Improving the convergence of Markov chains via permutations and projections

Michael Choi

Joint work with Max Hird (UCL) and Youjia Wang (NUS)

April 2025

Department of Statistics and Data Science
Faculty of Science

Introduction

1. Paper: arXiv:2411.08295, joint work with Max Hird (UCL) and Youjia Wang (NUS)
2. Much of my recent research focus is on information theory of Markov chains, in particular **information projections** and **information geometry** of Markov chains and MCMC algorithms.

Introduction

1. Paper: arXiv:2411.08295, joint work with Max Hird (UCL) and Youjia Wang (NUS)
2. Much of my recent research focus is on information theory of Markov chains, in particular **information projections** and **information geometry** of Markov chains and MCMC algorithms.
3. This paper is interesting from at least the following two perspectives. First, we study new information projections of Markov chains. This unveils previously unknown geometric structure in the space of transition matrices of Markov chains.

Introduction

1. Paper: arXiv:2411.08295, joint work with Max Hird (UCL) and Youjia Wang (NUS)
2. Much of my recent research focus is on information theory of Markov chains, in particular **information projections** and **information geometry** of Markov chains and MCMC algorithms.
3. This paper is interesting from at least the following two perspectives. First, we study new information projections of Markov chains. This unveils previously unknown geometric structure in the space of transition matrices of Markov chains.
4. Second, we use this discovered geometric structure "for good". Precisely, these information projections give rise to improved MCMC samplers, thus leading to new algorithmic development in MCMC.

Setup

1. Throughout this paper, we consider Markov chains on a finite state space $\mathcal{X}$.

Setup

1. Throughout this paper, we consider Markov chains on a finite state space $\mathcal{X}$.
2. We denote by $\mathcal{L} = \mathcal{L}(\mathcal{X})$ to be the set of transition matrices on $\mathcal{X}$.

Setup

1. Throughout this paper, we consider Markov chains on a finite state space $\mathcal{X}$.
2. We denote by $\mathcal{L} = \mathcal{L}(\mathcal{X})$ to be the set of transition matrices on $\mathcal{X}$.
3. We write $\mathcal{S}(\pi) \subseteq \mathcal{L}$ to be the set of π-stationary transition matrices, that is,
$$\mathcal{S}(\pi) := \{P \in \mathcal{L};\ \pi P = \pi\}.$$

Setup

1. Throughout this paper, we consider Markov chains on a finite state space $\mathcal{X}$.
2. We denote by $\mathcal{L} = \mathcal{L}(\mathcal{X})$ to be the set of transition matrices on $\mathcal{X}$.
3. We write $\mathcal{S}(\pi) \subseteq \mathcal{L}$ to be the set of π-stationary transition matrices, that is,
$$\mathcal{S}(\pi) := \{P \in \mathcal{L};\ \pi P = \pi\}.$$
4. For $P \in \mathcal{S}(\pi)$, we write P^* to be the time-reversal or the $\ell^2(\pi)$-adjoint of P.

Setup

1. Throughout this paper, we consider Markov chains on a finite state space $\mathcal{X}$.
2. We denote by $\mathcal{L} = \mathcal{L}(\mathcal{X})$ to be the set of transition matrices on $\mathcal{X}$.
3. We write $\mathcal{S}(\pi) \subseteq \mathcal{L}$ to be the set of π-stationary transition matrices, that is,
$$\mathcal{S}(\pi) := \{P \in \mathcal{L};\ \pi P = \pi\}.$$
4. For $P \in \mathcal{S}(\pi)$, we write P^* to be the time-reversal or the $\ell^2(\pi)$-adjoint of P.
5. We write $\mathcal{L}(\pi) \subseteq \mathcal{L}$ to be the set of π-reversible transition matrices, that is,
$$\mathcal{L}(\pi) := \{P \in \mathcal{L};\ P^* = P\}.$$

Setup: isometric involution matrices with respect to π

1. Q is said to be an **isometric involution** on $\mathcal{X}$ with respect to π as in Andrieu and Livingstone '21 if and only if Q satisfies $Q^2 = I$ and $Q^* = Q$.

Setup: isometric involution matrices with respect to π

1. Q is said to be an **isometric involution** on $\mathcal{X}$ with respect to π as in Andrieu and Livingstone '21 if and only if Q satisfies $Q^2 = I$ and $Q^* = Q$.
2. Examples of Q: $I, \pm(2\Pi - I)$, where Π is the transition matrix with each row equals to π.

Setup: isometric involution matrices with respect to π

1. Q is said to be an **isometric involution** on $\mathcal{X}$ with respect to π as in Andrieu and Livingstone '21 if and only if Q satisfies $Q^2 = I$ and $Q^* = Q$.
2. Examples of Q: $I, \pm(2\Pi - I)$, where Π is the transition matrix with each row equals to π.
3. We write $\mathcal{I}(\pi)$ to be the set of isometric involution matrices on $\mathcal{X}$ with respect to π.

Setup: isometric involution matrices with respect to π

1. Q is said to be an **isometric involution** on $\mathcal{X}$ with respect to π as in Andrieu and Livingstone '21 if and only if Q satisfies $Q^2 = I$ and $Q^* = Q$.
2. Examples of Q: $I, \pm(2\Pi - I)$, where Π is the transition matrix with each row equals to π.
3. We write $\mathcal{I}(\pi)$ to be the set of isometric involution matrices on $\mathcal{X}$ with respect to π.
4. Let $Q \in \mathcal{I}(\pi)$. $L \in \mathcal{S}(\pi)$ is said to be (π, Q)-self-adjoint if and only if

$$L^* = QLQ.$$

Setup: isometric involution matrices with respect to π

1. Q is said to be an **isometric involution** on $\mathcal{X}$ with respect to π as in Andrieu and Livingstone '21 if and only if Q satisfies $Q^2 = I$ and $Q^* = Q$.
2. Examples of Q: $I, \pm(2\Pi - I)$, where Π is the transition matrix with each row equals to π.
3. We write $\mathcal{I}(\pi)$ to be the set of isometric involution matrices on $\mathcal{X}$ with respect to π.
4. Let $Q \in \mathcal{I}(\pi)$. $L \in \mathcal{S}(\pi)$ is said to be (π, Q)-self-adjoint if and only if

$$L^* = QLQ.$$

5. We write $\mathcal{L}(\pi, Q) \subseteq \mathcal{L}$ to be the set of (π, Q)-self-adjoint transition matrices. In the special case of $Q = I$, we recover that $\mathcal{L}(\pi, I) = \mathcal{L}(\pi)$.

Setup: "equi-probability" permutation matrices with respect to π

1. Let $\mathbf{P}$ be the set of permutations on $\mathcal{X}$. Let $\psi \in \mathbf{P}$ be a permutation, and Q_ψ be the induced permutation matrix with entries $Q_\psi(x, y) := \delta_{y=\psi(x)}$ for all $x, y \in \mathcal{X}$, where δ is the Dirac mass function.

Setup: "equi-probability" permutation matrices with respect to π

1. Let $\mathbf{P}$ be the set of permutations on $\mathcal{X}$. Let $\psi \in \mathbf{P}$ be a permutation, and Q_ψ be the induced permutation matrix with entries $Q_\psi(x, y) := \delta_{y=\psi(x)}$ for all $x, y \in \mathcal{X}$, where δ is the Dirac mass function.
2. Define a set of "equi-probability" permutations with respect to π to be

$$\Psi(\pi) := \{\psi \in \mathbf{P};\ \forall x \in \mathcal{X},\ \psi(\psi(x)) = x,\ \pi(x) = \pi(\psi(x))\}.$$

Setup: "equi-probability" permutation matrices with respect to π

1. Let $\mathbf{P}$ be the set of permutations on $\mathcal{X}$. Let $\psi \in \mathbf{P}$ be a permutation, and Q_ψ be the induced permutation matrix with entries $Q_\psi(x, y) := \delta_{y=\psi(x)}$ for all $x, y \in \mathcal{X}$, where δ is the Dirac mass function.
2. Define a set of "equi-probability" permutations with respect to π to be

$$\Psi(\pi) := \{\psi \in \mathbf{P};\ \forall x \in \mathcal{X},\ \psi(\psi(x)) = x,\ \pi(x) = \pi(\psi(x))\}.$$

3. $\Psi(\pi)$ is non-empty for all π since the identity permutation always belongs to $\Psi(\pi)$.

Linking isometric involution with "equi-probability" permutation

1. An isometric involution matrix need not be a transition matrix, e.g. $\pm(2\Pi - I)$. But what is an isometric involution transition matrix?

Linking isometric involution with "equi-probability" permutation

1. An isometric involution matrix need not be a transition matrix, e.g. $\pm(2\Pi - I)$. But what is an isometric involution transition matrix?

Theorem 1 (C., Hird and Wang '24)

$$\mathcal{I}(\pi) \cap \mathcal{L} = \{Q_\psi;\ \psi \in \Psi(\pi)\}.$$

Linking isometric involution with "equi-probability" permutation

1. An isometric involution matrix need not be a transition matrix, e.g. $\pm(2\Pi - I)$. But what is an isometric involution transition matrix?

Theorem 1 (C., Hird and Wang '24)

$$\mathcal{I}(\pi) \cap \mathcal{L} = \{Q_{\psi};\ \psi \in \Psi(\pi)\}.$$

2. Takeaway message #1: on a finite state space, the set of isometric involution transition matrices with respect to π equals to the set of equi-probability permutation matrices with respect to π.

Remarks

1. Why is the notion of isometric involution important? It turns out that, many state-of-the-art non-reversible MCMC algorithms, while being non-reversible with respect to π, are in fact (π, Q)-self-adjoint. This motivates the study in Andrieu and Livingstone (AoS, '21).

Remarks

1. Why is the notion of isometric involution important? It turns out that, many state-of-the-art non-reversible MCMC algorithms, while being non-reversible with respect to π, are in fact (π, Q)-self-adjoint. This motivates the study in Andrieu and Livingstone (AoS, '21).
2. We shall reveal more connections between equi-probability permutations and equi-energy samplers by Kou et al. (AoS, '06).

The KL divergence between Markov chains

1. Let us first recall the KL divergence between transition matrices of Markov chains.

The KL divergence between Markov chains

1. Let us first recall the KL divergence between transition matrices of Markov chains.

Definition 2 (KL divergence between Markov chains)

For given π and transition matrices $M, L \in \mathcal{L}$, we define the KL-divergence from L to M with respect to π as

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

where several standard conventions apply.

The KL divergence between Markov chains

1. Let us first recall the KL divergence between transition matrices of Markov chains.

Definition 2 (KL divergence between Markov chains)

For given π and transition matrices $M, L \in \mathcal{L}$, we define the KL-divergence from L to M with respect to π as

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

where several standard conventions apply.

2. Note that π need not be the stationary distribution of L or M. In particular, when M is assumed to be π-stationary, $D^{\pi}_{KL}(M \| L)$ can be interpreted as the KL divergence rate from L to M.

Projection onto $\mathcal{L}(\pi, Q)$

1. At MCQMC 2024, Max Hird raised to me the following question: given $P \in \mathcal{S}(\pi)$ and Q be an isometric involution transition matrix with respect to π, what is a projection of P onto the set of (π, Q)-self-adjoint transition matrices $\mathcal{L}(\pi, Q)$ under D^{π}_{KL}?

Projection onto $\mathcal{L}(\pi, Q)$

1. At MCQMC 2024, Max Hird raised to me the following question: given $P \in \mathcal{S}(\pi)$ and Q be an isometric involution transition matrix with respect to π, what is a projection of P onto the set of (π, Q)-self-adjoint transition matrices $\mathcal{L}(\pi, Q)$ under D^{π}_{KL}?
2. That is, what is

$$\underset{M \in \mathcal{L}(\pi, Q)}{\arg\min}\ D^{\pi}_{KL}(P \| M)?$$

Projection onto $\mathcal{L}(\pi, Q)$

1. At MCQMC 2024, Max Hird raised to me the following question: given $P \in \mathcal{S}(\pi)$ and Q be an isometric involution transition matrix with respect to π, what is a projection of P onto the set of (π, Q)-self-adjoint transition matrices $\mathcal{L}(\pi, Q)$ under D^{π}_{KL}?
2. That is, what is
$$\operatorname*{arg\,min}_{M \in \mathcal{L}(\pi, Q)} D^{\pi}_{KL}(P \| M)?$$
3. In the above setting, we define
$$\overline{P}(Q) := \frac{1}{2}(P + QP^*Q).$$

This is known as a mixture of permuted Markov chains in Dubail '24.

Pythagorean identity

Theorem 3 (C., Hird and Wang '24)

Let $P \in \mathcal{S}(\pi)$ *and* $Q \in \mathcal{I}(\pi) \cap \mathcal{L}$. *For any* $M \in \mathcal{L}(\pi, Q)$, *we have*

$$D_{KL}^{\pi}(P \| M) = D_{KL}^{\pi}(P \| \overline{P}(Q)) + D_{KL}^{\pi}(\overline{P}(Q) \| M).$$

Pythagorean identity

Theorem 3 (C., Hird and Wang '24)

Let $P \in \mathcal{S}(\pi)$ and $Q \in \mathcal{I}(\pi) \cap \mathcal{L}$. For any $M \in \mathcal{L}(\pi, Q)$, we have

$$D^{\pi}_{KL}(P \| M) = D^{\pi}_{KL}(P \| \overline{P}(Q)) + D^{\pi}_{KL}(\overline{P}(Q) \| M).$$

1. The above result implies that

$$\overline{P}(Q) = \operatorname*{arg\,min}_{M \in \mathcal{L}(\pi, Q)} D^{\pi}_{KL}(P \| M).$$

Pythagorean identity

Theorem 3 (C., Hird and Wang '24)

Let $P \in \mathcal{S}(\pi)$ and $Q \in \mathcal{I}(\pi) \cap \mathcal{L}$. For any $M \in \mathcal{L}(\pi, Q)$, we have

$$D_{KL}^{\pi}(P \| M) = D_{KL}^{\pi}(P \| \overline{P}(Q)) + D_{KL}^{\pi}(\overline{P}(Q) \| M).$$

1. The above result implies that

$$\overline{P}(Q) = \underset{M \in \mathcal{L}(\pi, Q)}{\arg\min}\ D_{KL}^{\pi}(P \| M).$$

2. Takeaway message #2: $\overline{P}(Q)$ is the unique information projection of P onto the set of (π, Q)-self-adjoint transition matrices.

Remarks

1. Pythagorean identities are cornerstone results in the field of information geometry, see e.g. the work of Amari or Wolfer and Watanabe '21 for results on information geometry of transition matrices.

Remarks

1. Pythagorean identities are cornerstone results in the field of information geometry, see e.g. the work of Amari or Wolfer and Watanabe '21 for results on information geometry of transition matrices.
2. In the paper, we also establish the Pythagorean identity under the squared-Frobenius norm, that is, we have results of the form

$$\|P - M\|_F^2 = \left\|P - \overline{P}(Q)\right\|_F^2 + \left\|\overline{P}(Q) - M\right\|_F^2,$$

where $\|\cdot\|_F$ is the matrix Frobenius norm.

Remarks

1. Pythagorean identities are cornerstone results in the field of information geometry, see e.g. the work of Amari or Wolfer and Watanabe '21 for results on information geometry of transition matrices.
2. In the paper, we also establish the Pythagorean identity under the squared-Frobenius norm, that is, we have results of the form

$$\|P - M\|_F^2 = \left\|P - \overline{P}(Q)\right\|_F^2 + \left\|\overline{P}(Q) - M\right\|_F^2,$$

   where $\|\cdot\|_F$ is the matrix Frobenius norm.
3. More generally, I think the result can be generalized to some Bregman divergences between matrices (not in the paper).

Comparison theory

1. In the context of MCMC, Markov chain samplers such as Metropolis-Hastings or Gibbs samplers are frequently used to sample approximately from π.

Comparison theory

1. In the context of MCMC, Markov chain samplers such as Metropolis-Hastings or Gibbs samplers are frequently used to sample approximately from π.
2. Given ergodic $P \in \mathcal{S}(\pi)$ and an isometric involution transition matrix Q, there are now quite a few transition matrices associated with these objects:
   - P
   - QP
   - PQ
   - QPQ
   - $\overline{P}(Q) = \frac{1}{2}(P + QP^*Q)$
   - the mixture $\alpha P + (1-\alpha)QP^*Q$, where $\alpha \in [0,1]$.

   Which one is the "best"? E.g. which transition matrix converges to π the fastest or performs the "best" under some suitable metrics?

Comparison theory

Theorem 4 (C., Hird and Wang '24)

Under most commonly used metrics and suitable assumptions (e.g. $P \in \mathcal{L}(\pi)$),

$$\overline{P}(Q)$$

is preferred over the others.

Comparison theory

Theorem 4 (C., Hird and Wang '24)

Under most commonly used metrics and suitable assumptions (e.g. $P \in \mathcal{L}(\pi)$),

$$\overline{P}(Q)$$

is preferred over the others.

1. The convergence metrics studied in the paper include
   - SLEM: Second Largest Eigenvalue in Modulus
   - right spectral gap
   - Dobrushin coefficient
   - Kemeny's constant
   - worst-case asymptotic variance
   - average asymptotic variance

Comparison theory

Theorem 4 (C., Hird and Wang '24)

Under most commonly used metrics and suitable assumptions (e.g. $P \in \mathcal{L}(\pi)$),

$$\overline{P}(Q)$$

is preferred over the others.

1. The convergence metrics studied in the paper include
   - SLEM: Second Largest Eigenvalue in Modulus
   - right spectral gap
   - Dobrushin coefficient
   - Kemeny's constant
   - worst-case asymptotic variance
   - average asymptotic variance
2. Takeaway message #3: $\overline{P}(Q)$ is a preferred sampler of π compared with P or other competing transition matrices, based on most commonly used metrics.

Alternating projections

1. Let $m \in \mathbb{N}$ and suppose that we have a sequence of isometric involution transition matrices $Q_i \in \mathcal{I}(\pi) \cap \mathcal{L}$ for $i \in [\![0, m-1]\!]$. Is there a way to combine these Q_i to further improve the convergence to equilibrium?

Alternating projections

1. Let $m \in \mathbb{N}$ and suppose that we have a sequence of isometric involution transition matrices $Q_i \in \mathcal{I}(\pi) \cap \mathcal{L}$ for $i \in [\![0, m-1]\!]$. Is there a way to combine these Q_i to further improve the convergence to equilibrium?
2. One natural idea in this context is alternating projections. Specifically, given a $P \in \mathcal{L}(\pi)$, we first project it onto the space $\mathcal{L}(\pi, Q_0)$ to obtain $R_1 = R_1(Q_0, \ldots, Q_{m-1}, P) := \overline{P}(Q_0)$.

Alternating projections

1. Let $m \in \mathbb{N}$ and suppose that we have a sequence of isometric involution transition matrices $Q_i \in \mathcal{I}(\pi) \cap \mathcal{L}$ for $i \in [\![0, m-1]\!]$. Is there a way to combine these Q_i to further improve the convergence to equilibrium?
2. One natural idea in this context is alternating projections. Specifically, given a $P \in \mathcal{L}(\pi)$, we first project it onto the space $\mathcal{L}(\pi, Q_0)$ to obtain $R_1 = R_1(Q_0, \ldots, Q_{m-1}, P) := \overline{P}(Q_0)$.
3. Second, we project R_1 onto the space $\mathcal{L}(\pi, Q_1)$ to obtain $R_2 = R_2(Q_0, \ldots, Q_{m-1}, P) := \overline{R_1}(Q_1)$.

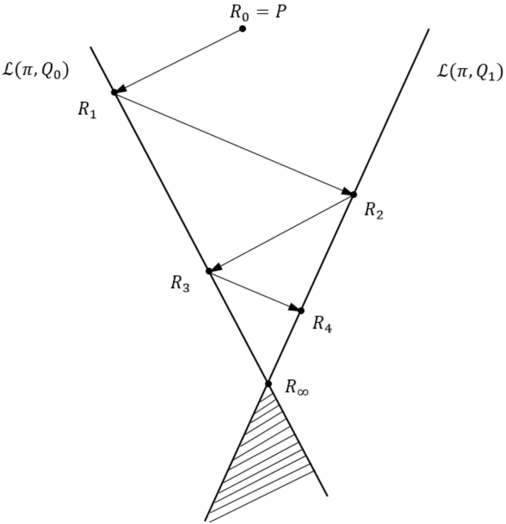
Alternating projections

1. Let $m \in \mathbb{N}$ and suppose that we have a sequence of isometric involution transition matrices $Q_i \in \mathcal{I}(\pi) \cap \mathcal{L}$ for $i \in [\![0, m-1]\!]$. Is there a way to combine these Q_i to further improve the convergence to equilibrium?
2. One natural idea in this context is alternating projections. Specifically, given a $P \in \mathcal{L}(\pi)$, we first project it onto the space $\mathcal{L}(\pi, Q_0)$ to obtain $R_1 = R_1(Q_0, \ldots, Q_{m-1}, P) := \overline{P}(Q_0)$.
3. Second, we project R_1 onto the space $\mathcal{L}(\pi, Q_1)$ to obtain $R_2 = R_2(Q_0, \ldots, Q_{m-1}, P) := \overline{R_1}(Q_1)$.
4. Third, we project R_2 onto the space $\mathcal{L}(\pi, Q_2)$ to obtain $R_3 = R_3(Q_0, \ldots, Q_{m-1}, P) := \overline{R_2}(Q_2)$. We proceed iteratively and the projection order is deterministic in a cycle in the order of $Q_0, \ldots, Q_{m-1}$. Precisely, for $n \in \mathbb{N}$, we define

$$R_n = R_n(Q_0, \ldots, Q_{m-1}, P) := \overline{R_{n-1}}(Q_{(n-1) \mod m})$$

with the initial condition $R_0 := P$.

Visualization of alternating projections with $m = 1$



Alternating projections

1. Main benefit of alternating projections: at each step, the projected transition matrix is improved when compared with the previous unprojected transition matrix, in view of the results in the previous section. In other words, we make use of the geometric structure to yield better Markov chain sampler.

Alternating projections

1. Main benefit of alternating projections: at each step, the projected transition matrix is improved when compared with the previous unprojected transition matrix, in view of the results in the previous section. In other words, we make use of the geometric structure to yield better Markov chain sampler.
2. Denote the intersections of $\mathcal{L}(\pi, Q_0), \ldots, \mathcal{L}(\pi, Q_{m-1})$ to be

$$\mathcal{E} = \mathcal{E}(\pi, Q_0, \ldots, Q_{m-1}) := \bigcap_{k=0}^{m-1} \mathcal{L}(\pi, Q_k).$$

Let R_∞ be an information projection of $P \in \mathcal{L}(\pi)$ onto $\mathcal{E}$, that is,

$$R_\infty = R_\infty(Q_0, \ldots, Q_{m-1}, P) := \operatorname*{arg\,min}_{N \in \mathcal{E}} D_{KL}^{\pi}(P \| N).$$

It can be shown that R_∞ is unique.

Alternating projections

1. Main benefit of alternating projections: at each step, the projected transition matrix is improved when compared with the previous unprojected transition matrix, in view of the results in the previous section. In other words, we make use of the geometric structure to yield better Markov chain sampler.
2. Denote the intersections of $\mathcal{L}(\pi, Q_0), \ldots, \mathcal{L}(\pi, Q_{m-1})$ to be

   $$\mathcal{E} = \mathcal{E}(\pi, Q_0, \ldots, Q_{m-1}) := \bigcap_{k=0}^{m-1} \mathcal{L}(\pi, Q_k).$$

   Let R_∞ be an information projection of $P \in \mathcal{L}(\pi)$ onto $\mathcal{E}$, that is,

   $$R_\infty = R_\infty(Q_0, \ldots, Q_{m-1}, P) := \operatorname*{arg\,min}_{N \in \mathcal{E}} D_{KL}^{\pi}(P \| N).$$

   It can be shown that R_∞ is unique.
3. Using the theory of alternating projections, we prove that R_n converges to the limit R_∞ elementise.

Alternating projections

Theorem 5 (C., Hird and Wang '24)

Let $m, n \in \mathbb{N}$. Let $P \in \mathcal{L}(\pi)$ and $Q_i \in \mathcal{I}(\pi) \cap \mathcal{L}$ for $i \in [\![0, m-1]\!]$ be a sequence of isometric involution transition matrices. Define R_n as earlier. The following limit exists (pointwise or in total variation):

$$\lim_{n \to \infty} R_n = R_\infty,$$

and

$$R_\infty \in \mathcal{L}(\pi) \cap \mathcal{E}, \quad \mathrm{Tr}(R_\infty) = \mathrm{Tr}(P).$$

Alternating projections

Theorem 5 (C., Hird and Wang '24)

Let $m, n \in \mathbb{N}$. Let $P \in \mathcal{L}(\pi)$ and $Q_i \in \mathcal{I}(\pi) \cap \mathcal{L}$ for $i \in [\![0, m-1]\!]$ be a sequence of isometric involution transition matrices. Define R_n as earlier. The following limit exists (pointwise or in total variation):

$$\lim_{n\to\infty} R_n = R_\infty,$$

and

$$R_\infty \in \mathcal{L}(\pi) \cap \mathcal{E}, \quad \mathrm{Tr}(R_\infty) = \mathrm{Tr}(P).$$

1. Takeaway message #4: consider using the alternating projections R_n or even R_∞ if possible, since these cannot be worse off than the original P.

Remarks

1. Alternating projections of Gibbs samplers have been studied in Diaconis et al. '10 and more recently in a paper by Qian Qin (AoAP, '24).

Remarks

1. Alternating projections of Gibbs samplers have been studied in Diaconis et al. '10 and more recently in a paper by Qian Qin (AoAP, '24).
2. Assume that we can simulate P, can we simulate the trajectories of the Markov chain with transition matrix R_n? In the paper, we have devised a recursive simulation procedure to do so.

Remarks

1. Alternating projections of Gibbs samplers have been studied in Diaconis et al. '10 and more recently in a paper by Qian Qin (AoAP, '24).
2. Assume that we can simulate P, can we simulate the trajectories of the Markov chain with transition matrix R_n? In the paper, we have devised a recursive simulation procedure to do so.
3. What is the rate of convergence of R_n towards R_∞? It turns out to depend on the angle between suitable subspaces, see the paper.

The maximum speed limit in (School Zone of) Singapore

Image Source: Land Transport Authority, Singapore
Image Link: https://www.lta.gov.sg/content/ltagov/en/getting_around/driving_in_singapore/driving_rules_and_regulations.html

The maximum speed limit of projection sampler

1. What are the situations that we have $R_n = \Pi$? This is an ideal situation since we have an exact sampler by simulating directly R_n! This is unlike most MCMC samplers which are approximate samplers of π.

A necessary condition of $R_n = \Pi$ in terms of trace

Corollary 6 (C., Hird and Wang '24)

If $R_n = \Pi$ for some $n \in \mathbb{N} \cup \{\infty\}$, then

$$\mathrm{Tr}(P) = 1.$$

This implies that, for π-reversible positive-definite transition matrices, $R_n \neq \Pi$ since $\mathrm{Tr}(P) > 1$.

A necessary condition of $\overline{P}(Q) = \Pi$ via the Sylvester's equation

Let $\lambda(P)$ be the set of eigenvalues of the matrix P. A direct application of the Sylvester's equation yields the following result:

Corollary 7 (C., Hird and Wang '24)

Let $P \in \mathcal{S}(\pi)$ and $Q \in \mathcal{I}(\pi) \cap \mathcal{L}$ be an isometric involution transition matrix. If $\overline{P}(Q) = \Pi$, then

$$\lambda(P) \cap \lambda(-P^*) \neq \emptyset,$$

that is, P and $-P^*$ have at least one common eigenvalue.

$\mathrm{Tr}(P) = 1$ is necessary and sufficient of $R_\infty = \Pi$ under some additional assumptions

Let π be the discrete uniform distribution, and P be a symmetric doubly stochastic matrix.

Theorem 8 (C., Hird and Wang '24)

Under the above assumptions and suitable choices of the permutation matrices, $R_\infty = \Pi$ if and only if $\mathrm{Tr}(P) = 1$.

Takeaway message #5: consider transforming the transition matrix so that it satisfies $\mathrm{Tr}(P) = 1$. Depending on P, even if it may not lead to $R_n = \Pi$, but at least it seems to make the Markov chain "better-conditioned".

Tuning strategies: overview

1. The isometric involution transition matrix $Q \in \mathcal{I}(\pi) \cap \mathcal{L}$ can be understood as a parameter in these algorithms, and the improvement depends on the tuning of Q.

Tuning strategies: overview

1. The isometric involution transition matrix $Q \in \mathcal{I}(\pi) \cap \mathcal{L}$ can be understood as a parameter in these algorithms, and the improvement depends on the tuning of Q.
2. For instance, the choice of $Q = I$ is always feasible, yet it leads to no improvement when $P \in \mathcal{L}(\pi)$ since $\overline{P}(I) = P$.

Tuning strategies: overview

1. The isometric involution transition matrix $Q \in \mathcal{I}(\pi) \cap \mathcal{L}$ can be understood as a parameter in these algorithms, and the improvement depends on the tuning of Q.
2. For instance, the choice of $Q = I$ is always feasible, yet it leads to no improvement when $P \in \mathcal{L}(\pi)$ since $\overline{P}(I) = P$.
3. On the other hand, we have seen in previous section that depending on P it might be possible to achieve $R_n = \Pi$ or $R_\infty = \Pi$ with suitable choices of Qs.

Tuning Q via optimization and Markov chain assignment problems

1. The first strategy seeks to find an optimal Q that minimizes the discrepancy between $\overline{P}(Q)$ and Π or more generally between R_n and Π.

Tuning Q via optimization and Markov chain assignment problems

1. The first strategy seeks to find an optimal Q that minimizes the discrepancy between $\overline{P}(Q)$ and Π or more generally between R_n and Π.
2. Precisely, we would like to find Q that minimizes the π-weighted KL divergence for a given $P \in \mathcal{S}(\pi)$:

$$Q_{*,KL} = Q_{*,KL}(P) := \underset{Q \in \mathcal{I}(\pi) \cap \mathcal{L}}{\arg\min}\ D^{\pi}_{KL}(\overline{P}(Q) \| \Pi).$$

Tuning Q via optimization and Markov chain assignment problems

1. The first strategy seeks to find an optimal Q that minimizes the discrepancy between $\overline{P}(Q)$ and Π or more generally between R_n and Π.
2. Precisely, we would like to find Q that minimizes the π-weighted KL divergence for a given $P \in \mathcal{S}(\pi)$:

$$Q_{*,KL} = Q_{*,KL}(P) := \underset{Q \in \mathcal{I}(\pi) \cap \mathcal{L}}{\arg\min} \; D^{\pi}_{KL}(\overline{P}(Q) \| \Pi).$$

3. The above optimization problems may not be solved in realistic time frame in practice, since π may involve normalization constant that is non-tractable.

Tuning Q via optimization and Markov chain assignment problems

1. The first strategy seeks to find an optimal Q that minimizes the discrepancy between $\overline{P}(Q)$ and Π or more generally between R_n and Π.
2. Precisely, we would like to find Q that minimizes the π-weighted KL divergence for a given $P \in \mathcal{S}(\pi)$:

$$Q_{*,KL} = Q_{*,KL}(P) := \underset{Q \in \mathcal{I}(\pi) \cap \mathcal{L}}{\arg\min} \; D^{\pi}_{KL}(\overline{P}(Q) \| \Pi).$$

3. The above optimization problems may not be solved in realistic time frame in practice, since π may involve normalization constant that is non-tractable.
4. Luckily, the Pythagorean identity comes to the rescue!

Tuning Q via optimization and Markov chain assignment problems

1. The Pythagorean identity gives that

$$Q_{*,KL} = \underset{Q \in \mathcal{I}(\pi) \cap \mathcal{L}}{\arg\max}\, D^{\pi}_{KL}(P \| \overline{P}(Q)) = \underset{\psi \in \Psi(\pi)}{\arg\max}\, D^{\pi}_{KL}(P \| \overline{P}(Q_{\psi})).$$

Tuning Q via optimization and Markov chain assignment problems

1. The Pythagorean identity gives that

$$Q_{*,KL} = \underset{Q \in \mathcal{I}(\pi) \cap \mathcal{L}}{\arg\max}\, D^{\pi}_{KL}(P \| \overline{P}(Q)) = \underset{\psi \in \Psi(\pi)}{\arg\max}\, D^{\pi}_{KL}(P \| \overline{P}(Q_{\psi})).$$

2. The rightmost maximization problem can be understood as a Markov chain assignment problem constrained to choosing equi-probability permutations within the set $\Psi(\pi)$. While in general assignment problems can be solved in polynomial time in $|\mathcal{X}|$, this may still be computationally infeasible in practice since $|\mathcal{X}|$ might be exponentially large in many models of interest in the context of MCMC, e.g. Ising or Potts model.

Tuning Q via optimization and Markov chain assignment problems

1. The earlier considerations can be generalized to consider multidimensional Markov chain assignment problems. Specifically, we seek to solve, for $m, l \in \mathbb{N}$,

$$\underset{\psi_i \in \Psi(\pi), \forall i \in [\![0, m-1]\!]}{\arg\min} D_{KL}^{\pi}(R_l(Q_{\psi_0}, \ldots, Q_{\psi_{m-1}}, P) \| \Pi)$$

$$= \underset{\psi_i \in \Psi(\pi), \forall i \in [\![0, m-1]\!]}{\arg\max} \sum_{j=0}^{l-1} D_{KL}^{\pi}(R_j(Q_{\psi_0}, \ldots, Q_{\psi_{m-1}}, P) \| R_{j+1}(Q_{\psi_0}, \ldots, Q_{\psi_{m-}}$$

Tuning Q via optimization and Markov chain assignment problems

1. The earlier considerations can be generalized to consider multidimensional Markov chain assignment problems. Specifically, we seek to solve, for $m, l \in \mathbb{N}$,

$$\underset{\psi_i \in \Psi(\pi),\, \forall i \in [\![0, m-1]\!]}{\arg\min} D_{KL}^{\pi}(R_l(Q_{\psi_0}, \ldots, Q_{\psi_{m-1}}, P) \| \Pi)$$

$$= \underset{\psi_i \in \Psi(\pi),\, \forall i \in [\![0, m-1]\!]}{\arg\max} \sum_{j=0}^{l-1} D_{KL}^{\pi}(R_j(Q_{\psi_0}, \ldots, Q_{\psi_{m-1}}, P) \| R_{j+1}(Q_{\psi_0}, \ldots, Q_{\psi_{m-}}$$

2. Multidimensional assignment problems are in general NP-hard to solve, but there are useful heuristics in practice, see the paper for more discussions.

Tuning Q adaptively in a single run

1. Let $H : \mathcal{X} \to \mathbb{R}$ be a target energy function, and π_β be its associated Gibbs distribution at inverse temperature $\beta \geqslant 0$, that is, for $x \in \mathcal{X}$,

$$\pi_\beta(x) := \frac{e^{-\beta H(x)}}{Z_\beta},$$

where $Z_\beta := \sum_{x \in \mathcal{X}} e^{-\beta H(x)}$ is the normalization constant. Thus, we see that $\pi_\beta(x) = \pi_\beta(y)$ if and only if $H(x) = H(y)$, that is, equi-probability is the same as equi-energy.

Tuning Q adaptively in a single run

1. Let $H : \mathcal{X} \to \mathbb{R}$ be a target energy function, and π_β be its associated Gibbs distribution at inverse temperature $\beta \geqslant 0$, that is, for $x \in \mathcal{X}$,

$$\pi_\beta(x) := \frac{e^{-\beta H(x)}}{Z_\beta},$$

where $Z_\beta := \sum_{x \in \mathcal{X}} e^{-\beta H(x)}$ is the normalization constant. Thus, we see that $\pi_\beta(x) = \pi_\beta(y)$ if and only if $H(x) = H(y)$, that is, equi-probability is the same as equi-energy.

2. The second tuning strategy lies in adjusting Q adaptively on the fly as the algorithm progresses. In short, the algorithm learns the equi-probability or equi-energy permutation adaptively on the fly.

Tuning Q adaptively in multiple runs

1. We use an exploration Markov chain, such as the proposal chain in Metropolis-Hastings or the Metropolis-Hastings chain at high temperature, for $k \in \mathbb{N}$ times. Each time an equi-energy permutation matrix $(Q_n)_{n=0}^k$ is generated or learnt on the fly.

Tuning Q adaptively in multiple runs

1. We use an exploration Markov chain, such as the proposal chain in Metropolis-Hastings or the Metropolis-Hastings chain at high temperature, for $k \in \mathbb{N}$ times. Each time an equi-energy permutation matrix $(Q_n)_{n=0}^k$ is generated or learnt on the fly.
2. We then combine these permutation matrices using alternating projections.

Tuning Q adaptively in multiple runs

1. We use an exploration Markov chain, such as the proposal chain in Metropolis-Hastings or the Metropolis-Hastings chain at high temperature, for $k \in \mathbb{N}$ times. Each time an equi-energy permutation matrix $(Q_n)_{n=0}^k$ is generated or learnt on the fly.
2. We then combine these permutation matrices using alternating projections.
3. This idea is inspired by the equi-energy sampler of Kou et al. (AoS, '06).

Tuning Q adaptively in multiple runs

1. Animation time! Ising model on the line. We shall see that the adaptive projection sampler is able to hop between the two modes (all-black and all-white), and the standard Metropolis-Hastings struggles to transverse between the two modes.

Tuning Q adaptively in multiple runs

1. Animation time! Ising model on the line. We shall see that the adaptive projection sampler is able to hop between the two modes (all-black and all-white), and the standard Metropolis-Hastings struggles to transverse between the two modes.
2. Animation credit and computational assistance: Zheyuan Lai (NUS)

Recap and outlook

1. This paper presents a new technique of improving the convergence of Markov chains via permutations and projections.

Recap and outlook

1. This paper presents a new technique of improving the convergence of Markov chains via permutations and projections.
2. In a broad sense, this technique can be understood as a preconditioning technique to improve the mixing of Markov chain.

Recap and outlook

1. This paper presents a new technique of improving the convergence of Markov chains via permutations and projections.
2. In a broad sense, this technique can be understood as a preconditioning technique to improve the mixing of Markov chain.
3. I find it intriguing since we are using various mathematical ideas “for good” in the context of MCMC: ideas from information projections, alternating projections, Sylvester’s equations, assignment problems all naturally come together to help us design improved MCMC samplers.

Recap and outlook

1. This paper presents a new technique of improving the convergence of Markov chains via permutations and projections.
2. In a broad sense, this technique can be understood as a preconditioning technique to improve the mixing of Markov chain.
3. I find it intriguing since we are using various mathematical ideas "for good" in the context of MCMC: ideas from information projections, alternating projections, Sylvester's equations, assignment problems all naturally come together to help us design improved MCMC samplers.
4. Some ideas for future work: extending the methodology to more general state space and more general processes such as diffusion processes? Can we design better samplers for problems in theoretical computer science where sampling is crucial? Implications for Sequential Monte Carlo? You are more than welcome to join me if interested!

Thank you! Question(s)?