Complexity of local MCMC algorithms for high-dimensional variable selection

Speaker: Quan Zhou
Texas A&M University, USA

Collaborators

G. Roberts (Warwick)

J. Rosenthal (Toronto)

D. Vats (IIT Kanpur)

J. Yang (Oxford)

MCMC in high-dimensional statistics

- MCMC (Markov chain Monte Carlo) methods are frequently used in Bayesian statistics for sampling from posterior distributions.
- Let $\mathcal{S}$ denote the state space. In model selection problems, typically $|\mathcal{S}|$ (the cardinality of $\mathcal{S}$) depends on a parameter p (number of variables).

Example (variable selection)

$\mathcal{S} = \{0, 1\}^p$ and $|\mathcal{S}| = 2^p$.

Example (structure learning)

$\mathcal{S}$ is the space of p-vertex directed acyclic graphs, and $|\mathcal{S}|$ grows super-exponentially with p.

MCMC in high-dimensional statistics

- In high-dimensional settings with $p \gg n$ (n denotes the sample size), some *sparsity* constraint needs to be imposed, but usually $|\mathcal{S}|$ still grows *super-polynomially* with p.
- Compared with other approximate methods for posterior computation, e.g. variational Bayes [1], are MCMC algorithms efficient enough?

Definition (rapid mixing)

We say an MCMC algorithm is rapidly mixing if its mixing time grows only polynomially with p and n.

Metropolis-Hastings (MH) algorithms

Let $\mathcal{S}$ be a finite state space and π be a probability distribution defined on $\mathcal{S}$ (assume $\pi(x) > 0$ for each x). Given an irreducible transition matrix K, we define

$$P(x,y) = \begin{cases} K(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{K(y,x)}{K(x,y)}\right\}, & \text{if } x \neq y, \\ 1 - \sum_{x' \neq x} P(x,x'), & \text{if } x = y. \end{cases}$$

- P is *reversible* with respect to π.
- To simulate a Markov chain with transition matrix P, we only need to know an *un-normalized* version of π.

Metropolis-Hastings (MH) algorithms

Let $\mathcal{S}$ be a finite state space and π be a probability distribution defined on $\mathcal{S}$ (assume $\pi(x) > 0$ for each x). Given an irreducible transition matrix K, we define

$$P(x,y) = \begin{cases} K(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{K(y,x)}{K(x,y)}\right\}, & \text{if } x \neq y, \\ 1 - \sum_{x' \neq x} P(x,x'), & \text{if } x = y. \end{cases}$$

- $K(x,\cdot)$ is called the proposal distribution.
- $\min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{K(y,x)}{K(x,y)}\right\}$ is called the acceptance probability.

Local proposals

Let $\mathcal{N}(x) = \{y \in \mathcal{S} \colon K(x, y) > 0\}$ denote the neighborhood of x.

- In practice, $|\mathcal{N}(\cdot)|$ usually grows *polynomially* with p.
- Most "standard" MH algorithms use random walk proposals,

$$K(x, y) = \frac{\mathbb{1}_{\mathcal{N}(x)}(y)}{|\mathcal{N}(x)|}, \quad \forall\, x, y \in \mathcal{S}.$$

Locally informed proposals

Making proposals informed (Zanella [2])

Choose some $f : (0, \infty) \to (0, \infty)$, and define a new proposal transition matrix K_f by

$$K_f(x, y) = \frac{f\left(\pi(y)/\pi(x)\right)}{Z(x)} \mathbb{1}_{\mathcal{N}(x)}(y), \text{ where } Z(x) = \sum_{x' \in \mathcal{N}(x)} f\left(\frac{\pi(x')}{\pi(x)}\right).$$

In words, we propose $y \in \mathcal{N}(x)$ with probability $\propto f\left(\pi(y)/\pi(x)\right)$.

Locally informed proposals

$$K_f(x,y) = \frac{f\left(\pi(y)/\pi(x)\right)}{Z(x)} \mathbb{1}_{\mathcal{N}(x)}(y), \text{ where } Z(x) = \sum_{x' \in \mathcal{N}(x)} f\left(\frac{\pi(x')}{\pi(x)}\right).$$

- Intuitively, we prefer non-decreasing f.
- The calculation of $Z(x)$ requires us to evaluate π for each $y \in \mathcal{N}(x)$.
- Similar ideas are used in other MCMC methods [3, 4, 5] and some non-MCMC methods as well [6, 7, 8].

Can informed MCMC methods achieve a sufficiently fast convergence rate that offsets the cost of computing $Z(x)$ in each iteration?

What can we say about π?

Definition (selection consistency)

We say a Bayesian model selection procedure has selection consistency if, for some $x^* \in \mathcal{S}$, $\pi(x^*) \to 1$ in probability w.r.t. the true data generating process. (Here, $\mathcal{S}, \pi, x^*$ are all implicitly indexed by n.)

If π concentrates on a single point x^*, the mixing time of an MCMC algorithm is equivalent to the expected hitting time of x^* [9, 10].

Selection consistency can often be proved by showing that π is unimodal (w.r.t. a local neighborhood relation) and the peak is "sharp".

Sparse linear regression

Consider the linear regression model $y = X\beta + \epsilon$.

- X is an $n \times p$ matrix (n: sample size; p: number of variables).
- We are mostly interested in the case $p \gg n$.
- $\beta \in \mathbb{R}^p$ is assumed to be *sparse*: most entries are zero or "negligible".
- ϵ represents normal i.i.d. errors.

Sparse linear regression

Let γ denote the set of variables that have non-negligible effects. The goal of "variable selection" is to identify γ from the data.

Model space

Due to the sparsity assumption, we can assume γ takes value in the space

$$\mathcal{M}(s_0) = \{\gamma \subseteq \{1, 2, \ldots, p\} : |\gamma| \leq s_0\},$$

for some constant s_0 (which may grow with p).

Local MH algorithms for variable selection

Add-delete-swap neighborhood

For each $\gamma \in \mathcal{M}(s_0)$, define

$$\mathcal{N}_{\text{add}}(\gamma) = \{\gamma' \in \mathcal{M}(s_0) : \gamma' = \gamma \cup \{j\} \text{ for some } j \notin \gamma\},$$
$$\mathcal{N}_{\text{del}}(\gamma) = \{\gamma' \in \mathcal{M}(s_0) : \gamma' = \gamma \setminus \{k\} \text{ for some } k \in \gamma\},$$
$$\mathcal{N}_{\text{swap}}(\gamma) = \{\gamma' \in \mathcal{M}(s_0) : \gamma' = (\gamma \cup \{j\}) \setminus \{k\} \text{ for some } j \notin \gamma,\ k \in \gamma\}.$$

Note $|\mathcal{N}_{\text{add}}(\gamma) \cup \mathcal{N}_{\text{del}}(\gamma)| = p$ and $|\mathcal{N}_{\text{swap}}(\gamma)| = (p - |\gamma|)|\gamma|$.

Local MH algorithms for variable selection

Using the addition/deletion/swap moves, we can define a random walk MH algorithm as follows.

Symmetric RWMH for variable selection

Given current state γ,

- with probability $1/2$, propose a state from $\mathcal{N}_{\text{add}}(\gamma) \cup \mathcal{N}_{\text{del}}(\gamma)$ randomly with equal probability;
- with probability $1/2$, propose a state from $\mathcal{N}_{\text{swap}}(\gamma)$ randomly with equal probability.

Challenge I: π can be highly "irregular"

Let γ^* denote the true set of "influential" covariates, and let $\gamma \neq \gamma^*$.
Even if n is sufficiently large,

- moving from γ to $\gamma \cup \{j\}$ for some $j \in \gamma^* \setminus \gamma$ may not increase the posterior probability,
- moving from γ to $\gamma \setminus \{k\}$ for some $k \in \gamma \setminus \gamma^*$ may not increase the posterior probability.
- Reason: dependence among the p variables.
- But (with high probability) there always exists one addition or deletion move at γ which can increase the posterior probability.

Rapid mixing of RWMH

Yang et al. (*Ann. Stat.*, 2016) proved that, under mild high-dimensional assumptions, the symmetric RWMH algorithm for Bayesian variable selection is *rapidly mixing*.

- The order of their mixing time bound is roughly $pns_0^2 \log p$.
- The proof relies on the canonical path method of Sinclair [11]; see [12, 13] for the general theory.

How fast can the mixing of an informed algorithm be?

Challenge II: a naive informed scheme can easily fail

A naive informed proposal

Let $\mathcal{N}(\gamma) = \mathcal{N}_{\text{add}}(\gamma) \cup \mathcal{N}_{\text{del}}(\gamma) \cup \mathcal{N}_{\text{swap}}(\gamma)$, and

$$K(\gamma, \gamma') \propto \pi(\gamma') \mathbb{1}_{\mathcal{N}(\gamma)}(\gamma').$$

Suppose $\gamma^* = \{1, 2\}$ and the current model is $\gamma = \emptyset$. Then

$$P(\emptyset, \{1\}) \leq \frac{\pi(\{1\})}{\pi(\emptyset)} K(\{1\}, \emptyset) \leq \frac{\pi(\{1\})}{\pi(\emptyset)} \frac{\pi(\emptyset)}{\pi(\{1, 2\})} = \frac{\pi(\{1\})}{\pi(\{1, 2\})},$$

which tends to be extremely small for large n.

Challenge II: a naive informed scheme can easily fail

Recall the general definition of a locally informed proposal scheme.

$$K_f(\gamma, \gamma') = \frac{f\left(\pi(\gamma')/\pi(\gamma)\right)}{Z(\gamma)} \mathbb{1}_{\mathcal{N}(\gamma)}(\gamma'), \text{ where } Z(\gamma') = \sum_{\tilde{\gamma} \in \mathcal{N}(\gamma)} f\left(\frac{\pi(\tilde{\gamma})}{\pi(\gamma)}\right).$$

A main challenge is that we can say almost nothing about the behavior of the mapping $\gamma \mapsto Z(\gamma)$, for most choices of f, e.g. $f(x) = x^c$.

Solution

Choose some bounded f so that Z is also bounded.

Our algorithm: LIT-MH

We propose an informed MCMC algorithm for variable selection still using the add-delete-swap neighborhood, named LIT-MH (Metropolis–Hastings with Locally Informed and Thresholded proposals).

Step 1: partition the neighborhood

$$K_{\text{lit}}(\gamma, \gamma') = \frac{1}{3} \sum_{\star = \text{`add', `del', `swap'}} \frac{w_\star(\gamma' \mid \gamma)}{Z_\star(\gamma)} \mathbb{1}_{\mathcal{N}_\star(\gamma)}(\gamma'),$$

$$Z_\star(\gamma) = \sum_{\tilde{\gamma} \in \mathcal{N}_\star(\gamma)} w_\star(\tilde{\gamma} \mid \gamma),$$

where $w_\star(\gamma' \mid \gamma) \in [0, \infty)$ denotes the proposal weight of $\gamma' \in \mathcal{N}_\star(\gamma)$ given current state γ.

Our algorithm: LIT-MH

Step 2: assign bounded proposal weights

The proposal weight of $\gamma' \in \mathcal{N}(\gamma)$ is calculated by

$$w_\star(\gamma' \mid \gamma) = p^{\ell_\star} \vee \frac{\pi(\gamma')}{\pi(\gamma)} \wedge p^{L_\star}, \qquad \text{for } \star = \text{'add', 'del', 'swap',}$$

where $-\infty \leq \ell_\star \leq L_\star \leq \infty$ are some constants that may depend on the type of move.

Main result

Theorem (dimension-free mixing of LIT-MH)

Define the mixing time of the LIT-MH chain by

$$T_{\mathrm{mix}} = \sup_{\gamma \in \mathcal{M}(s_0)} \min\{t \geq 0 \colon \|P_{\mathrm{lit}}^t(\gamma, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq 1/4\},$$

where $\|\cdot\|_{\mathrm{TV}}$ *denotes the total variation distance. Under some mild high-dimensional assumptions and assuming the parameters of the LIT-MH proposal scheme are properly chosen (see our paper for details), we have*

$$T_{\mathrm{mix}} \leq Cn$$

for some universal constant C*.*

Main result

- The result holds under the high-dimensional assumptions used by Yang et al. [14]. Recall that they showed the mixing time of RWMH is $O(pns_0^2 \log p)$. Since $|\mathcal{N}(\cdot)|$ grows at rate ps_0, the total complexity of LIT-MH is smaller than the bound of [14] for RWMH.
- We only need to require $s_0 \log p = O(n)$, which is a "standard" asymptotic regime in high-dimensional statistical theory [15, 16, 17].
- The mixing time bound of LIT-MH derived in our paper is actually slightly smaller than $O(n)$.

Simulation study I: find the posterior mode

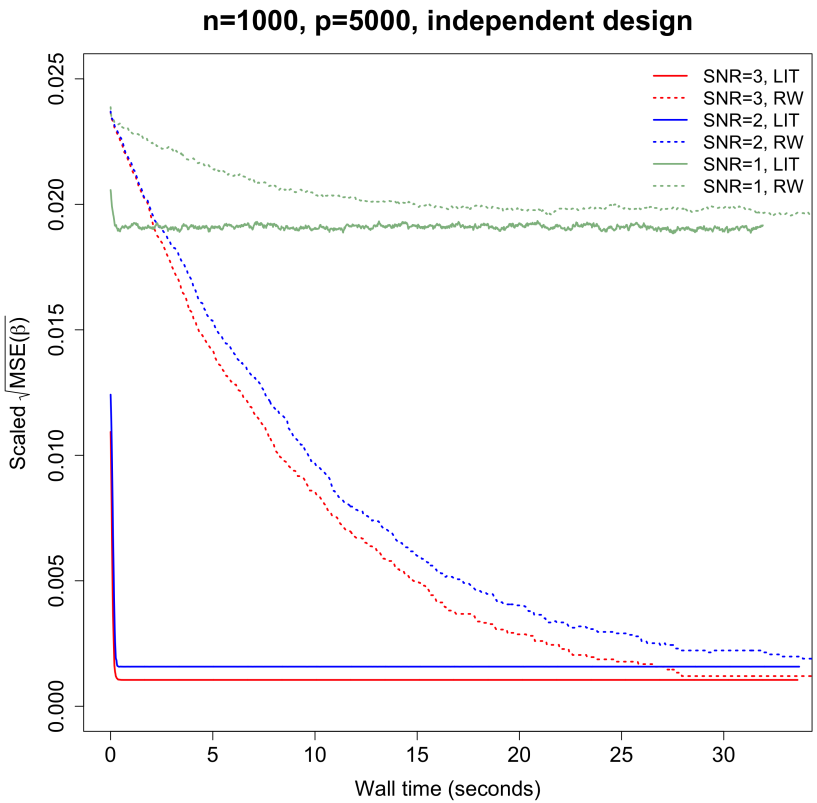
First, we considered the simulation settings of Yang et al. [14] with $|\gamma^*| = 10$. The sampler is initialized at some randomly generated $\gamma^{(0)}$ with $|\gamma^{(0)}| = 10$. When the signal-to-noise is sufficiently large, LIT-MH finds the posterior mode much faster than RWMH.

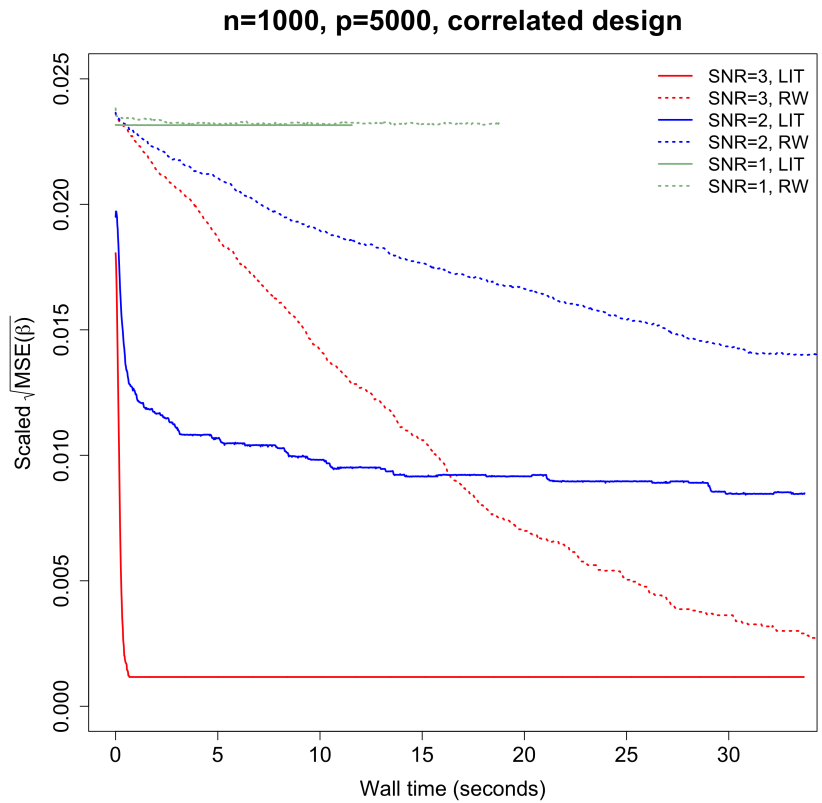
- $n = 1000, p = 5000$, independent design. RWMH: about *15* seconds; LIT-MH: *0.1* second.
- $n = 1000, p = 5000$, correlated design. RWMH: *20 to 40* seconds; LIT-MH: *0.1 to 0.2* second.

When $\gamma^* = \emptyset$, π tends to be very flat and RWMH tends to perform better.

Simulation study I: Rao-Blackwellization

No extra computational cost for Rao-Blackwellized estimation of β.





Simulation study II: exploring multimodal distributions

- The design matrix X has i.i.d. rows, but each row is sampled from $N(0, \Sigma_{d,p})$ where $\Sigma_{d,p} = \mathrm{diag}(\Sigma_d, \ldots, \Sigma_d)$ is block-diagonal. Each block Σ_d has dimension $d \times d$, and $(\Sigma_d)_{jk} = e^{-|j-k|/3}$.
- We fix $n = 1000$, $p = 5000$ and $d = 20$.
- The response y is simulated by $y = X\beta^* + z$ with $z \sim N(0, I_n)$. We generate β^* by first randomly sampling 100 nonzero entries and then sampling $\beta^*_{\gamma^*} \sim N(0, \sigma^2_\beta I_{100})$.

Simulation study II: exploring multimodal distributions

		RWMH (200K iterations)	LIT-MH (2K iterations)
$\sigma_\beta = 0.1$	Time	78.1	9.95
	Acc. Rate	0.012	0.495
	ESS/Time	4.83	34.5
$\sigma_\beta = 0.3$	Time	80.4	27.9
	Acc. Rate	0.0037	0.578
	ESS/Time	3.57	19.8
$\sigma_\beta = 0.5$	Time	81.8	42.5
	Acc. Rate	0.0021	0.485
	ESS/Time	2.45	15.0

Table: "ESS/Time" is the *effective sample size per second* calculated using $\|X\beta^{(k)}\|_2^2$. All statistics are averaged over 20 data sets.

Real GWAS data analysis

We applied our method to two real GWAS (genome-wide association study) datasets obtained from dbGaP (accession no: phs000308.v1.p1, phs000238.v1.p1). The response y is the cup-to-disk ratio measurement.

- After quality control, we end up with $n = 5,418$ and $p = 328,129$.
- RWMH has effective sample size *1.95 per minute*.
- An approximate implementation of LIT-MH has effective sample size *33.5 per minute*.
- We were able to recover 5 known GWAS hits for ocular traits located in 4 different regions.

Drift-condition approach to the analysis of LIT-MH

Drift condition

For any function g, let $(Pg)(\gamma) = \sum_{\gamma'} g(\gamma')P(\gamma, \gamma')$. If for some set $\mathsf{A} \subset \mathcal{M}(s_0)$, function $V : \mathcal{M}(s_0) \to [1, \infty)$ and constant $\lambda \in (0, 1)$,

$$(PV)(\gamma) \leq \lambda V(\gamma), \quad \forall\, x \in \mathsf{A},$$

we say the P satisfies a drift condition on A, which implies that the entry time of the Markov chain into A^c has a "thin-tailed" distribution [24].

We bound the mixing time of LIT-MH by showing that P_{lit} satisfies a two-stage drift condition.

Drift-condition approach to the analysis of LIT-MH

- To our knowledge, drift condition is rarely used in the mixing time analysis of high-dimensional *discrete* statistical problems such as variable selection.
- To establish a drift condition, we need to bound the expected change in the drift function by considering all possible moves of the chain.
- For informed MH algorithms, we need to find good bounds for the normalizing constants of proposal distributions.

Two-stage drift condition

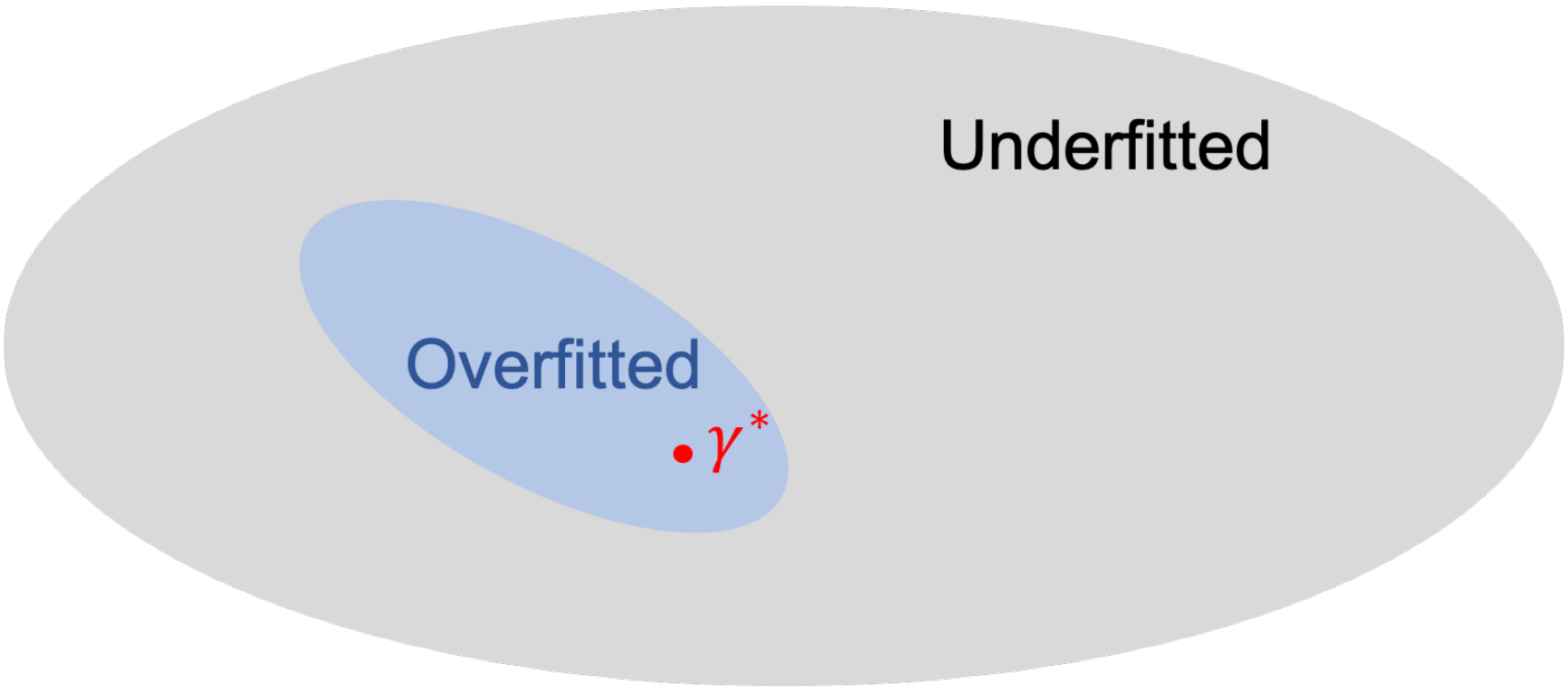
Overfitted and underfitted models

Let $\mathcal{O} = \{\gamma \in \mathcal{M}(s_0) \colon \gamma^* \subseteq \gamma\}$. Models in $\mathcal{O}$ are said to be overfitted, and other models are underfitted.

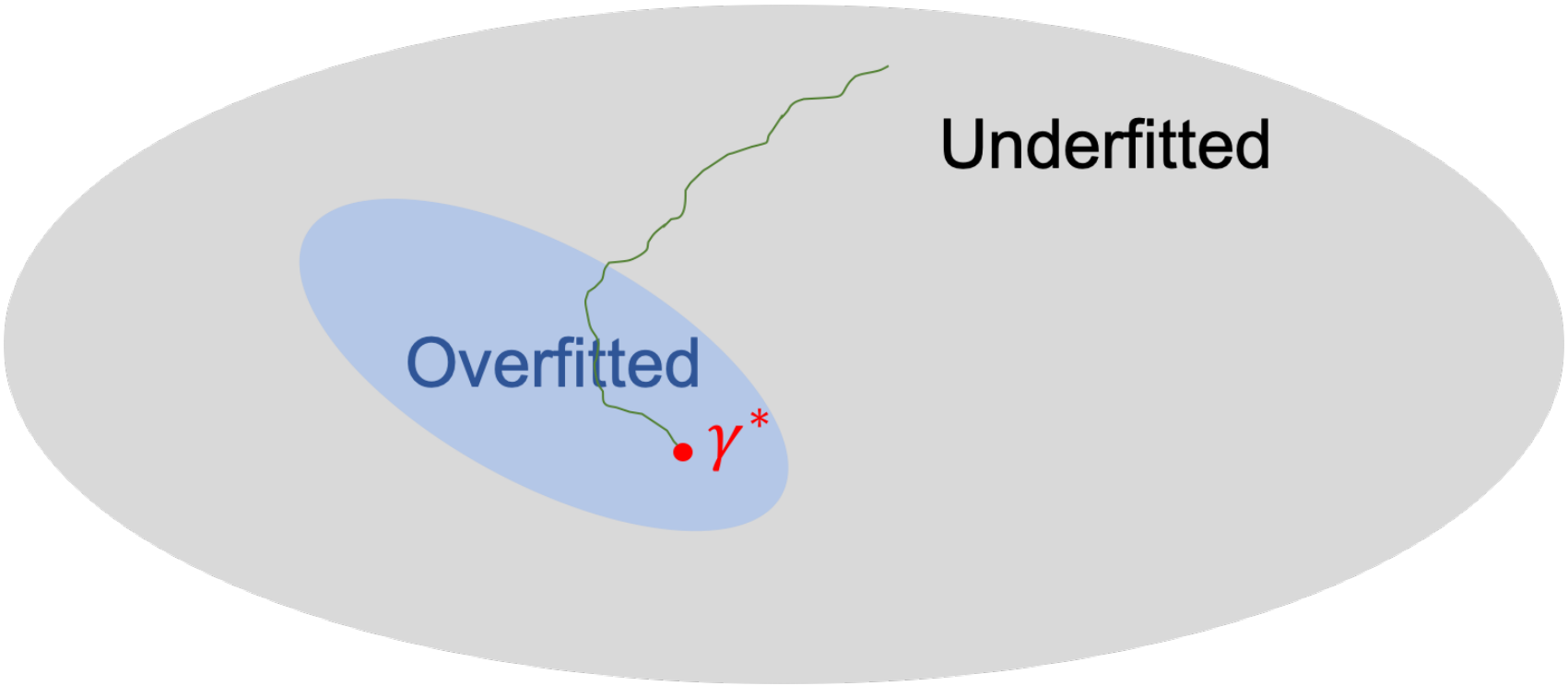
Two-stage drift condition of LIT-MH

- P_{lit} satisfies a drift condition on $\mathcal{O}^c$,
- P_{lit} satisfies another drift condition on $\mathcal{O} \setminus \{\gamma^*\}$,
- $P_{\text{lit}}(\gamma, \mathcal{O}^c)$ is very small for any $\gamma \in \mathcal{O}$.

Two-stage drift condition



Two-stage drift condition



A path of P_{lit}.

Two drift functions

We consider the prior used by Yang et al. [14], which yields the posterior,

$$\pi(\gamma) \propto p^{-\kappa|\gamma|} \left(1 - R^2_\gamma + g^{-1}\right)^{-n/2} \mathbb{1}_{\mathcal{M}(s_0)}(\gamma),$$

where κ and g are hyperparameters and R^2_γ denotes the coefficient of determinant for regressing y on the covariates in γ.

- The term $p^{-\kappa|\gamma|}$ penalizes the *model size*.
- The term $\left(1 - R^2_\gamma + g^{-1}\right)^{-n/2}$ penalizes the *lack of fit*.
- Other priors can be used as well.

Two drift functions

The two drift functions we choose are given by

$$V_1(\gamma) = \left\{1 + g^{-1}(1 - R^2_\gamma)\right\}^{1/\log(1+g)},$$
$$V_2(\gamma) = e^{|\gamma \setminus \gamma^*|/s_0}.$$

- V_1 is used for the drift condition on underfitted models.
- V_2 is used for the drift condition on overfitted models.

Two drift functions

The two drift functions we choose are given by

$$V_1(\gamma) = \left\{1 + g^{-1}(1 - R^2_\gamma)\right\}^{1/\log(1+g)},$$
$$V_2(\gamma) = e^{|\gamma \setminus \gamma^*|/s_0}.$$

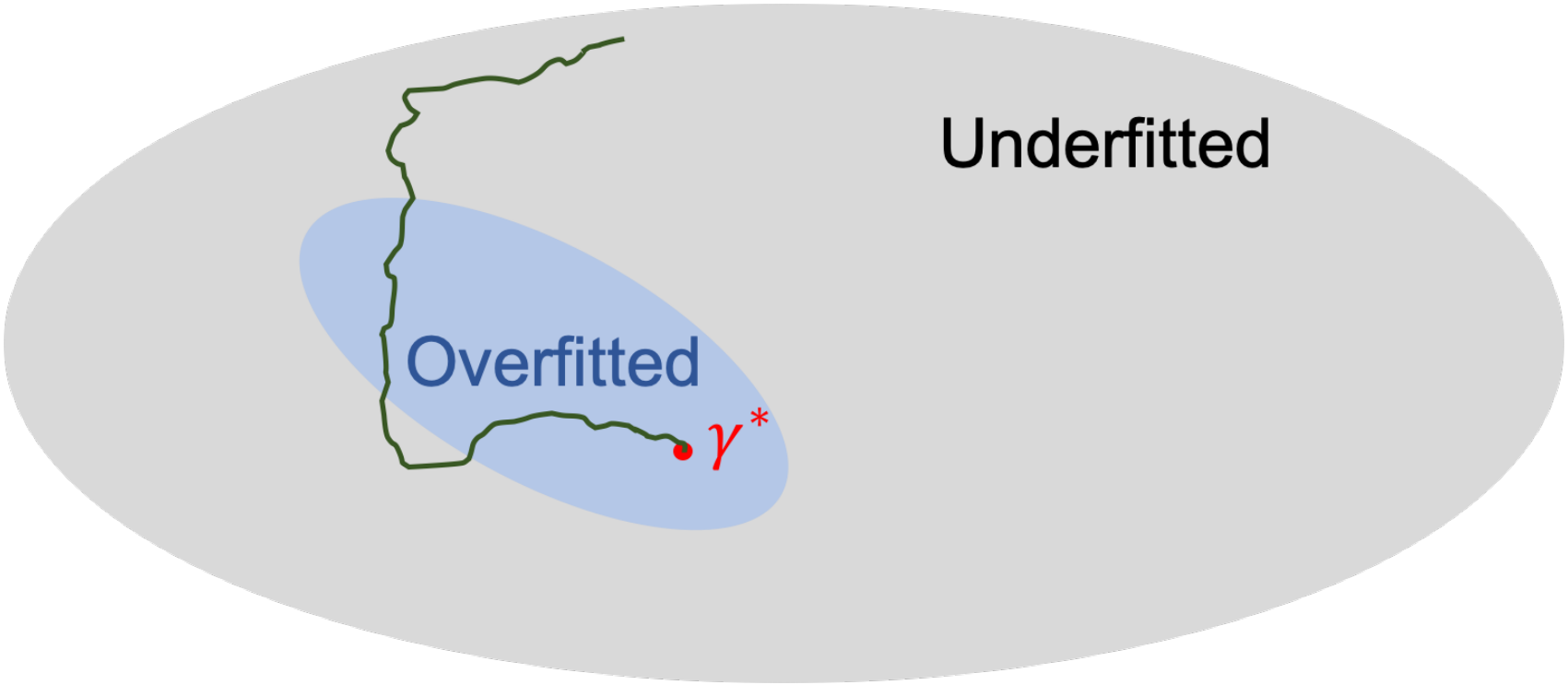
Intuition:

- When the model is underfitted, the chain tends to drift towards overfitted models to increase R^2_γ.
- When the model is underfitted, the chain tends to move towards γ^* by removing covariates in $\gamma \setminus \gamma^*$.
- Using a single drift function such as $V(\gamma) = \exp(|\gamma \triangle \gamma^*|)$ will probably fail as the behavior of the chain on $\mathcal{O}^c$ is "hard to predict".

How to bound the mixing time

- Let τ^* denote the hitting time of the true model γ^*. If we can bound $\mathbb{E}[\alpha^{-\tau^*} \mid \text{ started at some } \gamma^{(0)}]$ for some $\alpha \in (0,1)$, we can use the result of [25] to derive a mixing time bound.
- For our problem, directly bounding the generating function seems difficult. So we start by finding a tail bound instead.
- We split the path of the chain into disjoint "excursions" in $\mathcal{O}$ and $\mathcal{O}^c$. For each excursion in $\mathcal{O}$, there is some positive probability that the chain can hit γ^*, and then we can use a union bound to handle the tail probability of τ^* [26].
- The two-stage drift condition is conceptually similar to the classical drift-and-minorization approach [26, 27].

Two-stage drift condition



The chain hits γ^* during its second excursion in $\mathcal{O}$.

General results for the two-stage drift condition

Assumption on P

$(\mathsf{X}_t)_{t\in\mathbb{N}}$ is a Markov chain defined on a state space $(\mathcal{X}, \mathcal{E})$ where the σ-algebra $\mathcal{E}$ is countably generated. The transition kernel P is reversible with respect to a stationary distribution π, and the spectrum of P is non-negative.

Two-stage drift condition

Suppose that there exist two drift functions $V_1, V_2 \colon \mathcal{X} \to [1, \infty)$, constants $\lambda_1, \lambda_2 \in (0, 1)$, a set $\mathsf{A} \in \mathcal{E}$ and a point $x^* \in \mathsf{A}$ such that

$$\text{(i)} \quad PV_1 \leq \lambda_1 V_1 \ \text{ on } \mathsf{A}^c, \qquad \text{(ii)} \quad PV_2 \leq \lambda_2 V_2 \ \text{ on } \mathsf{A} \setminus \{x^*\}.$$

General results for the two-stage drift condition

Theorem (mixing time bound with the two-stage drift condition)

In addition to the two-stage drift condition, suppose that A *satisfies the following conditions, for any* $x \in \mathsf{A}$*, for some finite constants* M, K*.*

(iii) $V_1(x) = 1$*, and if* $P(x, \mathsf{A}^c) > 0$*,* $\mathbb{E}_x[V_1(\mathsf{X}_1) \mid \mathsf{X}_1 \in \mathsf{A}^c] \leq M/2$*.*

(iv) $V_2(x) \leq K$*, and if* $P(x, \mathsf{A}^c) > 0$*,* $\mathbb{E}_x[V_2(\mathsf{X}_1) \mid \mathsf{X}_1 \in \mathsf{A}^c] \geq V_2(x)$*.*

(v) $P(x, \mathsf{A}^c) \leq q$ *for some* $q < \min\{1 - \lambda_1, (1 - \lambda_2)/K\}$*.*

Then, for every $x \in \mathcal{X}$ *and* $t \in \mathbb{N}$*, we have*

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq 4\alpha^{t+1} \left(1 + M^{-1} V_1(x)\right),$$

where α *is a constant in* $(1 - q/4, 1)$ *and can be computed by*

$$\alpha = \frac{1 + \rho^r}{2} = \frac{1 + M^r/u}{2}, \quad \rho = \frac{qK}{1 - \lambda_2}, \quad u = \frac{1}{1 - q/2}, \quad r = \frac{\log u}{\log(M/\rho)}.$$

Concluding remarks

- LIT-MH is a simple but highly efficient solution to the variable selection problem. It attains a provable dimension-free mixing rate.
- Local evaluation of π can be easily parallelized.
- LIT-MH can be combined with other MCMC techniques such as blocking, tempering, lifting, etc.
- The methodology can be generalized to other model selection problems, e.g. structure learning.
- A key step of the theoretical analysis is to establish a unimodal condition, which also gives insights on how to devise efficient MCMC algorithms for model selection.

Thank you!

References I

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

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

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

[4] 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.

[5] 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.

[6] 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.

References II

[7] 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.

[8] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300, 2020.

[9] Yuval Peres and Perla Sousi. Mixing times are hitting times of large sets. *Journal of Theoretical Probability*, 28(2):488–519, 2015.

[10] Robert M Anderson, Haosui Duanmu, Aaron Smith, and Jun Yang. Drift, minorization, and hitting times. *arXiv preprint arXiv:1910.05904*, 2019.

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

[12] Laurent Saloff-Coste. Lectures on finite Markov chains. In *Lectures on probability theory and statistics*, pages 301–413. Springer, 1997.

[13] David Aldous and Jim Fill. Reversible Markov chains and random walks on graphs, 2002.

References III

[14] 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.

[15] Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

[16] Naveen Naidu Narisetty and Xuming He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

[17] Sayantan Banerjee, Ismaël Castillo, and Subhashis Ghosal. Bayesian inference in high-dimensional models. *arXiv preprint arXiv:2101.04491*, 2021.

[18] Søren F Jarner and Wai Kong Yuen. Conductance bounds on the l^2 convergence rate of Metropolis algorithms on unbounded state spaces. *Advances in Applied Probability*, pages 243–266, 2004.

[19] James Johndrow and Aaron Smith. Fast mixing of Metropolis-Hastings with unimodal targets. *Electronic Communications in Probability*, 23, 2018.

[20] Neal Madras and Dana Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, pages 581–606, 2002.

References IV

[21] Yongtao Guan and Stephen M Krone. Small-world MCMC and convergence to multi-modal distributions: From slow mixing to fast mixing. *Annals of Applied Probability*, 17(1):284–304, 2007.

[22] Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

[23] Quan Zhou and Hyunwoong Chang. Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes. *arXiv preprint arXiv:2101.04084*, 2021.

[24] Galin L Jones and James P Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334, 2001.

[25] Daniel Jerison. *The drift and minorization method for reversible Markov chains*. PhD thesis, Stanford University, 2016.

[26] Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

[27] Gareth O Roberts and Richard L Tweedie. Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and their applications*, 80(2):211–229, 1999.