THE FAST AND THE CURIOUS: MODERN MARKOV CHAIN MONTE CARLO

IRSA's 2023 Conference
May 18 to 20

Informed MCMC Sampling: Theory and Algorithms

Presenter: Quan Zhou
Department of Statistics, Texas A&M University

INSTITUTE for RESEARCH IN STATISTICS AND ITS APPLICATIONS

COLLEGE OF LIBERAL ARTS

Acknowledgment

This talk is based on my recent works co-authored with

- Guanxun Li (PhD student/postdoc), Texas A&M University
- Hyunwoong Chang (PhD student), Texas A&M University
- Aaron Smith, University of Ottawa
- Jun Yang, Oxford University/University of Copenhagen
- Dootika Vats, Indian Institute of Technology Kanpur
- Gareth Roberts, University of Warwick
- Jeffrey Rosenthal, University of Toronto

The research presented in this talk is supported by NSF DMS-2245591, DMS-2311307.

Outline of the talk

- Introduction
- General theoretical guarantees
  - A unimodal condition
  - Rapid mixing of random walk MH
  - Even faster mixing of informed MH
- Examples
  - Variable selection
  - Graphical structure learning
- A better solution than MH: importance tempering
  - Informed importance tempering
  - More sophisticated algorithms (e.g. multiple-try)
  - Comment on Peskun's ordering

High-dimensional model selection

Let $\mathcal{X}$ be the state space for a model selection problem with p variables. Let $|\mathcal{X}|$ denote the cardinality.

Examples:

- Variable selection (sparse linear regression): $\mathcal{X} = 2^{\{1,\dots,p\}}$, $|\mathcal{X}| = 2^p$.
- Structure learning: $\mathcal{X}$ is the space of p-vertex DAGs, and $|\mathcal{X}|$ grows super-exponentially with p.

In high-dimensional settings, sparsity constraints need to be imposed, but usually $|\mathcal{X}|$ still grows *super-polynomially* in p.

Goal: generate samples from the posterior distribution π.

MCMC for model selection

Local MCMC samplers

Most MCMC samplers for model selection problems are "local": at x, we propose the next state from a "small" set $\mathcal{N}(x) \subset \mathcal{X}$ such that $|\mathcal{N}(x)|$ is *polynomial* in p.

Example: a typical path in variable selection.

$$\{1,2\} \xrightarrow{\text{add covariate 4}} \{1,2,4\} \xrightarrow{\text{swap covariate 2 with 3}} \{1,3,4\}$$

$$\xrightarrow{\text{delete covariate 4}} \{1,3\} \xrightarrow{\text{delete covariate 1}} \{3\}$$

Do we have theoretical guarantees?

Rapid mixing

An MCMC algorithm is rapidly mixing if its mixing time grows polynomially with p (number of variables) and n (sample size).

Challenges in the complexity analysis of MCMC for model selection:

1. State space is discrete, and samplers only use local moves.
2. Need mixing time bounds in high-dimensional settings.

Two steps:

1. Analyze the landscape of π using high-dim statistical theory.
2. Bound the mixing time using tools from Markov chain theory.

Two classes of proposals

Consider local Metropolis-Hastings (MH) algorithms. Let $K(x, \cdot)$ denote the proposal distribution at state x.

Random walk (uninformed) proposal

Recall $\mathcal{N}(x)$ is the set of neighboring states of x. Let $K(x, y) = 1/|\mathcal{N}(x)|$ for every $y \in \mathcal{N}(x)$. That is, we randomly propose a state from $\mathcal{N}(x)$ with equal probability.

Informed proposal

Let $K(x, y)$ depend on $\pi(y)$. For example, set $K(x, y) \propto \pi(y)$ for $y \in \mathcal{N}(x)$.

Do we have theoretical guarantees?

On the informed proposals:

- Similar ideas are used in many MCMC methods [Titsias and Yau, 2017, Zanella and Roberts, 2019, Zanella, 2020, Griffin et al., 2021] and some non-MCMC methods [Hans et al., 2007, Shin et al., 2018].
- To implement an informed proposal at x, we need to evaluate π for each $y \in \mathcal{N}(x)$.
- Can informed MCMC methods achieve a sufficiently fast convergence rate that offsets the cost of local evaluation of π?
- How to choose the proposal weighting scheme?
- Any other scheme that is more efficient than the MH implementation?

Our mixing time bounds

Define mixing time by $T_{\text{mix}} = \max_x \min\{t : ||P^t(x, \cdot) - \pi||_{\text{TV}} \leq 1/4\}$.

Under a "unimodal condition" on π,

- For random walk MH, the mixing time is $O(N \log \pi_{\min}^{-1})$ where
  - $\pi_{\min} = \min_{x \in \mathcal{X}} \pi(x)$,
  - N is the maximum neighborhood size.
- There exists an informed MH with mixing time $O(\log \pi_{\min}^{-1})$.

(Precise statements will be given later.)

Recall that the per-iteration cost of informed MH is N times that of random walk MH.

Set-up for general finite state spaces

Notation:

- $\mathcal{X}$: a finite state space.
- π: a positive probability measure.
- $\mathcal{N}$: a neighborhood function; i.e., $\mathcal{N}(x)$ is the set of states that the sampler may move to from x. Assume (i) $x \notin \mathcal{N}(x)$, and (ii) $x \in \mathcal{N}(y)$ whenever $y \in \mathcal{N}(x)$.
- $x^* = \arg\max_{x \in \mathcal{X}} \pi(x)$: the global mode.

To obtain a rapid mixing guarantee, we will assume π is "unimodal with light tails."

A unimodal condition

Define two parameters:

$$N = \max_{x \in \mathcal{X}} |\mathcal{N}(x)|,$$

$$R = \min_{x \neq x^*} \max_{y \in \mathcal{N}(x)} \frac{\pi(y)}{\pi(x)}.$$

That is, for any $x \neq x^*$, there exists $y \in \mathcal{N}(x)$ such that $\pi(y)/\pi(x) \geq R$.

- If $R > 1$, we say π is unimodal (w.r.t. $\mathcal{N}$).
- If $R > N$, π has "sub-exponential tails":

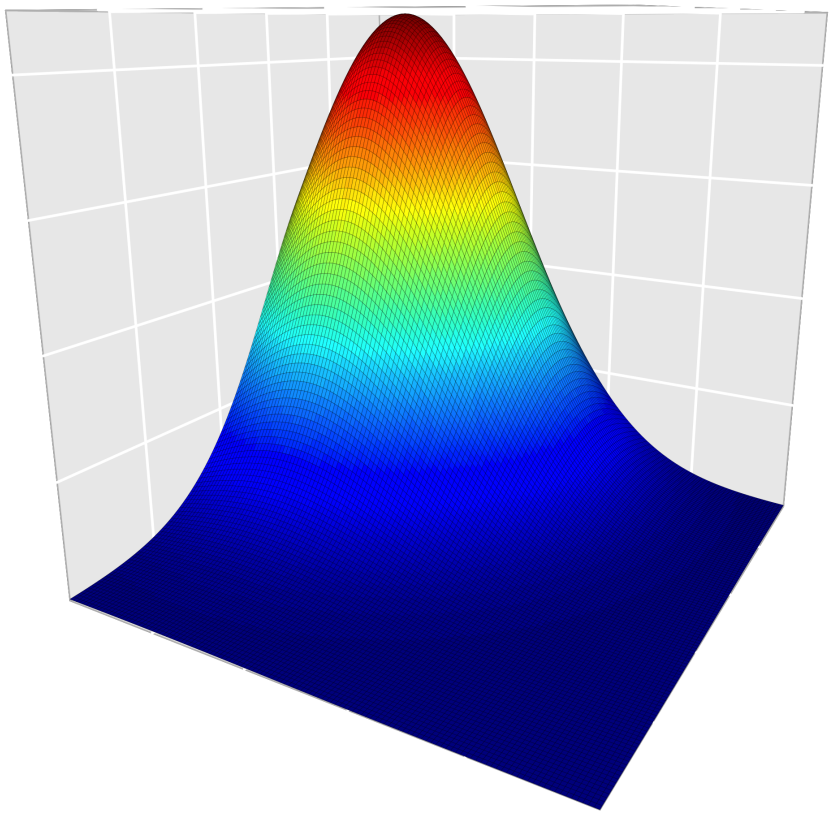
$$\pi(\{x : \mathrm{dist}(x, x^*) \geq k\}) \leq e^{-ck}, \qquad \text{where } c = \log(R/N).$$

We will assume $R > N \geq 1$.

On the unimodal condition

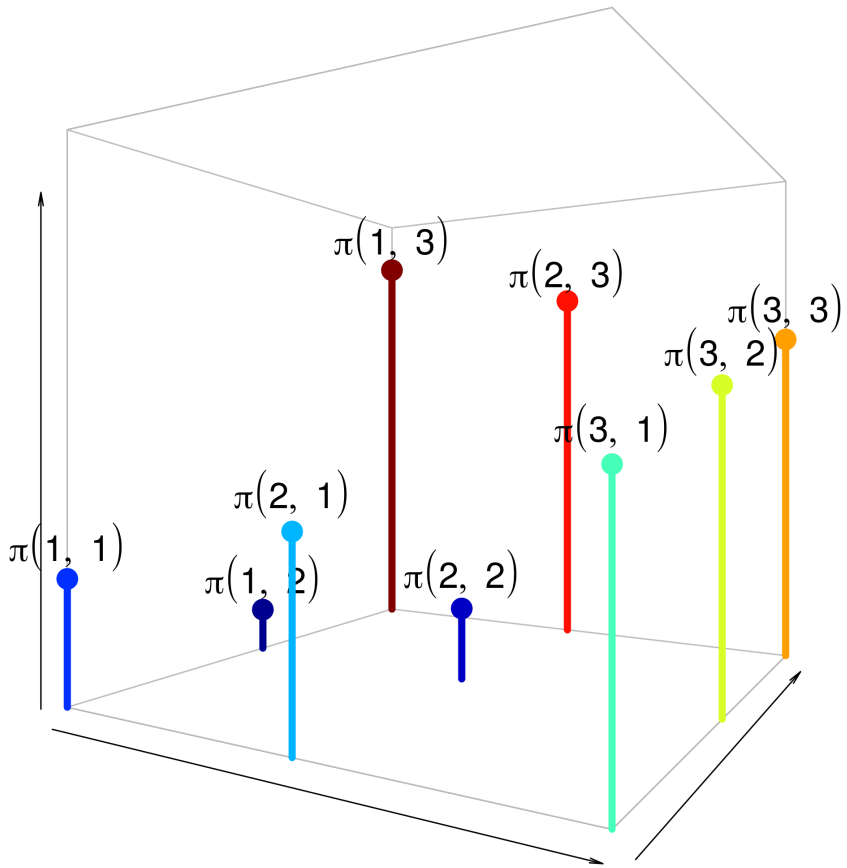
- It is supported by the high-dimensional statistical theory.
- Log-concavity is widely used in the theory of sampling and optimization algorithms on $\mathbb{R}^p$ [Dalalyan, 2017, Dwivedi et al., 2018, Cheng et al., 2018, Mangoubi and Smith, 2019, Shen and Lee, 2019]. Log-concavity implies unimodality and sub-exponential tails, and is conceptually stronger than our unimodal condition.
- If there exist very "sharp" sub-optimal local modes separated from each other, rapid mixing may be impossible.
- Unimodal distribution is a building block for general multimodal distributions (use state decomposition theorems).

Unimodal distributions



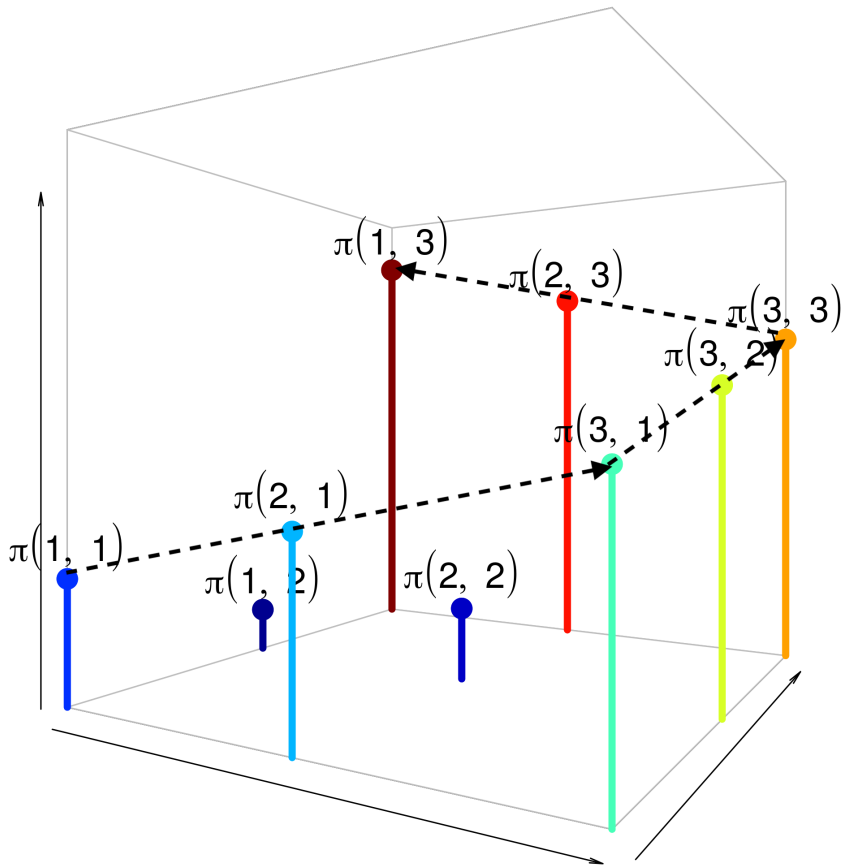
A bivariate normal distribution.

Unimodal distributions



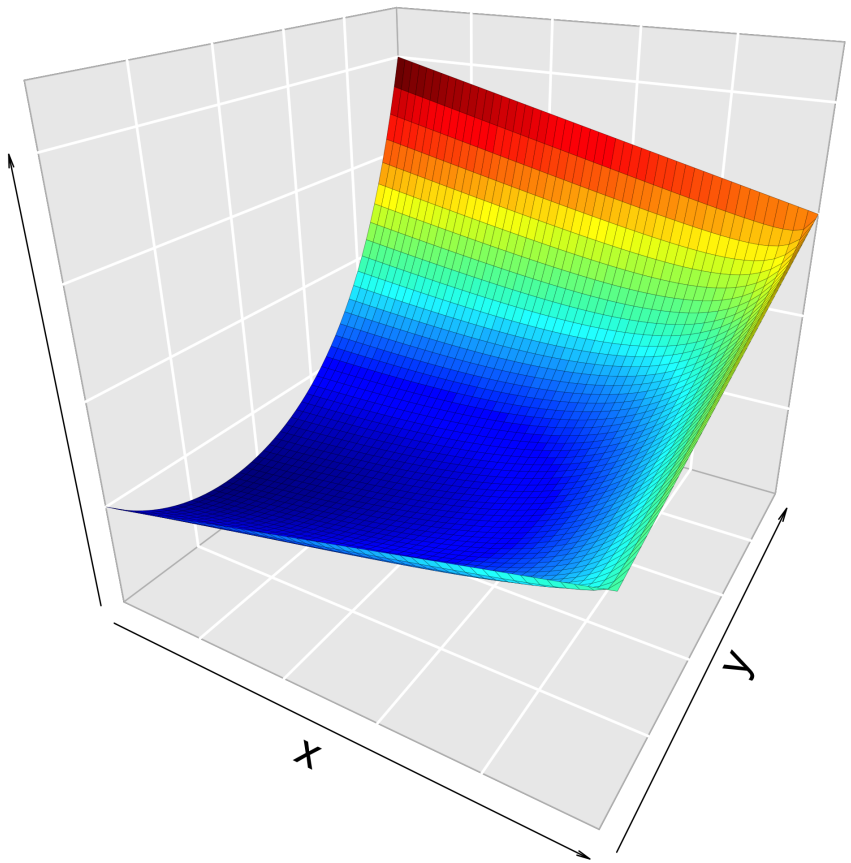
A unimodal distribution on $\mathcal{X} = \{1, 2, 3\}^2$ (the neighborhood of each point is the set of closest points).

Unimodal distributions



Moving from $(1,1)$ to $(1,3)$.

Unimodal distributions



A continuous version.

Mixing of random walk MH

Let $P_{\text{lazy}} = (P + I)/2$ denote the lazy version of a transition matrix P.

Theorem (Rapid mixing of RWMH)

Suppose $\rho = R/N > 1$. For random walk MH,

$$T_{\text{mix}}(P^{\text{RW}}_{\text{lazy}}) \leq c_\rho \, N \log \pi^{-1}_{\min},$$

where $\pi_{\min} = \min_{x \in \mathcal{X}} \pi(x)$ and c_ρ is a small constant only depending on ρ (c_ρ decreases as ρ increases).

This improves the bounds in Yang et al. [2016] (variable selection) and Zhuo and Gao [2021] (stochastic block model). Both works derive the bounds by verifying $R/N \to \infty$ and using the canonical path method of Sinclair [1992].

Mixing of informed MH

Theorem (Rapid mixing of informed MH)

Let $\rho = R/N$ and suppose $R > N^2$. There exists a locally informed MH algorithm P^{inf} such that

$$T_{\text{mix}}(P^{\text{inf}}_{\text{lazy}}) \leq 2c_\rho \log \pi^{-1}_{\text{min}}.$$

This can be proved by using Theorem 2 of [Zhou and Chang, 2023].

The bound does not involve neighborhood size N! But, constructing an informed algorithm that attains this bound in practice may be difficult.

How to check the unimodal condition?

For most high-dimensional model selection problems, establishing $R > N^c$ for any given $c > 0$ is essentially the same as establishing $R > 1$ (i.e., π is unimodal).

More precisely, the high-dimensional assumptions used to prove $R > N^c$ are essentially the same as the assumptions used to prove $R > 1$ (one only needs to modify some universal constants in the assumptions).

However, proving unimodality is often much more difficult than proving the posterior selection consistency.

High-dimensional variable selection

Let s_0 be a sparsity parameter and define the model space by

$$\mathcal{X} = \{x \subseteq \{1, \ldots, p\} \colon |x| \leq s_0\}.$$

- In high-dimensional settings, we typically let $s_0 \to \infty$.
- For add-delete-swap MH samplers, $N = \max_x |\mathcal{N}(x)| \leq p^2$.
- Let $x^* \in \mathcal{X}$ denote the true model. Yang et al. [2016] proved that, under mild assumptions,
  - $\pi(x^*)$ converges to 1 in probability (strong selection consistency);
  - with high probability, $R \geq p^\nu$ for some universal constant $\nu > 2$ (this implies strong selection consistency);
  - with high probability, the mixing time of the random walk add-delete-swap sampler is $O(pns_0^2 \log p)$.

LIT-MH for high-dimensional variable selection

In [Zhou et al., 2022], we propose an informed MH algorithm using add-delete-swap proposals, named LIT-MH (Metropolis–Hastings with Locally Informed and Thresholded proposals).

Theorem (Dimension-free mixing of LIT-MH)

Under the high-dimensional assumptions of [Yang et al., 2016] and assuming the parameters of the LIT-MH proposal are properly chosen,

$$T_{\text{mix}}(P^{\text{LIT}}_{\text{lazy}}) \leq Cn$$

for some universal constant $C \in (0, \infty)$.

High-dimensional structure learning

DAG model

A p-variate directed acyclic graph (DAG) encodes the conditional independence (CI) relations among p node variables.

Structure learning

Learn the underlying DAG model of a p-variate probability distribution from n observations.

Defining the model space is tricky.

Markov equivalence class

Two DAGs are Markov equivalent if they encode the same set of CI relations, e.g. $i \to j \to k$ and $i \leftarrow j \to k$.

High-dimensional structure learning

In Zhou and Chang [2023], we proved the unimodal condition and rapid mixing of an RWMH sampler on the space of sparse equivalence classes.

- Sparisty is imposed by using in-degree and out-degree constraints.
- We only consider equivalence classes that have a member DAG that satisfy the node degree cosntraints.
- Challenges:
  - construct a proper local neighborhood relation $\mathcal{N}$,
  - prove that π is unimodal w.r.t. $\mathcal{N}$; in particular, the degree constraints make the analysis very difficult.
- Open problems: rapid mixing on the DAG space and order space.

Potential issues with informed MH algorithms

Devising an efficient informed MH algorithm for model selection problems can be surprisingly challenging.

- It is easy to assign larger proposal probabilities to states with larger posterior, e.g. $K(x, y) \propto \pi(y)$.
- But it is difficult to control the acceptance probability.
- Informed MH algorithms can mix even more slowly than RWMH.

Henceforth, we will consider locally balanced proposals (a class of informed proposals).

Locally balanced proposals

Balancing function

We say $h\colon (0,\infty) \to (0,\infty)$ is a balancing function if

$$h(u) = u\, h(1/u), \quad \forall\, u > 0.$$

Examples: $h(u) = 1 + u, h(u) = \min\{1, u\}, h(u) = \sqrt{u}$.

Locally balanced proposals (Zanella, JASA 2020)

Let $\mathcal{N}\colon \mathcal{X} \to 2^{\mathcal{X}}$ be given and choose a balancing function h. Propose $y \in \mathcal{N}(x)$ with probability $\propto h\left(\frac{\pi(y)}{\pi(x)}\right)$, i.e.,

$$K_h(x,y) = \frac{h\left(\frac{\pi(y)}{\pi(x)}\right)}{Z_h(x)} \mathbb{1}_{\mathcal{N}(x)}(y), \text{ where } Z_h(x) = \sum_{x' \in \mathcal{N}(x)} h\left(\frac{\pi(y)}{\pi(x)}\right).$$

Potential issues with informed MH algorithms

The acceptance probability of moving from x to y is

$$\begin{aligned}
\alpha_h(x,y) &= \min\left\{1, \frac{\pi(y)K_h(y,x)}{\pi(x)K_h(x,y)}\right\} \\
&= \min\left\{1, \frac{uh(u^{-1})/Z_h(y)}{h(u)/Z_h(x)}\right\} && \left(\text{let } u = \frac{\pi(y)}{\pi(x)}\right) \\
&= \min\left\{1, \frac{Z_h(x)}{Z_h(y)}\right\}, && (\text{since } h \text{ is a balancing function}).
\end{aligned}$$

Let $y = \arg\max_{z\in\mathcal{N}(x)} \pi(z)$. The behavior of $Z_h(x)/Z_h(y)$ in model selection problems is unpredictable. For variable selection, if predictors are correlated, this ratio can easily be exceedingly small when $n \to \infty$.
See Zhou et al. [2022] for a toy example.

A simple solution

The acceptance probability of moving from x to y is

$$\begin{aligned}
\alpha_h(x,y) &= \min\left\{1, \frac{\pi(y)K_h(y,x)}{\pi(x)K_h(x,y)}\right\} \\
&= \min\left\{1, \frac{uh(u^{-1})/Z_h(y)}{h(u)/Z_h(x)}\right\} && \left(\text{let } u = \frac{\pi(y)}{\pi(x)}\right) \\
&= \min\left\{1, \frac{Z_h(x)}{Z_h(y)}\right\}, && (\text{since } h \text{ is a balancing function}).
\end{aligned}$$

Solution: Replace π with a new target distribution $\pi_h(x) \propto \pi(x)Z_h(x)$. Then the acceptance probability is always 1.

Informed importance tempering (IIT)

Algorithm:

- We draw samples $x^{(1)}, \ldots, x^{(t)}$ from the Markov chain K_h (transition matrix of the locally balanced proposal scheme).
- Since K_h is reversible w.r.t. π_h, we calculate the importance weights $\omega^{(1)}, \ldots, \omega^{(t)}$ by $\omega^{(k)} = \pi(x^{(k)})/\pi_h(x^{(k)}) \propto 1/Z_h(x)$.

This is also the main idea behind the tempered Gibbs sampler of Zanella and Roberts [2019], which uses balancing function $h(u) = 1 + u$.

"Importance tempering" just means to run an MCMC targeting some $\tilde{\pi}$ and use importance sampling to correct for the bias. This dates back to Hastings [1970].

Rapid convergence of IIT

An obvious advantage of IIT is that the chain is always moving (we assume the informed proposal satisfies $K_h(x, x) = 0$).

In Zhou and Smith [2022], we show that:

- If $R/N \to \infty$, IIT with $h(u) = 1 + u$ converges extremely fast and has overall complexity $O(N^2/R)$ (see our paper for definition).
- However, $h(u) = 1 + u$ is too aggresive and can be very inefficient for multimodal targets.
- $h(u) = \sqrt{u}$ performs well in a wider range of settings.

Extensions of IIT

This importance tempering trick turns out to be widely applicable. Most existing Metropolis-Hastings schemes can be converted to IIT versions.

Another perspective on MH algorithms

Actually, even the standard random walk MH algorithm is an importance tempering scheme where

- samples are accepted states, and
- sojourn times give unbiased estimates of importance weights.

See, e.g., Douc and Robert [2011].

Extensions of IIT

In Li et al. [2023], we propose the following IIT variants:

- IIT schemes that do not require posterior evaluation of all neighboring states;
- integration of IIT and simulated tempering algorithm;
- integration of IIT and pseudo-marginal methods;
- importance-tempered multiple-try algorithm, which is applicable to general state spaces.

IIT schemes appear to always converge faster than their MH counterparts in our numerical studies.

Example: importance tempering of MTM

Multiple-try Metropolis (MTM) algorithms are widely used but known to suffer from low acceptance rates [Yang and Liu, 2023]. Chang et al. [2022], Gagnon et al. [2022] proposed to use balancing functions to construct locally balanced MTM schemes, which tend to have high acceptance rates.

But importance tempering can be used to further enforce the sampler to always accept the proposal. The main idea is similar to that behind IIT.

Example: importance tempering of MTM

Locally balanced MTM on general state spaces

Let $Q(x, \cdot)$ denote an *uninformed* symmetric proposal with density q such that $q(x, y) = q(y, x)$. Let h be a balancing function.

An iteration of MTM at state x:

1. Draw $y_1, \ldots, y_m$ from $Q(x, \cdot)$.
2. Select y from $y_1, \ldots, y_m$ with probability $\propto h(\pi(y)/\pi(x))$.
3. Draw $x_1, \ldots, x_{m-1}$ from $Q(y, \cdot)$. Set $x_m = x$.
4. Accept y with probability

$$\min\left\{1, \frac{Z_h(x, y_1, \ldots, y_m)}{Z_h(y, x_1, \ldots, x_m)}\right\},$$

where $Z_h(x, y_1, \ldots, y_m) = \sum_{k=1}^{m} h(\pi(y_k)/\pi(x))$.

Example: importance tempering of MTM

Multiple-try importance tempering

In Step 4, we can actually just accept y and assign to the previous state x importance weight $1/Z_h(x, y_1, \ldots, y_m)$.

Caveat: in the next iteration, the m candidate neighboring states of y are NOT resampled (we have already generated them in Step 3).

Why is it correct? One can show that this algorithm is just a standard IIT algorithm on an augmented space with auxiliary variables being the m candidate neighboring states.

A comment on Peskun's ordering

Consider random walk MH algorithms with symmetric proposals. One can accept $x \to y$ with probability $\min\{1, \pi(y)/\pi(x)\}$, or $\pi(y)/(\pi(x) + \pi(y))$ (Barker's acceptance probability), etc.

The choice $\min\{1, \pi(y)/\pi(x)\}$ is most popular since Peskun's ordering guarantees that this is the optimal choice.

But if we use IIT to exactly calculate the importance weight, we no longer have such results. The optimal choice of the balancing function depends on the problem.

Concluding remarks

- Informed MCMC methods are useful and can be much more efficient than RWMH for model selection problems. Local evaluation of π can be easily parallelized.
- For any high-dimensional model selection problems, once one prove the unimodality condition, our bounds can be used to obtain the rapid mixing results for RWMH and informed MH.
- Informed proposal scheme needs to be chosen with caution.
- Importance tempering seems always better than Metropolis-Hastings for utilizing informed proposals.

Thank you!

- Q. Zhou and H. Chang. "Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes." *Annals of Statistics*, `arXiv:2101.04084`.
- Q. Zhou, J. Yang, D. Vats, G. Roberts and J. Rosenthal. "Dimension-free mixing of high-dimensional Bayesian variable selection." *JRSSB*, `arXiv:2105.05719`.
- Q. Zhou and A. Smith. "Rapid convergence of informed importance tempering." *AISTATS* (oral presentation), `arXiv:2107.10827`.
- G. Li, A. Smith and Q. Zhou. "Importance is Important: A Guide to Informed Importance Tempering Methods." `arXiv:2304.06251`.

References I

Hyunwoong Chang, Changwoo Lee, Zhao Tang Luo, Huiyan Sang, and Quan Zhou. Rapidly mixing multiple-try Metropolis algorithms for model selection problems. *Advances in Neural Information Processing Systems*, 35:25842–25855, 2022.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Randal Douc and Christian P Robert. A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *The Annals of Statistics*, 39(1):261–277, 2011.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.

Philippe Gagnon, Florian Maire, and Giacomo Zanella. Improving multiple-try Metropolis with local balancing. *arXiv preprint arXiv:2211.11613*, 2022.

JE Griffin, KG Łatuszyński, and MFJ Steel. In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. *Biometrika*, 108(1):53–69, 2021.

References II

Chris Hans, Adrian Dobra, and Mike West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.

WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97, 1970.

Guanxun Li, Aaron Smith, and Quan Zhou. Importance is important: A guide to informed importance tempering methods. *arXiv preprint arXiv:2304.06251*, 2023.

Oren Mangoubi and Aaron Smith. Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 586–595. PMLR, 2019.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *arXiv preprint arXiv:1909.05503*, 2019.

Minsuk Shin, Anirban Bhattacharya, and Valen E Johnson. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053, 2018.

Alistair Sinclair. Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, probability and Computing*, 1(4):351–370, 1992.

Michalis K Titsias and Christopher Yau. The Hamming ball sampler. *Journal of the American Statistical Association*, 112(520):1598–1611, 2017.

References III

Xiaodong Yang and Jun S Liu. Convergence rate of multiple-try Metropolis independent sampler. *Statistics and Computing*, 33(4):79, 2023.

Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

Giacomo Zanella and Gareth Roberts. Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society Series B*, 81(3):489–517, 2019.

Quan Zhou and Hyunwoong Chang. Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes. *Annals of Statistics*, to appear, 2023.

Quan Zhou and Aaron Smith. Rapid convergence of informed importance tempering. pages 10939–10965, 2022.

Quan Zhou, Jun Yang, Dootika Vats, Gareth O Roberts, and Jeffrey S Rosenthal. Dimension-free mixing for high-dimensional bayesian variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1751–1784, 2022.

Bumeng Zhuo and Chao Gao. Mixing time of Metropolis-Hastings for Bayesian community detection. *Journal of Machine Learning Research*, 22:10–1, 2021.