag{50}$$

where $A_n^*(\mathbf{w}) = A_n^*(\mathbf{w}, S_1,\dots,S_m,\mathcal{B}) \in \mathcal{A}$ is defined to be, for all $x, y \in \mathcal{X}$,

$$A_n^*(x,y) := \log(\otimes_{j=1}^{m}\overline{P}^{(S_j)})(x,y).$$

In the special case that $n = 1$, we recover that $A_1^* = A^*$.

For the problem (47), we denote the Lagrangian $L : \mathbb{R}_+ \times \mathcal{A} \times \mathbb{R}_+^n$ to be

$$L(r, A, \mathbf{w}) := r + \sum_{i=1}^{n} w_i(\widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) - r), \tag{51}$$

where $\mathbf{w}$ is the associated Lagrangian multiplier.

From the Pythagorean identity (50), the dual problem of (47) can be written as

$$\max_{\mathbf{w}\in\mathbb{R}_+^n} \min_{r\ge 0,\ A\in\mathcal{A}} L(r,A,\mathbf{w}) = \max_{\mathbf{w}\in\mathcal{S}_n} \min_{A\in\mathcal{A}} \sum_{i=1}^{n} w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) = \max_{\mathbf{w}\in\mathcal{S}_n} \sum_{i=1}^{n} w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A_n^*(\mathbf{w})). \tag{52}$$

The main results in this section are that strong duality holds for problem (47), and problem (45) and (46) are equivalent. We write the results in the following theorem.

Theorem 9.2. *1. The strong duality holds for problem* (47) *and there exists* $\mathbf{w}^* \in \mathcal{S}_n$ *such that*

$$\min_{A\in\mathcal{A}}\max_{P\in\mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A) = \max_{\mathbf{w}\in\mathcal{S}_n} \sum_{i=1}^{n} w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A^*_n(\mathbf{w})) = \sum_{i=1}^{n} w_i^* \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A^*_n(\mathbf{w}^*)).$$

2. Suppose the pair $(A, r) \in \mathcal{A}\times\mathbb{R}_+$ *minimizes the primal problem* (47) *and* $\mathbf{w}^* \in \mathcal{S}_n$ *maximizes the dual problem* (52)*, then the following complementary slackness results hold: for* $i \in [\![n]\!]$*, we have*

$$\widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) \begin{cases} = r, & \text{if } w_i^* > 0; \\ \le r, & \text{if } w_i^* = 0. \end{cases}$$

3. Problems (45) *and* (46) *are equivalent, i.e.*

$$\min_{Q\in\mathcal{F}}\max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q) = \min_{A\in\mathcal{A}}\max_{P\in\mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A).$$

4. The same $\mathbf{w}^* \in \mathcal{S}_n$ *from item* (1) *satisfies*

$$\min_{Q\in\mathcal{F}}\max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q) = \max_{\mathbf{w}\in\mathcal{S}_n} \sum_{i=1}^{n} w_i D^{\pi}_{\mathrm{KL}}(P_i\| \otimes_{k=1}^{m} \overline{P}(\mathbf{w})^{(S_k)}) = \sum_{i=1}^{n} w_i^* D^{\pi}_{\mathrm{KL}}(P_i\| \otimes_{k=1}^{m} \overline{P}(\mathbf{w}^*)^{(S_k)}).$$

5. The map

$$\mathcal{S}_n \ni \mathbf{w} \mapsto \sum_{i=1}^{n} w_i D^{\pi}_{\mathrm{KL}}(P_i\| \otimes_{k=1}^{m} \overline{P}(\mathbf{w})^{(S_k)})$$

is concave in $\mathbf{w}$*.*

Proof. We first show item (1), i.e., strong duality holds for problem (47). We shall show that the Slater's qualification is verified (see Section 5.2.3 of (Boyd and Vandenberghe, 2004) and Appendix A of (Beck, 2017)), which requires that the constraints in (47) are strictly feasible. We take any A and

$$r = \max_{i\in[\![n]\!]} \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) + 1 > \widetilde{D}^{\pi}_{\mathrm{KL}}(P_l\|A),\ \forall l \in [\![n]\!],$$

hence the strong duality holds. Therefore we have

$$\min_{A\in\mathcal{A}}\max_{P\in\mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A) = \max_{\mathbf{w}\in\mathcal{S}_n} \sum_{i=1}^{n} w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A^*_n(\mathbf{w})) = \sum_{i=1}^{n} w_i^* \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A^*_n(\mathbf{w}^*)).$$

As the strong duality in item (1) holds, by Section 5.5.2 of (Boyd and Vandenberghe, 2004), the *complementary slackness* condition holds, i.e.

$$w_i^*(\widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) - r) = 0,$$

which is equivalent to

$$\widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A) \begin{cases} = r, & \text{if } w_i^* > 0; \\ \le r, & \text{if } w_i^* = 0, \end{cases}$$

for all $i \in [\![n]\!]$, hence it proves item (2).

We proceed to prove item (3). Let $j \in [\![n]\!]$ be an index where $w_j^* > 0$, we want to show

$$\widetilde{D}^{\pi}_{\mathrm{KL}}(P_j\|A^*_n(\mathbf{w}^*)) = \max_{l\in[\![n]\!]} \widetilde{D}^{\pi}_{\mathrm{KL}}(P_l\|A^*_n(\mathbf{w}^*)).$$

As it is clear to see that $\widetilde{D}^{\pi}_{\mathrm{KL}}(P_j\|A^*_n(\mathbf{w}^*)) \le \max_{l\in[\![n]\!]} \widetilde{D}^{\pi}_{\mathrm{KL}}(P_l\|A^*_n(\mathbf{w}^*))$, we then assume the contrary that

$$\widetilde{D}^{\pi}_{\mathrm{KL}}(P_j\|A^*_n(\mathbf{w}^*)) < \max_{l\in[\![n]\!]} \widetilde{D}^{\pi}_{\mathrm{KL}}(P_l\|A^*_n(\mathbf{w}^*)).$$

That is, there exists an index l^* such that

$$\widetilde{D}^{\pi}_{\mathrm{KL}}(P_j\|A^*_n(\mathbf{w}^*)) < \widetilde{D}^{\pi}_{\mathrm{KL}}(P_{l^*}\|A^*_n(\mathbf{w}^*)).$$

By strong duality, we have $w^*_{l^*} = 0$, then by complementary slackness in item (2), we have

$$\widetilde{D}^{\pi}_{\mathrm{KL}}(P_{l^*}\|A^*_n(\mathbf{w}^*)) \le r = \widetilde{D}^{\pi}_{\mathrm{KL}}(P_j\|A^*_n(\mathbf{w}^*)) < \widetilde{D}^{\pi}_{\mathrm{KL}}(P_{l^*}\|A^*_n(\mathbf{w}^*)),$$

which leads to a contradiction. It therefore yields

$$\widetilde{D}^{\pi}_{\mathrm{KL}}(P_j\|A^*_n(\mathbf{w}^*)) = \max_{l\in[\![n]\!]} \widetilde{D}^{\pi}_{\mathrm{KL}}(P_l\|A^*_n(\mathbf{w}^*)).$$

By recalling the definition of generalized KL divergence and (48), we have

$$\begin{aligned}
\min_{Q\in\mathcal{F}}\max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q) &\ge \min_{A\in\mathcal{A}}\max_{P\in\mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A) = \max_{\mathbf{w}\in\mathcal{S}_n}\sum_{i=1}^n w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A^*_n(\mathbf{w})) \\
&= \max_{l\in[\![n]\!]} \widetilde{D}^{\pi}_{\mathrm{KL}}(P_l\|A^*_n(\mathbf{w}^*)) = \widetilde{D}^{\pi}_{\mathrm{KL}}(P_j\|A^*_n(\mathbf{w}^*)) \\
&= \max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|\otimes_{k=1}^m \overline{P}(\mathbf{w}^*)^{(S_k)}) \ge \min_{Q\in\mathcal{F}}\max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q),
\end{aligned}$$

therefore we obtain

$$\min_{Q\in\mathcal{F}}\max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q) = \min_{A\in\mathcal{A}}\max_{P\in\mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A),$$

hence problem (45) and problem (46) are equivalent. Therefore, for the $\mathbf{w}^* \in \mathcal{S}_n$ in item (1), we have

$$\begin{aligned}
\min_{Q\in\mathcal{F}}\max_{P\in\mathcal{B}} D^{\pi}_{\mathrm{KL}}(P\|Q) = \min_{A\in\mathcal{A}}\max_{P\in\mathcal{B}} \widetilde{D}^{\pi}_{\mathrm{KL}}(P\|A) &= \sum_{i=1}^n w^*_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A^*_n(\mathbf{w}^*)) \\
&= \sum_{i=1}^n w^*_i D^{\pi}_{\mathrm{KL}}(P_i\|\otimes_{k=1}^m \overline{P}(\mathbf{w}^*)^{(S_k)}),
\end{aligned}$$

which proves item (4).

We then show item (5). From (52), we have

$$\sum_{i=1}^n w_i D^{\pi}_{\mathrm{KL}}(P_i\|\otimes_{k=1}^m \overline{P}(\mathbf{w})^{(S_k)}) = \sum_{i=1}^n w_i \widetilde{D}^{\pi}_{\mathrm{KL}}(P_i\|A^*_n) = \min_{r\ge 0,\ A\in\mathcal{A}} L(r, A, \mathbf{w}),$$

hence the map

$$\mathcal{S}_n \ni \mathbf{w} \mapsto \sum_{i=1}^n w_i D^{\pi}_{\mathrm{KL}}(P_i\|\otimes_{k=1}^m \overline{P}(\mathbf{w})^{(S_k)})$$

is concave since it is the Lagrangian dual function of problem (47) (see Section 5.1.2 of (Boyd and Vandenberghe, 2004)). □

10 An information-theoretic game

Inspired by the reversiblization entropy games in (Choi and Wolfer, 2025), we cast the minimax problem as a two-player zero-sum game between Nature and a probabilist. Nature chooses a transition probability matrix $P \in \mathcal{B}$, while the probabilist chooses an approximating factorizable transition matrix $Q \in \mathcal{F} = \mathcal{F}(\mathbf{S})$. The payoff is the KL divergence $D^{\pi}_{\mathrm{KL}}(P\|Q)$, which Nature aims to *maximize* while the probabilist aims to *minimize*.

In the *pure strategy game*, Nature selects a single $P \in \mathcal{B}$ and the probabilist selects a single $Q \in \mathcal{F}$. In the *mixed strategy game*, Nature is permitted to randomize over $\mathcal{B}$ according to a probability distribution $\mu \in \mathcal{P}(\mathcal{B})$ (which corresponds to a weight vector $\mathbf{w} \in \mathcal{S}_n$), while the probabilist still chooses a single $Q \in \mathcal{F}$.

We adapt the following notation for some related minimax and maximin values:

$$\begin{aligned}
\overline{V} = \overline{V}(\mathbf{S}, \mathcal{B}) &= \min_{Q \in \mathcal{F}} \max_{\mu \in \mathcal{P}(\mathcal{B})} \int_{\mathcal{B}} D_{\mathrm{KL}}^{\pi}(P \| Q) \mu(\mathrm{d}P), \\
\underline{V} = \underline{V}(\mathbf{S}, \mathcal{B}) &= \max_{\mu \in \mathcal{P}(\mathcal{B})} \min_{Q \in \mathcal{F}} \int_{\mathcal{B}} D_{\mathrm{KL}}^{\pi}(P \| Q) \mu(\mathrm{d}P), \\
\overline{v} = \overline{v}(\mathbf{S}, \mathcal{B}) &= \min_{Q \in \mathcal{F}} \max_{P \in \mathcal{B}} D_{\mathrm{KL}}^{\pi}(P \| Q), \\
\underline{v} = \underline{v}(\mathbf{S}, \mathcal{B}) &= \max_{P \in \mathcal{B}} \min_{Q \in \mathcal{F}} D_{\mathrm{KL}}^{\pi}(P \| Q).
\end{aligned}$$

From item (4) of Theorem 9.2, the pure-strategy minimax value $\overline{v}$ is equivalent to the dual problem:

$$\overline{v} = \min_{Q \in \mathcal{F}} \max_{P \in \mathcal{B}} D_{\mathrm{KL}}^{\pi}(P \| Q) = \max_{\mathbf{w} \in \mathcal{S}_n} \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^{m} \overline{P}^{(S_j)}). \tag{53}$$

The following theorem establishes the existence of a mixed-strategy Nash equilibrium (see Section 3 of (Osborne and Rubinstein, 1994)), which is a foundational result in game theory.

Theorem 10.1 (Existence of mixed strategy Nash equilibrium)**.** *Consider the two-person mixed strategy game with respect to parameters* $(\mathbf{S}, \mathcal{B})$*,*

1. *The mixed strategy Nash equilibrium always exists. That is, the value of the game is well-defined and given by*

$$\overline{V}(\mathbf{S}, \mathcal{B}) = \underline{V}(\mathbf{S}, \mathcal{B}) = \max_{\mathbf{w} \in \mathcal{S}_n} \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^{m} \overline{P}^{(S_j)}).$$

2. *The mixed strategy Nash equilibrium is attained at* (Q^*, μ^*)*, where* μ^* *is represented by the optimal weight vector* $\mathbf{w}^* \in \mathcal{S}_n$ *and* Q^* *is the information projection of the corresponding weighted average* $\overline{P}(\mathbf{w}^*)$ *onto* $\mathcal{F}$*, i.e.*

$$Q^* = \otimes_{j=1}^{m} \overline{P}(\mathbf{w}^*)^{(S_j)}.$$

Proof. We first show existence in item (1). By Proposition 3.10 of (Choi and Wolfer, 2025), we have the standard minimax inequalities $\overline{v}(\mathbf{S}, \mathcal{B}) \geq \overline{V}(\mathbf{S}, \mathcal{B}) \geq \underline{V}(\mathbf{S}, \mathcal{B})$. We can also establish a lower bound for $\underline{V}$ by restricting Nature's strategy space from all probability measures $\mathcal{P}(\mathcal{B})$ to the simplex of finite measures $\mathcal{S}_n$:

$$\begin{aligned}
\underline{V} = \underline{V}(\mathbf{S}, \mathcal{B}) &= \max_{\mu \in \mathcal{P}(\mathcal{B})} \min_{Q \in \mathcal{F}} \int_{\mathcal{B}} D_{\mathrm{KL}}^{\pi}(P \| Q) \mu(\mathrm{d}P) \\
&\geq \max_{\mathbf{w} \in \mathcal{S}_n} \min_{Q \in \mathcal{F}} \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| Q) \\
&= \max_{\mathbf{w} \in \mathcal{S}_n} \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^{m} \overline{P}^{(S_j)}) = \overline{v},
\end{aligned}$$

where the second last equality comes from Lemma 2.2 and the final equality comes from (53). We have thus shown the chain of inequalities $\overline{v} \geq \overline{V} \geq \underline{V} \geq \overline{v}$, which enforces equality throughout. This implies $\overline{V} = \underline{V}$, confirming that the mixed-strategy Nash equilibrium exists.

Item (2) follows from item (1). At the mixed-strategy Nash equilibrium, the pair of optimal strategies (Q^*, μ^*) is composed of Nature's optimal strategy μ^*, which is represented by the optimal weight vector $\mathbf{w}^* \in \mathcal{S}_n$, and the probabilist's optimal pure strategy $Q^* \in \mathcal{F}$. Nature's strategy $\mathbf{w}^*$ is the solution to the dual maximization problem as in item (4) of Theorem 9.2, identifying the "worst-case" mixture in $\mathcal{B}$. In response to this specific mixture, the probabilist's unique best response Q^* is the information projection of the corresponding weighted average model $\overline{P}(\mathbf{w}^*)$ onto the set of factorizable $\mathcal{F}$, which is explicitly given by $Q^* = \otimes_{j=1}^{m} \overline{P}(\mathbf{w}^*)^{(S_j)}$. □

11 A projected subgradient algorithm

From Theorem 9.2, since problems (45) and (46) are equivalent (item (3)), hence by item (4), it suffices to solving the following convex minimization problem:

$$\min_{\mathbf{w}\in\mathcal{S}_n} \quad h(\mathbf{w}), \tag{54}$$

where $h(\mathbf{w}) = -\sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{w})^{(S_k)})$ is convex from item (5). We now compute a subgradient of h, through which we aim to propose a projected subgradient algorithm with theoretical guarantee.

Theorem 11.1 (Subgradient of h and an upper bound of its l^2-norm)**.** *A subgradient of h at $\mathbf{v}\in\mathcal{S}_n$ is given by $\mathbf{g}=\mathbf{g}(\mathbf{v})=(g_1,\ldots,g_n)\in\mathbb{R}^n$, where for all $i\in[\![n]\!]$, we have*

$$g_i = g_i(\mathbf{v}) = D_{\mathrm{KL}}^{\pi}(P_n \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}) - D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}).$$

The subgradient $\mathbf{g}$ satisfies that, for all $\mathbf{w},\mathbf{v}\in\mathcal{S}_n$,

$$h(\mathbf{w}) \geq h(\mathbf{v}) + \sum_{i=1}^n g_i \cdot (w_i - v_i).$$

Moreover, the l^2-norm of $\mathbf{g}(\mathbf{v})$ is bounded above by

$$\|\mathbf{g}\|_2^2 = \sum_{i=1}^n g_i^2 \leq n\left(|\mathcal{X}| \sup_{\mathbf{v}\in\mathcal{S}_n;\ i\in[\![n]\!];\ P_i(x,y)>0} P_i(x,y)\ln\frac{P_i(x,y)}{\otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}(x,y)}\right)^2 := B.$$

Proof. By the Pythagorean identity (Lemma 2.2), we have

$$\sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{w})^{(S_k)}) \leq \sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)})$$

for any $\mathbf{w},\mathbf{v}\in\mathcal{S}_n$. Hence,

$$\begin{aligned}
h(\mathbf{w}) - h(\mathbf{v}) &= -\sum_{i=1}^n w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{w})^{(S_k)}) + \sum_{i=1}^n v_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}) \\
&\geq -\sum_{i=1}^n (w_i - v_i) D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}) \\
&= -\sum_{i=1}^n (w_i - v_i) D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}) + \sum_{i=1}^n (w_i - v_i) D_{\mathrm{KL}}^{\pi}(P_n \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}) \\
&= \sum_{i=1}^n (w_i - v_i) g_i,
\end{aligned}$$

where the second last equation holds because $\mathbf{w},\mathbf{v}\in\mathcal{S}_n$, and hence $\sum_{i=1}^n (w_i - v_i) = 0$.

We proceed to prove the upper bound on the l^2-norm. We first show the upper bound of the KL divergence term:

$$\begin{aligned}
D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}) &= \sum_{x\in\mathcal{X}} \pi(x) \sum_{y\in\mathcal{X}} P_i(x,y)\ln\frac{P_i(x,y)}{\otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}(x,y)} \\
&\leq |\mathcal{X}| \sup_{\mathbf{v}\in\mathcal{S}_n;\ i\in[\![n]\!];\ P_i(x,y)>0} P_i(x,y)\ln\frac{P_i(x,y)}{\otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)}(x,y)} = \sqrt{\frac{B}{n}},
\end{aligned}$$

and then we have

$$\begin{aligned}
\|\mathbf{g}\|_2^2 = \sum_{i=1}^n g_i^2 &\leq \sum_{i=1}^n \max\left\{D_{\mathrm{KL}}^{\pi}(P_n \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)})^2, D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)})^2\right\} \\
&\leq n \max_{l\in[\![n]\!]} D_{\mathrm{KL}}^{\pi}(P_l \| \otimes_{k=1}^m \overline{P}(\mathbf{v})^{(S_k)})^2 \leq n\cdot\sqrt{\frac{B}{n}}^2 = B.
\end{aligned}$$

□

Inspired by Algorithm 1 of (Choi and Wolfer, 2025), we propose a projected subgradient algorithm to solve problem (54). In Algorithm 5, we conduct the projected subgradient algorithm for t iterations. At each iteration, we first update the weight parameters via subgradient,

$$\mathbf{v}^{(i)} = \mathbf{w}^{(i-1)} - \eta \cdot \mathbf{g}(\mathbf{w}^{(i-1)}),$$

where $\eta > 0$ is the stepsize of the algorithm while we take $\mathbf{g}$ as in Theorem 11.1, the subgradient of h. In the second step, the updated weight $\mathbf{v}^{(i)}$ is to be projected onto the n-probability-simplex $\mathcal{S}_n$, i.e.

$$\mathbf{w}^{(i)} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathcal{S}_n} \|\mathbf{w} - \mathbf{v}^{(i)}\|_2^2,$$

which can be accomplished by existing projection algorithms onto a simplex (see e.g. (Condat, 2016)). Note that the subgradient algorithm is not a descent algorithm, hence the monotonicity of $h(\mathbf{w})$ among different iterations is not guaranteed, see Section 13.1 for examples.

Algorithm 5 A projected subgradient algorithm to solve problem (54)

Require: Initial weight value $\mathbf{w}^{(0)} \in \mathcal{S}_n$, set $\{P_i\}_{i=1}^n$, target distribution π, stepsize $\eta > 0$, and number of iterations t

1: **for** $i = 1$ to t **do**
2: $\quad \mathbf{v}^{(i)} \leftarrow \mathbf{w}^{(i-1)} - \eta \cdot \mathbf{g}(\mathbf{w}^{(i-1)})$ ▷ Update via subgradient descent
3: $\quad \mathbf{w}^{(i)} \leftarrow \operatorname*{arg\,min}_{\mathbf{w} \in \mathcal{S}_n} \|\mathbf{w} - \mathbf{v}^{(i)}\|_2^2$ ▷ Project onto $\mathcal{S}_n$
4: **end for**
5: **Output:** The sequence $\left(\mathbf{w}^{(i)}\right)_{i=1}^t$

The rest of the section is devoted to providing a theoretical guarantee for Algorithm 5. We first prove an upper bound of Algorithm 5.

Theorem 11.2 (Upper bound of Algorithm 5)**.** *Consider Algorithm 5 with its outputs* $(\mathbf{w}^{(i)})_{i=1}^t$*, we have*

$$h(\overline{\mathbf{w}}^t) - h(\mathbf{w}^*) \leq \frac{n}{2\eta t} + \frac{\eta B}{2},$$

where $\overline{\mathbf{w}}^t = \frac{1}{t}\sum_{i=1}^t \mathbf{w}^{(i)}$ *and* $\mathbf{w}^*$ *is the optimal solution to problem* (54)*. Furthermore, if we choose constant stepsize* $\eta = \sqrt{\frac{n}{Bt}}$*, we have*

$$h(\overline{\mathbf{w}}^t) - h(\mathbf{w}^*) \leq \sqrt{\frac{nB}{t}}.$$

In addition, given any $\epsilon > 0$*, if we further choose*

$$t = \left\lceil \frac{nB}{\epsilon^2} \right\rceil,$$

then we can reach an ϵ*-close value to* $h(\mathbf{w}^*)$ *such that*

$$h(\overline{\mathbf{w}}^t) - h(\mathbf{w}^*) \leq \epsilon.$$

Proof. For all $i \in [\![t]\!]$, due to projection, we have

$$\begin{aligned}
\|\mathbf{w}^{(i+1)} - \mathbf{w}^*\|_2^2 &\leq \|\mathbf{v}^{(i+1)} - \mathbf{w}^*\|_2^2 = \|\mathbf{w}^{(i)} - \eta \cdot \mathbf{g}(\mathbf{w}^{(i)}) - \mathbf{w}^*\|_2^2 \\
&= \|\mathbf{w}^{(i)} - \mathbf{w}^*\|_2^2 + \eta^2 \|\mathbf{g}(\mathbf{w}^{(i)})\|^2 - 2\eta \mathbf{g}(\mathbf{w}^{(i)})(\mathbf{w}^{(i)} - \mathbf{w}^*) \\
&\leq \|\mathbf{w}^{(i)} - \mathbf{w}^*\|_2^2 + \eta^2 B - 2\eta \mathbf{g}(\mathbf{w}^{(i)})(\mathbf{w}^{(i)} - \mathbf{w}^*),
\end{aligned}$$

where the last inequality come from the upper bound in Theorem 11.1. We then apply the definition of subgradient $\mathbf{g}$ in Theorem 11.1, and it leads to

$$\begin{aligned}
h(\mathbf{w}^{(i)}) - h(\mathbf{w}^*) &\leq \mathbf{g}(\mathbf{w}^{(i)}) \cdot (\mathbf{w}^{(i)} - \mathbf{w}^*) \\
&\leq \frac{1}{2\eta}\left(\|\mathbf{w}^{(i)} - \mathbf{w}^*\|_2^2 - \|\mathbf{w}^{(i+1)} - \mathbf{w}^*\|_2^2\right) + \frac{\eta B}{2}.
\end{aligned}$$

We then take summation on i from 1 to t and obtain

$$\sum_{i=1}^{t}(h(\mathbf{w}^{(i)})-h(\mathbf{w}^*)) \le \frac{1}{2\eta}\left(\|\mathbf{w}^{(1)}-\mathbf{w}^*\|_2^2-\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2^2\right)+\frac{\eta Bt}{2}$$
$$\le \frac{1}{2\eta}\|\mathbf{w}^{(i)}-\mathbf{w}^*\|_2^2+\frac{\eta Bt}{2} \le \frac{n}{2\eta}+\frac{\eta Bt}{2},$$

where the last inequality holds because $\mathbf{w}^{(i)}, \mathbf{w}^* \in \mathcal{S}_n$. From the convexity of h, we have

$$h(\overline{\mathbf{w}}^t)-h(\mathbf{w}^*) \le \frac{1}{t}\left(\sum_{i=1}^{t}(h(\mathbf{w}^{(i)})-h(\mathbf{w}^*))\right) \le \frac{n}{2\eta t}+\frac{\eta B}{2}.$$

By AM-GM inequality, the right hand side is minimized when we choose stepsize $\eta=\sqrt{\frac{n}{Bt}}$, we then obtain

$$h(\overline{\mathbf{w}}^t)-h(\mathbf{w}^*) \le \sqrt{\frac{nB}{t}}.$$

□

We proceed to discuss the convergence rate of Algorithm 5. We define the π-weighted **total variation distance** between Q and P as

$$D_{\mathrm{TV}}^{\pi}(P\|Q) := \frac{1}{2}\sum_{x,y\in\mathcal{X}}\pi(x)|P(x,y)-Q(x,y)|,$$

and show the convergence rate of Algorithm 5.

Theorem 11.3 (Convergence rate of Algorithm 5)**.** *Consider Algorithm 5 and its outputs $(\mathbf{w}^{(i)})_{i=1}^t$, and the stepsize is chosen to be $\eta=\sqrt{\frac{n}{Bt}}$, we have*

$$D_{\mathrm{TV}}^{\pi}(\otimes_{k=1}^m \overline{P}(\overline{\mathbf{w}})^{(S_k)}\| \otimes_{k=1}^m \overline{P}(\mathbf{w}^*)^{(S_k)}) = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right).$$

Proof. From the convexity of KL divergence $D_{\mathrm{KL}}^{\pi}(\cdot\|\cdot)$ and Equation 3.25 of (Csiszár, 1972), we have a constant C such that

$$\begin{aligned}
&D_{\mathrm{TV}}^{\pi}(\otimes_{k=1}^m \overline{P}(\overline{\mathbf{w}})^{(S_k)}\| \otimes_{k=1}^m \overline{P}(\mathbf{w}^*)^{(S_k)})\\
&\le C\left(\sum_{i=1}^{n}\overline{w}_i^t D_{\mathrm{KL}}^{\pi}(P_i\| \otimes_{k=1}^m \overline{P}(\mathbf{w}^*)^{(S_k)}) - \sum_{i=1}^{n}\overline{w}_i^t D_{\mathrm{KL}}^{\pi}(P_i\| \otimes_{k=1}^m \overline{P}(\overline{\mathbf{w}}^{(i)})^{(S_k)})\right)\\
&\le C\left(\max_{i\in[\![n]\!]} D_{\mathrm{KL}}^{\pi}(P_i\| \otimes_{k=1}^m \overline{P}(\mathbf{w}^*)^{(S_k)}) + h(\overline{\mathbf{w}}^t)\right)\\
&= C(h(\overline{\mathbf{w}}^t)-h(\mathbf{w}^*)) = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right),
\end{aligned}$$

where the second last equality comes from the complementary slackness introduced in item (2) of Theorem 9.2, and the last equality comes from Theorem 11.2 as we choose the stepsize $\eta=\sqrt{\frac{n}{Bt}}$. □

Remark 11.4. *Theorem 11.2 and Theorem 11.3 establish the theoretical guarantee of Algorithm 5 through the averaged output $\overline{\mathbf{w}}^t$. However, in numerical experiments, we choose* $\arg\min_{i\in[\![t]\!]} h(\mathbf{w}^{(i)})$ *as the result for practical purpose, see Section 13.1.*

12 A max-min-max submodular optimization problem and a two-layer subgradient-greedy algorithm

Recall that in earlier sections we consider the minimax problem (45) and investigate its implications in the two-person game between Nature and probabilist. As the set $\mathcal{F}(\mathbf{S})$ depends on the choice of the partition $\mathbf{S}$, in this section we consider a max-min-max optimization problem of the form

$$\max_{\mathbf{S}\in(m+1)^{[\![d]\!]}} \min_{Q\in\mathcal{F}} \max_{\mu\in\mathcal{P}(\mathcal{B})} \int_{\mathcal{B}} D_{\mathrm{KL}}^{\pi}(P\|Q)\mu(\mathrm{d}P).$$

In words, we seek to find an optimal partition the maximizes the minimal worst-case information loss. We write

$$f(\mathbf{S},\mathbf{w}) := \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^{m} \overline{P}(\mathbf{w})^{(S_j)}), \tag{55}$$

and from the mixed-strategy Nash equilibrium (item (1) of Theorem 10.1), we can denote the inner part as

$$\begin{aligned}
f(\mathbf{S},\mathbf{w}^*(\mathbf{S})) &= \min_{Q\in\mathcal{F}} \max_{\mu\in\mathcal{P}(\mathcal{B})} \int_{\mathcal{B}} D_{\mathrm{KL}}^{\pi}(P\|Q)\mu(\mathrm{d}P) \\
&= \max_{\mathbf{w}\in\mathcal{S}_n} \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^{m} \overline{P}(\mathbf{w})^{(S_j)}), \quad \mathbf{S}\in(m+1)^{[\![d]\!]} \\
&= \sum_{i=1}^{n} w_i^* D_{\mathrm{KL}}^{\pi}(P_i \| \otimes_{j=1}^{m} \overline{P}(\mathbf{w}^*)^{(S_j)}), \quad \mathbf{S}\in(m+1)^{[\![d]\!]} \\
&= \sum_{i=1}^{n} w_i^* D_{\mathrm{KL}}^{\pi}(P_i \| (\otimes_{j=1}^{m-1} \overline{P}(\mathbf{w}^*)^{(S_j)}) \otimes \overline{P}(\mathbf{w}^*)^{(-\mathrm{supp}(\mathbf{S}))}), \quad \mathbf{S}\in m^{[\![d]\!]},
\end{aligned}$$

where we write

$$\mathbf{w}^* = \mathbf{w}^*(\mathbf{S}) = \arg\max_{\mathbf{w}\in\mathcal{S}_n} f(\mathbf{S},\mathbf{w}).$$

We furthermore choose the ground set $\mathbf{V}\in m^{[\![d]\!]}$ and cardinality constraint l, and instead consider the max-min-max optimization problem

$$\max_{\mathbf{S}\preceq\mathbf{V};\ |\mathrm{supp}(\mathbf{S})|\le l} f(\mathbf{S},\mathbf{w}^*(\mathbf{S})). \tag{56}$$

We then investigate the following map for fixed $\mathbf{w}\in\mathcal{S}_n$ through the lens of submodularity:

$$m^{[\![d]\!]} \ni \mathbf{S} \mapsto f(\mathbf{S}) = f(\mathbf{S},\mathbf{w}) := \sum_{i=1}^{n} w_i D_{\mathrm{KL}}^{\pi}(P_i \| (\otimes_{j=1}^{m-1} \overline{P}(\mathbf{w})^{(S_j)}) \otimes \overline{P}(\mathbf{w})^{(-\mathrm{supp}(\mathbf{S}))}). \tag{57}$$

Lemma 12.1. *The map* (57) *is orthant submodular.*

Proof. We shall prove that $\Delta_{e,j} f(\mathbf{S}) \ge \Delta_{e,j} f(\mathbf{T})$ from the definition of orthant submodularity, where we choose $\mathbf{S}\preceq\mathbf{T}$ and $e\notin\mathrm{supp}(\mathbf{T})$.

$$\begin{aligned}
\Delta_{e,j} f(\mathbf{S}) - \Delta_{e,j} f(\mathbf{T}) &= \sum_{i=1}^{n} w_i \left(H(\overline{P}^{(S_j\cup\{e\})}) - H(\overline{P}^{(S_j)}) + H(\overline{P}^{(-\mathrm{supp}(\mathbf{S})\cup\{e\})}) - H(\overline{P}^{(-\mathrm{supp}(\mathbf{S}))}) \right) \\
&\quad - \sum_{i=1}^{n} w_i \left(H(\overline{P}^{(T_j\cup\{e\})}) - H(\overline{P}^{(T_j)}) + H(\overline{P}^{(-\mathrm{supp}(\mathbf{T})\cup\{e\})}) - H(\overline{P}^{(-\mathrm{supp}(\mathbf{T}))}) \right) \\
&= \left[\left(H(\overline{P}^{(S_j\cup\{e\})}) - H(\overline{P}^{(S_j)}) \right) - \left(H(\overline{P}^{(T_j\cup\{e\})}) - H(\overline{P}^{(T_j)}) \right) \right] \\
&\quad + \left[\left(H(\overline{P}^{(-\mathrm{supp}(\mathbf{T}))}) - H(\overline{P}^{(-\mathrm{supp}(\mathbf{T})\cup\{e\})}) \right) - \left(H(\overline{P}^{(-\mathrm{supp}(\mathbf{S}))}) - H(\overline{P}^{(-\mathrm{supp}(\mathbf{S})\cup\{e\})}) \right) \right].
\end{aligned}$$

Since the map $S\mapsto H(\overline{P}^{(S)})$ is submodular (see item 4 of Theorem 2.10) and $\mathbf{S}\preceq\mathbf{T}$, then we have

$$\begin{aligned}
\left(H(\overline{P}^{(S_j\cup\{e\})}) - H(\overline{P}^{(S_j)}) \right) - \left(H(\overline{P}^{(T_j\cup\{e\})}) - H(\overline{P}^{(T_j)}) \right) &\ge 0, \\
\left(H(\overline{P}^{(-\mathrm{supp}(\mathbf{T}))}) - H(\overline{P}^{(-\mathrm{supp}(\mathbf{T})\cup\{e\})}) \right) - \left(H(\overline{P}^{(-\mathrm{supp}(\mathbf{S}))}) - H(\overline{P}^{(-\mathrm{supp}(\mathbf{S})\cup\{e\})}) \right) &\ge 0.
\end{aligned}$$

Therefore $\Delta_{e,j} f(\mathbf{S}) - \Delta_{e,j} f(\mathbf{T}) \ge 0$ and hence the map (57) is orthant submodular. □

In view of Theorem 2.9, since the map (57) is orthant submodular, then for any $\beta = \beta(\mathbf{w}) \in \mathbb{R}$, if $\mathbf{S} \preceq \mathbf{V}$, we have the following monotonically non-decreasing $(m-1)$-submodular function:

$$\begin{aligned} g(\mathbf{S}, \mathbf{w}) &:= f(\mathbf{S}) - \beta + \sum_{j=1}^{m-1} \sum_{e \in S_j} (f(V_1, \ldots, V_j, \ldots, V_{m-1})) - f(V_1, \ldots, V_j \backslash \{e\}, \ldots, V_{m-1})) \\ &= f(\mathbf{S}) - \beta + \sum_{i=1}^{n} \sum_{j=1}^{m-1} \sum_{e \in S_j} w_i \left[D_{\mathrm{KL}}^{\pi}(\overline{P}^{(V_j)} \| \overline{P}^{(V_j \backslash \{e\})} \otimes \overline{P}^{(e)}) - D_{\mathrm{KL}}^{\pi}(\overline{P}^{(-\mathrm{supp}(\mathbf{V}) \backslash \{e\})} \| \overline{P}^{(-\mathrm{supp}(\mathbf{V}))} \otimes \overline{P}^{(e)}) \right] \\ &= f(\mathbf{S}) - \beta + \sum_{j=1}^{m-1} \sum_{e \in S_j} \left[D_{\mathrm{KL}}^{\pi}(\overline{P}^{(V_j)} \| \overline{P}^{(V_j \backslash \{e\})} \otimes \overline{P}^{(e)}) - D_{\mathrm{KL}}^{\pi}(\overline{P}^{(-\mathrm{supp}(\mathbf{V}) \backslash \{e\})} \| \overline{P}^{(-\mathrm{supp}(\mathbf{V}))} \otimes \overline{P}^{(e)}) \right], \end{aligned} \tag{58}$$

where the last equality comes from the fact that $\mathbf{w} \in \mathcal{S}_n$.

We also obtain the following modular function:

$$c(\mathbf{S}, \mathbf{w}) = -\beta + \sum_{j=1}^{m-1} \sum_{e \in S_j} \left[D_{\mathrm{KL}}^{\pi}(\overline{P}^{(V_j)} \| \overline{P}^{(V_j \backslash \{e\})} \otimes \overline{P}^{(e)}) - D_{\mathrm{KL}}^{\pi}(\overline{P}^{(-\mathrm{supp}(\mathbf{V}) \backslash \{e\})} \| \overline{P}^{(-\mathrm{supp}(\mathbf{V}))} \otimes \overline{P}^{(e)}) \right], \tag{59}$$

where we take

$$\beta = \beta(\mathbf{w}) \leq - \sum_{j=1}^{m-1} \sum_{e \in S_j} \left[H(\overline{P}(\mathbf{w})^{(-\mathrm{supp}(\mathbf{V}) \cup \{e\})}) + H(\overline{P}(\mathbf{w})^{(e)}) \right] \tag{60}$$

and write $c(\mathbf{S}, \mathbf{w}) \leq C$ to ensure that $0 \leq c \leq C$. Therefore, for fixed $\mathbf{w} \in \mathcal{S}_n$,

$$f(\mathbf{S}, \mathbf{w}) = g(\mathbf{S}, \mathbf{w}) - c(\mathbf{S}, \mathbf{w}),$$

where f can be written as the difference between a $(m-1)$-submodular function and a non-negative modular function.

Remark 12.2. *If we consider the optimization problem* (56) *with fixed* $\mathbf{w} \in \mathcal{S}_n$*, i.e.,*

$$\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq l} f(\mathbf{S}) = f(\mathbf{S}, \mathbf{w}),$$

we can apply the generalized distorted greedy algorithm (Algorithm 3) with g *as in* (58)*,* c *as in* (59)*, and* β *as in* (60) *to solve the problem. Furthermore, Theorem 2.16 gives the following lower bound:*

$$f(\mathbf{S}_l, \mathbf{w}) \geq (1 - e^{-1}) g(\mathbf{OPT}, \mathbf{w}) - c(\mathbf{OPT}, \mathbf{w}),$$

where $\mathbf{S}_l = (S_{l,1}, \ldots, S_{l,m-1})$ *is the final output of Algorithm 3 and* $\mathbf{OPT} = \arg\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq l} f(\mathbf{S})$.

We propose Algorithm 6 to solve problem (56). Algorithm 6 is a two-layer subgradient-greedy algorithm, which combines the outer generalized distorted greedy algorithm (Algorithm 3) and the inner projected subgradient algorithm (Algorithm 5). Specifically, we conduct totally l rounds of generalized distorted greedy algorithm: at the i-th round, we first fix $\mathbf{S}_i$ and apply the projected subgradient algorithm on fixed $\mathbf{S}_i$ for K iterations to maximize the objective function $f(\mathbf{S}_i, \cdot)$; we then fix $\overline{\mathbf{w}}_{i+1} = \sum_{k=1}^{K} \mathbf{w}_{i+1}^{(k)}$ and perform generalized distorted greedy algorithm to obtain $\mathbf{S}_{i+1}$. We proceed to state and prove a lower bound of Algorithm 6 in Theorem 12.3.

Theorem 12.3 (Lower bound of Algorithm 6)**.** *Algorithm 6 provides the following lower bound:*

$$f(\mathbf{S}_l, \overline{\mathbf{w}}_l) > \frac{1}{l} \sum_{i=1}^{l} [\alpha_i g(\mathbf{OPT}(\overline{\mathbf{w}}_i), \overline{\mathbf{w}}_i) - c(\mathbf{OPT}(\overline{\mathbf{w}}_i), \overline{\mathbf{w}}_i)] - \mathcal{O}\left(l \left(\sqrt{\frac{nB}{K}} + C \right) \right),$$

where $(\mathbf{S}_l, \overline{\mathbf{w}}_l)$ *is the output of Algorithm 6,* $\alpha_i = (1 - \frac{1}{l})^{l-i}$*, and*

$$\mathbf{OPT}(\mathbf{w}) = \arg\max_{\mathbf{S} \preceq \mathbf{V};\ |\mathrm{supp}(\mathbf{S})| \leq l} f(\mathbf{S}, \mathbf{w}).$$

Algorithm 6 A two-layer subgradient-greedy algorithm to solve problem (56)

Require: f as in (55); g as in (58); c as in (59); subgradient $\mathbf{g}$ as in Theorem 11.1; cardinality constraint l; partition of ground set $\mathbf{V} = (V_1, \ldots, V_{m-1}) \in m^{[\![d]\!]}$; inner iteration number K

Initialize $\mathbf{S}_0 = (S_{0,1}, \ldots, S_{0,m-1}) \leftarrow \emptyset$ and $\mathbf{w}_0^{(K)} = (\frac{1}{m}, \ldots, \frac{1}{m})$
Compute bound B as in Theorem 11.1 and stepsize $\eta = \sqrt{\frac{n}{BK}}$
for $i = 0$ to $l-1$ **do**
  $\mathbf{w}_{i+1}^{(0)} \leftarrow \mathbf{w}_i^{(K)}$
  for $k = 0$ to $K-1$ **do**
    $\mathbf{v} \leftarrow \mathbf{w}_{i+1}^{(k)} - \eta \cdot \mathbf{g}(\mathbf{S}_i, \mathbf{w}_{i+1}^{(k)})$
    $\mathbf{w}_{i+1}^{(k+1)} \leftarrow \arg\min_{\mathbf{w} \in \mathcal{S}_n} \|\mathbf{w} - \mathbf{v}\|_2^2$
  end for
  $\overline{\mathbf{w}}_{i+1} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \mathbf{w}_{i+1}^{(k)}$
  $(j^*, e^*) \leftarrow \arg\max_{j \in [\![m-1]\!];\ e \in V_j \setminus S_{i,j}} \left\{ \left(1 - \frac{1}{l}\right)^{l-(i+1)} \Delta_{e,j} g(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - c(\{e\}, \overline{\mathbf{w}}_{i+1}) \right\}$
  if $\left(1 - \frac{1}{l}\right)^{l-(i+1)} \Delta_{e^*,j^*} g(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - c(\{e^*\}, \overline{\mathbf{w}}_{i+1}) > 0$ **then**
    $S_{i+1,j^*} \leftarrow S_{i,j^*} \cup \{e^*\}$
  else
    $S_{i+1,j^*} \leftarrow S_{i,j^*}$
  end if
  for $k \in [\![m-1]\!],\ k \neq j^*$ **do**
    $S_{i+1,k} \leftarrow S_{i,k}$
  end for
end for
Output: $\mathbf{S}_l$ and $\overline{\mathbf{w}}_l$

Proof. We define the distorted objective function $\Phi_i : m^{[\![d]\!]} \times \mathcal{S}_n \to \mathbb{R}$ to be

$$\Phi_i(\mathbf{S}, \overline{\mathbf{w}}_i) := \alpha_i g(\mathbf{S}, \overline{\mathbf{w}}_i) - c(\mathbf{S}, \overline{\mathbf{w}}_i) > \alpha_i f(\mathbf{S}, \overline{\mathbf{w}}_i) - c(\mathbf{S}, \overline{\mathbf{w}}_i),$$

where the inequality comes from the fact that $0 < \alpha_i \leq 1$.

We look into the difference of the distorted objective function

$$\Phi_{i+1}(\mathbf{S}_{i+1}, \overline{\mathbf{w}}_{i+1}) - \Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_i) = [\Phi_{i+1}(\mathbf{S}_{i+1}, \overline{\mathbf{w}}_{i+1}) - \Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1})] - [\Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - \Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_i)],$$

where the first term is the gain in the distorted greedy algorithm, and the second term is the weight update error.

We first refer to the proof of Theorem 2.16 and state the lower bound of the gain in the distorted greedy part

$$\Phi_{i+1}(\mathbf{S}_{i+1}, \overline{\mathbf{w}}_{i+1}) - \Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) \geq \frac{1}{l}(\alpha_{i+1} g(\mathbf{OPT}(\overline{\mathbf{w}}_{i+1}), \overline{\mathbf{w}}_{i+1}) - c(\mathbf{OPT}(\overline{\mathbf{w}}_{i+1}), \overline{\mathbf{w}}_{i+1})).$$

We then analyze the weight update error term. From Theorem 11.2, we have

$$f(\mathbf{S}_i, \mathbf{w}^*(\mathbf{S}_i)) - f(\mathbf{S}_i, \overline{\mathbf{w}}_m) \leq \sqrt{\frac{nB}{K}},\ \forall m \in [\![l]\!].$$

hence the lower bound of the weight update error is

$$\begin{aligned}
\Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - \Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_i) &= \alpha_i(f(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - f(\mathbf{S}_i, \overline{\mathbf{w}}_i)) - (c(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - c(\mathbf{S}_i, \overline{\mathbf{w}}_i)) \\
&> -\alpha_i \|f(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - f(\mathbf{S}_i, \overline{\mathbf{w}}_i)\| - C \\
&\geq -\alpha_i(\|f(\mathbf{S}_i, \mathbf{w}^*(\mathbf{S}_i)) - f(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1})\| + \|f(\mathbf{S}_i, \mathbf{w}^*(\mathbf{S}_i)) - f(\mathbf{S}_i, \overline{\mathbf{w}}_i)\|) - C \\
&\geq -2\alpha_i \sqrt{\frac{nB}{K}} - C.
\end{aligned}$$

Since $\Phi_0(\mathbf{S}_0) \geq 0$, then

$$f(\mathbf{S}_l, \overline{\mathbf{w}}_l) = \alpha_l \cdot g(\mathbf{S}_l, \overline{\mathbf{w}}_i) - c(\mathbf{S}_l, \overline{\mathbf{w}}_i) \geq \sum_{i=0}^{l-1} [\Phi_{i+1}(\mathbf{S}_{i+1}) - \Phi_i(\mathbf{S}_i)],$$

hence

$$
\begin{aligned}
f(\mathbf{S}_l, \overline{\mathbf{w}}_l) &\geq \sum_{i=0}^{l-1} [\Phi_{i+1}(\mathbf{S}_{i+1}, \overline{\mathbf{w}}_{i+1}) - \Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1})] + \sum_{i=0}^{l-1} [\Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_{i+1}) - \Phi_i(\mathbf{S}_i, \overline{\mathbf{w}}_i)] \\
&> \frac{1}{l} \sum_{i=1}^{l} [\alpha_i g(\mathbf{OPT}(\overline{\mathbf{w}}_i), \overline{\mathbf{w}}_i) - c(\mathbf{OPT}(\overline{\mathbf{w}}_i), \overline{\mathbf{w}}_i)] - 2\sqrt{\frac{nB}{K}} \sum_{i=0}^{l-1} \alpha_i - lC \\
&= \frac{1}{l} \sum_{i=1}^{l} [\alpha_i g(\mathbf{OPT}(\overline{\mathbf{w}}_i), \overline{\mathbf{w}}_i) - c(\mathbf{OPT}(\overline{\mathbf{w}}_i), \overline{\mathbf{w}}_i)] - \mathcal{O}\left(l\left(\sqrt{\frac{nB}{K}} + C\right)\right).
\end{aligned}
$$

□

13 Numerical experiments of Part II[2]

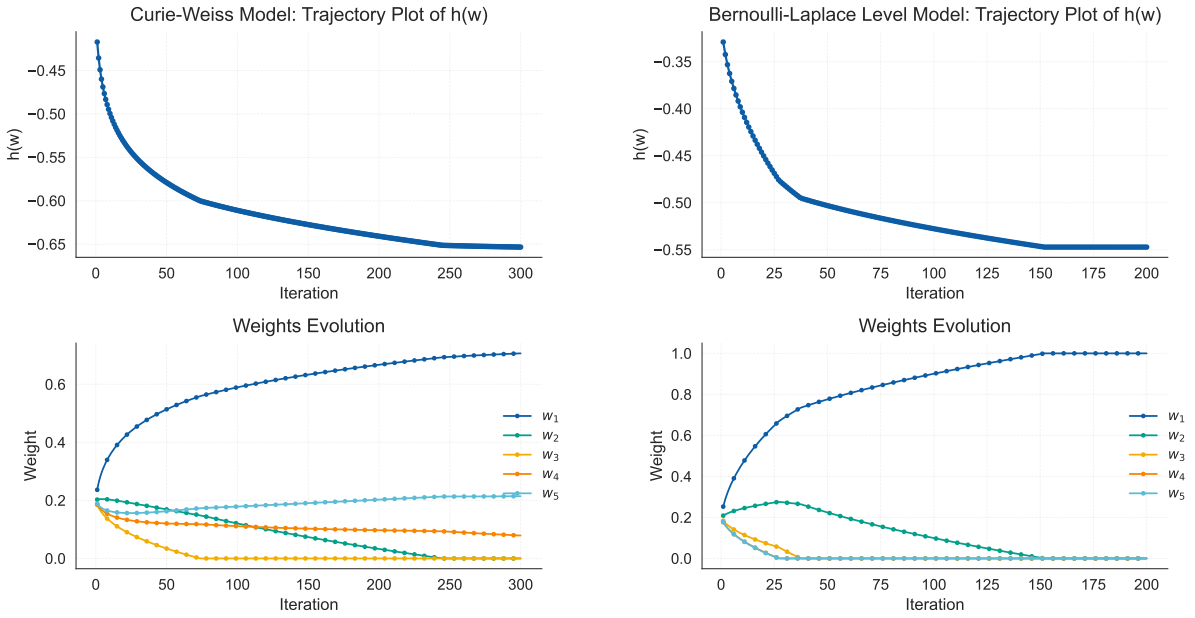
We conduct a series of numerical experiments to validate the theoretical framework and evaluate the performance of the proposed algorithms on the Curie-Weiss model and the Bernoulli-Laplace level model (see Section 2.4 for details). The experiments are designed to demonstrate the performance of the projected subgradient algorithm (Algorithm 5) to solve problem (54) and the two-layer subgradient-greedy algorithm (Algorithm 6) to solve problem (56).

13.1 Numerical experiments of Algorithm 5

We apply the projected subgradient algorithm (Algorithm 5) to solve the minimization problem (54) for both the Curie-Weiss and Bernoulli-Laplace level models. We start with a low-dimensional example. For both settings, we construct a 5-dimensional Markov chain with π-stationary transition probability matrix P on state space $\mathcal{X} = \{0, 1\}^5$. We then construct a family of $n = 5$ transition matrices with $\mathcal{B} = \{P, P^2, P^4, P^8, P^{16}\}$, which ensures that all matrices in $\mathcal{B}$ share the same stationary distribution π. We partition the state space into $\mathbf{S} = \{S_1, S_2, S_3\}$ ($m = 3$) such that $S_1 = \{1, 2\}$, $S_2 = \{3, 5\}$, and $S_3 = \{4\}$.

We initialize the algorithm with uniform weights $\mathbf{w}^{(0)} = (1/5, \ldots, 1/5)$. The step size is chosen according to the theoretical guarantee from Theorem 11.2, $\eta = \sqrt{\frac{n}{Bt}}$, where the subgradient norm bound B is estimated once at the beginning of the algorithm. The number of iterations until convergence is theoretically determined by $t = \lceil \frac{nB}{\epsilon^2} \rceil$, but t would be large with large B and small ϵ. Therefore for practical purpose, we only run a small number of iterations for demonstration. The trajectory plots of the projected subgradient algorithm and the evolution of weights of both models are shown in Figure 11. We also summarize the weights $\mathbf{w} \in \mathcal{S}_n$ and the corresponding objective value $h(\mathbf{w})$ in Table 20 for both Curie-Weiss and Bernoulli-Laplace models. We state and compare the optimal $\mathbf{w}$ during the optimization process $\arg\min_{i \in [\![t]\!]} h(\mathbf{w}^{(i)})$, the averaged value during the iterations $\overline{\mathbf{w}}^t$, initial uniform $\mathbf{w}^{(0)}$, extreme weight $\mathbf{w}_{\text{ex}}$ such that only $\mathbf{w}_{\text{ex},0} = 1$, and the final weight $\mathbf{w}^{(t)}$ of the iterations.

[2]The code is available at: `https://github.com/zheyuanlai/subgradient-greedy`.



(a) Curie-Weiss model

(b) Bernoulli-Laplace level model

Figure 11: Convergence of the projected subgradient algorithm for both models ($d = 5$).

$\mathbf{w}$, $h(\mathbf{w})$ / **Model**	**Curie-Weiss**	**Bernoulli-Laplace**
$\arg\min_{i\in[\![t]\!]} h(\mathbf{w}^{(i)})$	$(0.71, 0.00, 0.00, 0.08, 0.21)$	$(1.00, 0.00, 0.00, 0.00, 0.00)$
$\overline{\mathbf{w}}^t$	$(0.60, 0.08, 0.02, 0.11, 0.19)$	$(0.85, 0.11, 0.02, 0.01, 0.01)$
$\mathbf{w}^{(0)}$	$(0.20, 0.20, 0.20, 0.20, 0.20)$	$(0.20, 0.20, 0.20, 0.20, 0.20)$
$\mathbf{w}_{\text{ex}}$	$(1.00, 0.00, 0.00, 0.00, 0.00)$	$(1.00, 0.00, 0.00, 0.00, 0.00)$
$\mathbf{w}^{(t)}$	$(0.71, 0.00, 0.00, 0.08, 0.21)$	$(1.00, 0.00, 0.00, 0.00, 0.00)$
$\min_{i\in[\![t]\!]} h(\mathbf{w}^{(i)})$	-0.65	-0.55
$h(\overline{\mathbf{w}}^t)$	-0.62	-0.51
$h(\mathbf{w}^{(0)})$	-0.39	-0.31
$h(\mathbf{w}_{\text{ex}})$	-0.48	-0.55
$h(\mathbf{w}^{(t)})$	-0.65	-0.55

Table 20: Comparison of $h(\mathbf{w})$ values for different weight choices ($d = 5$)

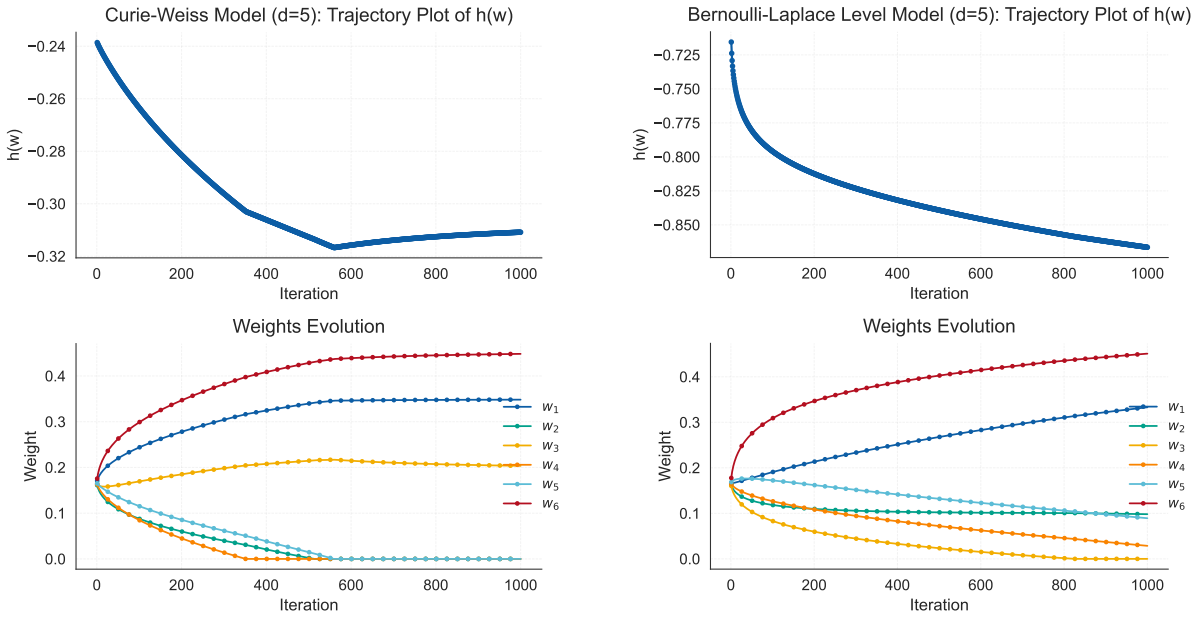
For the Curie-Weiss model (Figure 11a), the algorithm demonstrates rapid initial decrease, after the first 50 iterations, the objective value decreases with a slower rate, which totally converges after 250 iterations. The weights converge to a sparse distribution, with the final weight vector being approximately $\mathbf{w}^{(t)} = (0.71, 0.00, 0.00, 0.08, 0.21)$. This indicates that the final solution is approximately a convex combination of the base transition matrix P and the transition matrix with the highest mixing rate P^{16}, while the intermediate transition matrices have zero weights.

The Bernoulli-Laplace level model (Figure 11b) exhibits similar convergence behavior: the objective value decreases fast in the first 30 steps, then it moves slowly until fully converged after 150 iterations. The final weight vector converges to $\mathbf{w}^{(t)} = (1.00, 0.00, 0.00, 0.00, 0.00)$, indicating that the optimal solution is entirely the base transition matrix P.

We then conduct experiments associated with the family of transition matrices including lazy Markov chain (see e.g. (Shen et al., 2014) for background). Precisely, we choose

$$\mathcal{B} = \left\{P, P^2, P^4, \frac{1}{4}I + \frac{3}{4}P, \frac{1}{2}(I + P), \frac{3}{4}I + \frac{1}{4}P\right\},$$

where one readily verifies that all the transition matrices in family $\mathcal{B}$ share the same stationary distribution π. The trajectory plots are shown in Figure 12, and we also summarize the objective values of different $\mathbf{w}$'s in Table 21.



(a) Curie-Weiss model

(b) Bernoulli-Laplace level model

Figure 12: Trajectory plot of the projected subgradient algorithm for both models (incl. lazy chains).

$\mathbf{w}$, $h(\mathbf{w})$ / **Model**	**Curie-Weiss**	**Bernoulli-Laplace**
$\arg\min_{i\in[\![t]\!]} h(\mathbf{w}^{(i)})$	$(0.35, 0.00, 0.22, 0.00, 0.00, 0.44)$	$(0.33, 0.10, 0.00, 0.03, 0.09, 0.45)$
$\overline{\mathbf{w}}^t$	$(0.32, 0.03, 0.20, 0.02, 0.04, 0.40)$	$(0.26, 0.11, 0.03, 0.08, 0.13, 0.39)$
$\mathbf{w}^{(0)}$	$(0.17, 0.17, 0.17, 0.17, 0.17, 0.17)$	$(0.17, 0.17, 0.17, 0.17, 0.17, 0.17)$
$\mathbf{w}_{\text{ex}}$	$(1.00, 0.00, 0.00, 0.00, 0.00, 0.00)$	$(1.00, 0.00, 0.00, 0.00, 0.00, 0.00)$
$\mathbf{w}^{(t)}$	$(0.35, 0.00, 0.20, 0.00, 0.00, 0.45)$	$(0.33, 0.10, 0.00, 0.03, 0.09, 0.45)$
$\min_{i\in[\![t]\!]} h(\mathbf{w}^{(i)})$	-0.32	-0.87
$h(\overline{\mathbf{w}}^t)$	-0.34	-0.31
$h(\mathbf{w}^{(0)})$	-0.28	-0.29
$h(\mathbf{w}_{\text{ex}})$	-0.29	-0.55
$h(\mathbf{w}^{(t)})$	-0.31	-0.87

Table 21: Comparison of $h(\mathbf{w})$ values for different weight choices (incl. lazy chains)

For the Curie-Weiss model (Figure 12a), the algorithm exhibits an initial decrease followed by a slight increase towards convergence. Since the projected subgradient algorithm (Algorithm 5) is not a descent algorithm, then it is not guaranteed that h shows a non-decreasing trajectory. The final objective value reaches approximately -0.311, while the final weight learned by the algorithm is

$$\mathbf{w}^{(t)} = \Big(\underbrace{0.35}_{P}, \underbrace{0.00}_{P^2}, \underbrace{0.20}_{P^4}, \underbrace{0.00}_{\frac{1}{4}I+\frac{3}{4}P}, \underbrace{0.00}_{\frac{1}{2}(I+P)}, \underbrace{0.45}_{\frac{3}{4}I+\frac{1}{4}P}\Big),$$

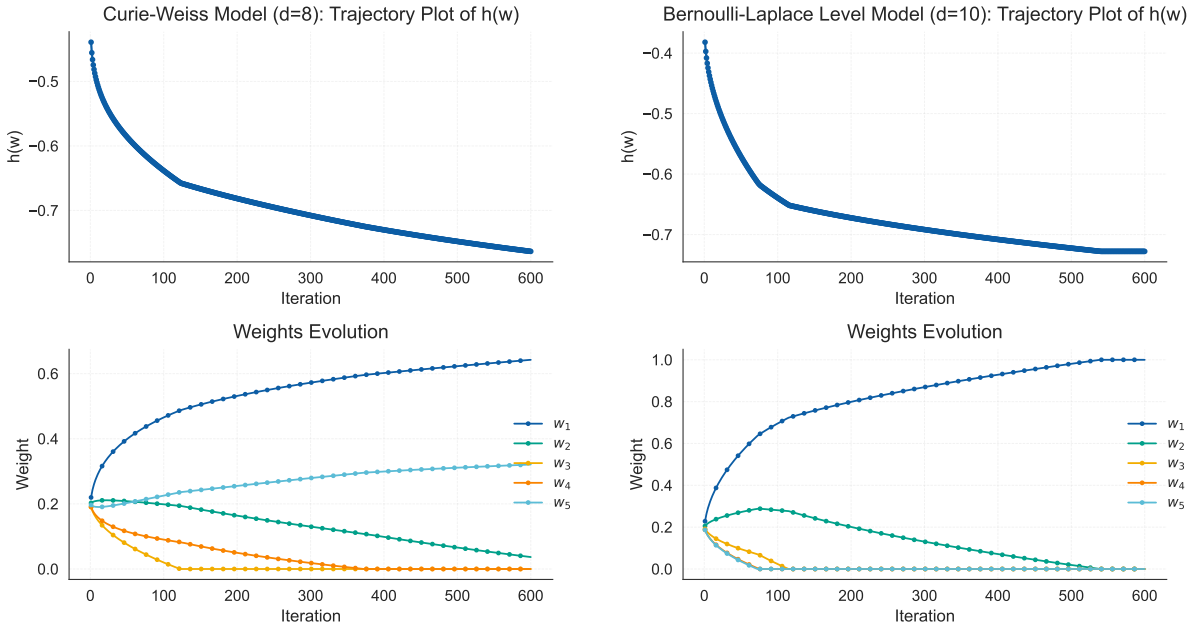
which is sparse and concentrates on three extremes: the base chain P, the most accelerated P^4, and the "laziest" member $\frac{3}{4}I + \frac{1}{4}P$. Intermediate options ($P^2$ and the moderately lazy mixtures) receive zero weight. This indicates that, within this family on the Curie-Weiss chain, the best trade-off for the minimax optimization is achieved by combining the slowest $\frac{3}{4}I + \frac{1}{4}P$ and fastest P^4 directions with the base chain P.

For the Bernoulli–Laplace level model (Figure 12b), we similarly observe rapid early descent and a stable plateau thereafter as in Figure 11b. The final objective is approximately -0.866 though has not reached convergence given the limited computational budget. The final weight is

$$\mathbf{w}^{(t)} = \Big(\underbrace{0.33}_{P}, \underbrace{0.10}_{P^2}, \underbrace{0.00}_{P^4}, \underbrace{0.03}_{\frac{1}{4}I+\frac{3}{4}P}, \underbrace{0.09}_{\frac{1}{2}(I+P)}, \underbrace{0.45}_{\frac{3}{4}I+\frac{1}{4}P}\Big),$$

which gives majority of weight on the base transition matrix P and the transition matrix associated with the most "lazy" chain $\frac{3}{4}I + \frac{1}{4}P$. This indicates that, within this family on the Bernoulli-Laplace chains, the best trade-off for the minimax optimization is achieved by combining the slowest direction $\frac{3}{4}I + \frac{1}{4}P$ and P^2 direction with the base chain P.

We proceed to simulate on higher-dimensional Markov chains associated with both models, with results presented in Figure 13. For these experiments, the family of transition matrices is $\mathcal{B} = \{P, P^2, P^4, P^8, P^{16}\}$ ($n = 5$). For the Bernoulli-Laplace level model, we conduct experiments on $d = 10$, while for the Curie-Weiss model, we only choose $d = 8$ in order to avoid numerical overflow. We also summarize the objective values of different $\mathbf{w}$'s in Table 22.



(a) Curie-Weiss model ($d = 8$)

(b) Bernoulli-Laplace level model ($d = 10$)

Figure 13: Trajectory plots of the projected subgradient algorithm for both models (higher dimension).

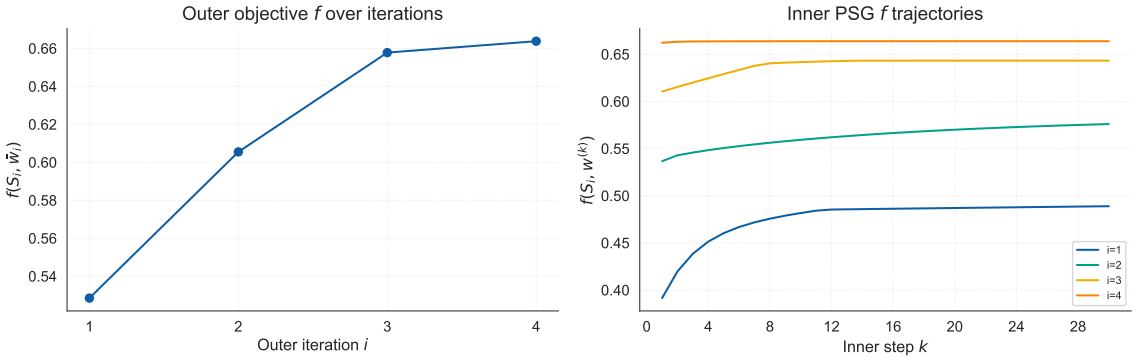
$\mathbf{w}$, $h(\mathbf{w})$ / **Model**	**Curie-Weiss**	**Bernoulli-Laplace**
$\arg\min_{i\in[\![t]\!]} h(\mathbf{w}^{(i)})$	$(0.64, 0.04, 0.00, 0.00, 0.32)$	$(1.00, 0.00, 0.00, 0.00, 0.00)$
$\overline{\mathbf{w}}^t$	$(0.55, 0.13, 0.01, 0.04, 0.27)$	$(0.83, 0.14, 0.02, 0.01, 0.01)$
$\mathbf{w}^{(0)}$	$(0.20, 0.20, 0.20, 0.20, 0.20)$	$(0.20, 0.20, 0.20, 0.20, 0.20)$
$\mathbf{w}_{\text{ex}}$	$(1.00, 0.00, 0.00, 0.00, 0.00)$	$(1.00, 0.00, 0.00, 0.00, 0.00)$
$\mathbf{w}^{(t)}$	$(0.64, 0.04, 0.00, 0.00, 0.32)$	$(1.00, 0.00, 0.00, 0.00, 0.00)$
$\min_{i\in[\![t]\!]} h(\mathbf{w}^{(i)})$	-0.76	-0.73
$h(\overline{\mathbf{w}}^t)$	-0.69	-0.67
$h(\mathbf{w}^{(0)})$	-0.44	-0.38
$h(\mathbf{w}_{\text{ex}})$	-0.27	-0.73
$h(\mathbf{w}^{(t)})$	-0.76	-0.73

Table 22: Comparison of $h(\mathbf{w})$ values for different weight choices (higher dimension)

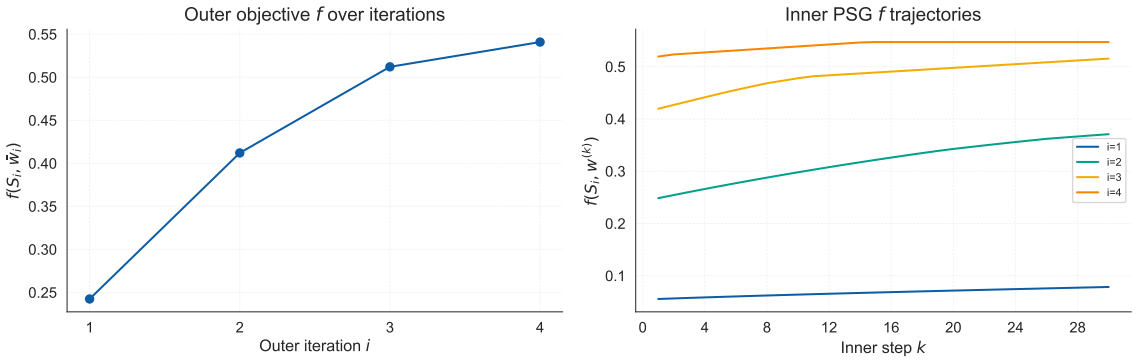
The experiments associated with the Bernoulli-Laplace level model (Figure 13b) exhibit similar trends as the 5-dimensional example (Figure 11b), as the objective value $h(\mathbf{w})$ decreases fast at start and then converges slower towards $\mathbf{w}^{(t)} = (1.00, 0.00, 0.00, 0.00, 0.00)$. For the Curie-Weiss model, the 8-dimensional example (Figure 13a) shows similar convergence trend as the 5-dimensional example (Figure 11a). However, as the B in Theorem 11.2 is large, we do not obtain the exact converging $\mathbf{w}^*$ with the same computational budget as the Bernoulli-Laplace model.

13.2 Numerical experiments of Algorithm 6

We apply Algorithm 6 to solve the maximization problem (56) on both the Curie-Weiss and Bernoulli-Laplace models. For both models, we construct a 5-dimensional Markov chain with state space $\mathcal{X} = \{0,1\}^5$ and π-stationary transition matrix P. We then construct $\mathcal{B} = \{P, P^2, P^4, P^8, P^{16}\}$ so that all matrices in $\mathcal{B}$ share the same stationary distribution π. We choose the ground set to be $\mathbf{V} = \{V_1, V_2\}$ such that $V_1 = \{1,2\}$ and $V_2 = \{3,5\}$. For the inner part, we execute $K = 30$ iterations of the projected subgradient algorithm. We summarize the running results of both models in Figure 14.



(a) Curie-Weiss model



(b) Bernoulli-Laplace level model

Figure 14: Trajectory plot of Algorithm 6 for both models ($d = 5$).

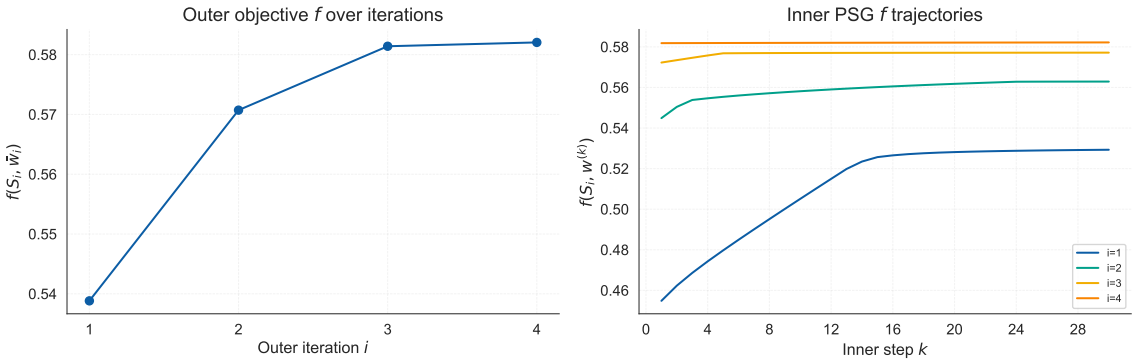
For the Curie-Weiss model (Figure 14a), the final weight is $\overline{\mathbf{w}}_l = (0.72, 0.00, 0.00, 0.00, 0.28)$, and the final partition set is $\mathbf{S}_l = \{S_1, S_2\}$, where $S_1 = \{2\}$ and $S_2 = \{3,5\}$. It shows that after the final round of Algorithm 6, the resultant weight vector of the max-min-max optimization problem is attained by combining the base transition matrix P and the transition matrix with the highest mixing rate P^{16}.

For the Bernoulli-Laplace level model (Figure 14b), the final weight is $\overline{\mathbf{w}}_l = (0.97, 0.03, 0.00, 0.00, 0.00)$, and the final partition set is $\mathbf{S}_l = \{S_1, S_2\}$, where $S_1 = \{2\}$ and $S_2 = \{3,5\}$. It shows that after the final round of Algorithm 6, the convex hull of family $\mathcal{B}$ concentrates on the base transition matrix P.

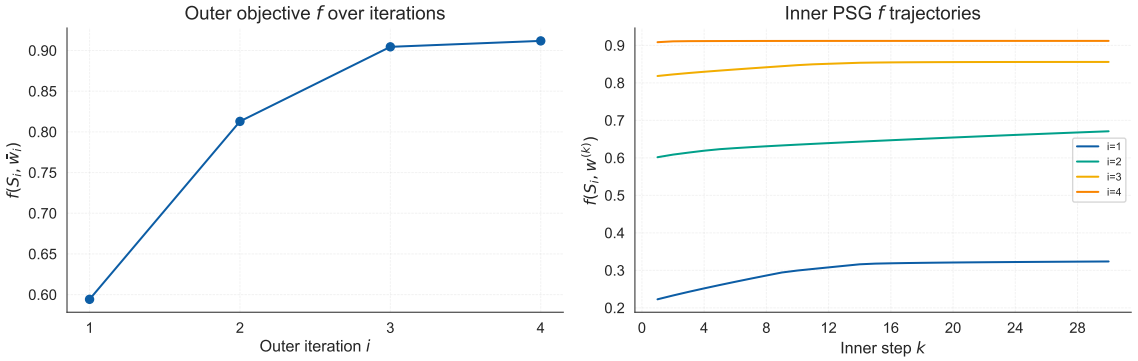
Similar to the numerical experiments in Section 13.1, we then look into the experiments associated with the family of transition matrices including lazy random walk, precisely, we choose

$$\mathcal{B} = \left\{P, P^2, P^4, \frac{1}{4}I + \frac{3}{4}P, \frac{1}{2}(I+P), \frac{3}{4}I + \frac{1}{4}P\right\}.$$

We summarize the results in Figure 15.



(a) Curie-Weiss model



(b) Bernoulli-Laplace level model

Figure 15: Trajectory plot of Algorithm 6 for both models (incl. lazy matrices).

For the Curie-Weiss model (Figure 15a), the final weight is

$$\overline{\mathbf{w}}_l = \Big(\underbrace{0.37}_{P}, \underbrace{0.00}_{P^2}, \underbrace{0.33}_{P^4}, \underbrace{0.00}_{\frac{1}{4}I+\frac{3}{4}P}, \underbrace{0.00}_{\frac{1}{2}(I+P)}, \underbrace{0.30}_{\frac{3}{4}I+\frac{1}{4}P} \Big),$$

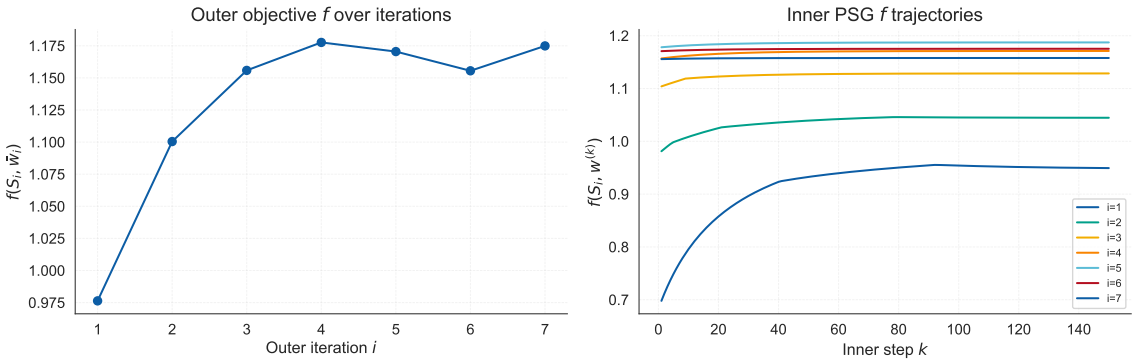
and the final partition set is $\mathbf{S}_l = \{S_1, S_2\}$, where $S_1 = \{2\}$ and $S_2 = \{3, 5\}$. The final weight vector $\overline{\mathbf{w}}_l$ concentrates on three modes, which indicates that the final weight is obtained by combining the slowest $\frac{3}{4}I + \frac{1}{4}P$ and the fastest P^4 directions with the base chain P.

For the Bernoulli-Laplace level model (Figure 15b), the final weight is

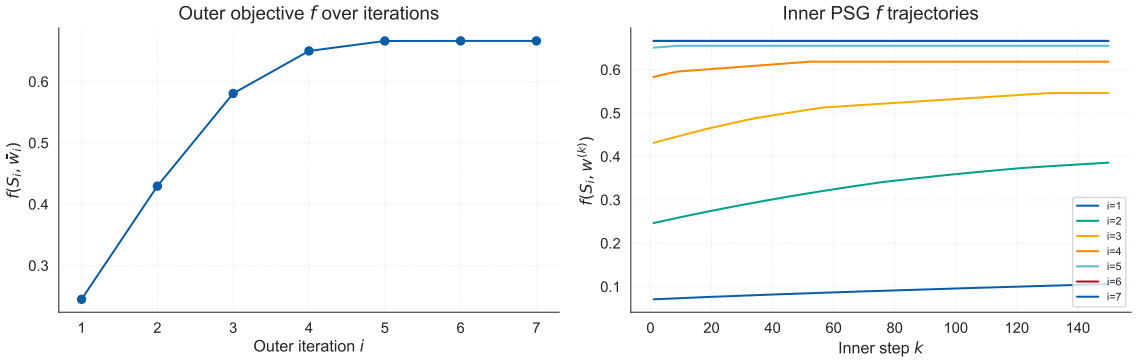
$$\overline{\mathbf{w}}_l = \Big(\underbrace{0.50}_{P}, \underbrace{0.00}_{P^2}, \underbrace{0.00}_{P^4}, \underbrace{0.00}_{\frac{1}{4}I+\frac{3}{4}P}, \underbrace{0.00}_{\frac{1}{2}(I+P)}, \underbrace{0.50}_{\frac{3}{4}I+\frac{1}{4}P} \Big),$$

and the final partition set is $\mathbf{S}_l = \mathbf{V}$, which means that Algorithm 6 selects the whole ground set as the subset. The final output $\overline{\mathbf{w}}_l$ concentrates on two matrices, which indicates that the final result is obtained by averaging the chain with the slowest mixing rate $\frac{3}{4}I + \frac{1}{4}P$ and the base chain P.

We proceed to analyze higher-dimensional cases of both models with $d = 8$ and cardinality constraint $l = 7$, and choose the ground set as $\mathbf{V} = \{V_1, V_2\}$, where $V_1 = \{1, 2, 3, 4\}$ and $V_2 = \{5, 6, 7\}$. We choose the family of the transition probability matrices to be $\mathcal{B} = \{P, P^2, P^4, P^8, P^{16}\}$. For the inner part, we execute $K = 150$ iterations of the projected subgradient algorithm. The trajectory plots of both models are summarized in Figure 16.



(a) Curie-Weiss model



(b) Bernoulli-Laplace level model

Figure 16: Trajectory plot of Algorithm 6 for both models ($d = 8$).

For the Curie-Weiss model (Figure 16a), the objective value $f(\mathbf{S}_i, \overline{\mathbf{w}}_i)$ is not monotonically non-decreasing, as both the generalized distorted greedy algorithm (Algorithm 3) and the projected subgradient algorithm (Algorithm 5) do not guarantee monotonicity. The final partition set is $\mathbf{S}_l = \mathbf{V}$, which means that the algorithm selects the ground set as the subset. After the final round of Algorithm 6, the final weight is $\overline{\mathbf{w}}_l = (0.70, 0.00, 0.00, 0.00, 0.30)$, which concentrates on the base transition matrix P and the matrix with fastest mixing P^{16}.

For the Bernoulli-Laplace level model (Figure 16b), the final weight is $\overline{\mathbf{w}}_l = (1.00, 0.00, 0.00, 0.00, 0.00)$ and the final partition set is $\mathbf{S}_l = \{S_1, S_2\}$, where $S_1 = \{1, 2, 3\}$ and $S_2 = \{5, 6, 7\}$. It shows that after the final round of Algorithm 6, the weight of the max-min-max optimization reaches closely to the base transition matrix P.

Acknowledgements

Zheyuan Lai acknowledges Professor Michael Choi, for his invaluable guidance and support throughout this project.

References

A. Beck. *First-Order Methods in Optimization*. SIAM, Philadelphia, PA, 2017.

I. Bogunovic, S. Mitrović, J. Scarlett, and V. Cevher. Robust submodular maximization: A non-uniform partitioning approach. In D. Precup and Y. W. Teh, editors, *Proc. Int. Conf. Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 508–516, Sydney, Australia, 2017. PMLR.

A. Bovier and F. Den Hollander. *Metastability: a potential-theoretic approach*, volume 351. Springer, 2016.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

M. C. H. Choi and G. Wolfer. Markov chain entropy games and the geometry of their Nash equilibria. *ALEA Lat. Am. J. Probab. Math. Stat.*, 22(2):925–, 2025.

M. C. H. Choi, Y. Wang, and G. Wolfer. Geometry and factorization of multivariate Markov chains with applications to MCMC acceleration. *arXiv preprint arXiv:2404.12589*, 2024.

L. Condat. Fast projection onto the simplex and the l_1 ball. *Math. Program.*, 158(1):575–585, 2016.

I. Csiszár. A class of measures of informativity of observation channels. *Period. Math. Hung.*, 2(1–4): 191–213, 1972.

A. Ene and H. Nguyen. Streaming algorithm for monotone k-submodular maximization with cardinality constraints. In *International Conference on Machine Learning*, pages 5944–5967. PMLR, 2022.

U. Feige, V. S. Mirrokni, and J. Vondrák. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

B. C. Geiger and C. Temmel. Lumpings of Markov chains, entropy rate preservation, and higher-order lumpability. *Journal of Applied Probability*, 51(4):1114–1132, 2014.

A. A. Gushchin and D. A. Zhdanov. A minimax result for f-divergences. In Y. Kabanov, R. Liptser, and J. Stoyanov, editors, *From Stochastic Calculus to Mathematical Finance: The Shiryaev Festschrift*, pages 287–294. Springer, Berlin, Heidelberg, 2006.

H. Hafez-Kolahi, B. Moniri, and S. Kasaei. Information-theoretic analysis of minimax excess risk. *IEEE Trans. Inf. Theory*, 69:4659–4674, 2022.

C. Harshaw, M. Feldman, J. Ward, and A. Karbasi. Submodular maximization beyond non-negativity: Guarantees, fast algorithms, and applications. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2634–2643. PMLR, 09–15 Jun 2019.

D. Haussler. A general minimax result for relative entropy. *IEEE Trans. Inf. Theory*, 43(4):1276–1280, 1997.

J. Jagalur-Mohan and Y. Marzouk. Batch greedy maximization of non-submodular functions: Guarantees and applications to experimental design. *Journal of Machine Learning Research*, 22(252):1–62, 2021.

K. Khare and H. Zhou. Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Ann. Appl. Probab.*, 19(2):737–777, 2009.

B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms*. Springer, Berlin, 4th edition, 2008.

D. Lacker. Independent projections of diffusions: Gradient flows for variational inference and optimal mean field approximations. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 2025. to appear.

J. Lee, M. Sviridenko, and J. Vondrák. Submodular maximization over multiple matroids via generalized exchange properties. *Mathematics of Operations Research*, 35(4):795–806, 2010.

D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Math. Program.*, 14(1):265–294, 1978.

J. B. Orlin, A. S. Schulz, and R. Udwani. Robust monotone submodular function maximization. *Math. Program.*, 172(1):505–537, 2018.

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, MA, 1994.

Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, Cambridge, 2025.

J. Shen, Y. Du, W. Wang, and X. Li. Lazy random walks for superpixel segmentation. *IEEE Trans. Image Process.*, 23(4):1451–1462, 2014.

M. Staib and S. Jegelka. Robust budget allocation via continuous submodular functions. *Appl. Math. Optim.*, pages 1–31, 2019.

J. Ward and S. Živnỳ. Maximizing k-submodular functions and beyond. *ACM Trans. Algorithms*, 12 (4):1–26, 2016.