The Annals of Statistics
2023, Vol. 51, No. 3, 1058–1085
https://doi.org/10.1214/23-AOS2280

COMPLEXITY ANALYSIS OF BAYESIAN LEARNING OF HIGH-DIMENSIONAL DAG MODELS AND THEIR EQUIVALENCE CLASSES

BY QUAN ZHOU[a] AND HYUNWOONG CHANG[b]

Department of Statistics, Texas A&M University, [a]*quan@stat.tamu.edu,* [b]*hwchang@stat.tamu.edu*

Structure learning via MCMC sampling is known to be very challenging because of the enormous search space and the existence of Markov equivalent DAGs. Theoretical results on the mixing behavior are lacking. In this work, we prove the rapid mixing of a random walk Metropolis–Hastings algorithm, which reveals that the complexity of Bayesian learning of sparse equivalence classes grows only polynomially in n and p, under some high-dimensional assumptions. A series of high-dimensional consistency results is obtained, including the strong selection consistency of an empirical Bayes model for structure learning. Our proof is based on two new results. First, we derive a general mixing time bound on finite-state spaces, which can be applied to local MCMC schemes for other model selection problems. Second, we construct high-probability search paths on the space of equivalence classes with node degree constraints by proving a combinatorial property of DAG comparisons. Simulation studies on the proposed MCMC sampler are conducted to illustrate the main theoretical findings.

1. Introduction.

1.1. *Gaussian DAG models and equivalence classes.*

A directed acyclic graph (DAG) encodes a set of conditional independence (CI) relations among node variables, which can be read off using the “d-separation” criterion [46]. Structure learning of DAG models from observational data plays a fundamental role in causal inference and has found many applications in machine learning and statistical data analysis [29]. In genomics, for example, DAG is a convenient device for conducting pathway analysis and inferring interactions among genes or proteins [17, 35].

Two DAGs with different edge sets can encode the same set of CI relations, in which case we say both belong to the same (Markov) equivalence class. For example, the DAGs $i \to j \to k$ and $i \leftarrow j \to k$ are Markov equivalent: both encode only one CI relation $i \perp\!\!\!\perp k \mid j$ (i.e., i, k are independent given j). But they are not Markov equivalent to $i \to j \leftarrow k$, since the latter encodes only one CI relation $i \perp\!\!\!\perp k$. A Gaussian DAG model represents a set of multivariate normal distributions that satisfy the CI constraints encoded by the DAG. Due to normality, Markov equivalence further implies distributional equivalence [19], and thus observational data alone cannot distinguish between Markov equivalent DAGs; this is a main challenge in devising efficient structure learning algorithms [10].

This paper is chiefly concerned with the following problem: given n i.i.d. observations from a p-variate DAG-perfect normal distribution, estimate the equivalence class of the underlying DAG model. This is a model selection problem where the model space is a collection of p-vertex equivalence classes. We are most interested in high-dimensional settings where p grows much faster than n and the true DAG model is sparse.

Received January 2021; revised January 2023.
MSC2020 subject classifications. 62F15, 62J05.
Key words and phrases. Finite Markov chains, greedy equivalence search (GES), locally informed proposals, Poincaré-type inequality, random walk Metropolis–Hastings, rapid mixing, strong selection consistency.

The structure learning problem can be greatly simplified if the topological ordering of the variables is known. By ordering, we mean a permutation $\sigma \in \mathbb{S}^p$, where $\mathbb{S}^p$ denotes the symmetric group on $\{1, \dots, p\}$, such that for any $i < j$, an edge connecting $\sigma(i)$ and $\sigma(j)$ is always directed as $\sigma(i) \to \sigma(j)$. Such a total ordering always exists, but may not be unique, for a DAG due to acyclicity. For example, for $1 \to 3 \leftarrow 2$, the ordering can be either $(1, 2, 3)$ or $(2, 1, 3)$. Any two different DAGs that share a same ordering cannot be Markov equivalent. This can be proved by contradiction: if the two DAGs are Markov equivalent, they must have the same skeleton [61], but the ordering uniquely determines the directions of all edges implying that the two DAGs must be the same. Henceforth, we refer to the problem as DAG selection when the ordering is known and reserve the term "structure learning" for learning equivalence classes when the ordering is unknown; the latter is the focus of this paper.

1.2. *Algorithms for Bayesian structure learning.* Most Bayesian structure learning methods aim to produce a posterior distribution of the DAG model or its equivalence class, which can be further used for making inference on quantities of interest via model averaging. To numerically approximate the posterior distribution, Markov chain Monte Carlo (MCMC) sampling is often invoked, and existing MCMC methods differ from each other mainly in three respects: the state space, the set of local operators and the proposal scheme. The local operators decide which states the sampler may move to in the next iteration (i.e., they define the "neighborhood" of each state). The proposal scheme refers to how the proposal probabilities of these neighboring states are assigned. Most existing algorithms use either random walk Metropolis–Hastings (MH) or Gibbs schemes, but we note that informed proposal schemes recently proposed in Zanella [66] and Zhou et al. [68] can be applied as well.

There are three popular choices of the state space: states can be DAGs, equivalence classes or orderings. The famous "structure MCMC" sampler is a random walk MH algorithm that searches the DAG space using addition, deletion and reversal of single edges [20, 36]. It is straightforward to implement (one only needs to check acyclicity when proposing local moves) but may not be efficient since the sampler can spend a lot of time traversing large equivalence classes. Various methods have been proposed to improve the performance by using more complicated local operators [22, 56] or blocked Gibbs schemes [21]. Directly searching the space of equivalence classes seems more efficient, but a major challenge is to construct a proper set of local operators [2, 10, 41, 48, 49]; see Madigan et al. [37] and Castellettiet al. [8] for MCMC samplers defined on the space of equivalence classes. Order MCMC methods [1, 15, 16] target a posterior distribution on the order space $\mathbb{S}^p$. They are motivated by the observation that given ordering, the conditional posterior distribution of DAGs can be evaluated relatively easily. More sophisticated MCMC schemes can be built by using partial orderings [30, 44]. It should be noted that the choice of the prior distribution typically depends on the state space, which results in essentially different posterior distributions on the three state spaces (see Section 7.1). We will focus on the space of equivalence classes.

In principle, by treating the logarithm of the posterior probability as a scoring criterion, deterministic score-based search algorithms can also be used to find the structure that maximizes the score (i.e., the maximum a posteriori estimate). This approach appears less popular in the Bayesian structure learning literature, probably because it cannot quantify the uncertainty in estimation. One of the most important score-based algorithms is the greedy equivalence search (GES) proposed by Meek [39] and Chickering [11], a two-stage greedy search algorithm defined on the space of equivalence classes. Nandy, Hauser and Maathuis [42] were the first to prove the high-dimensional consistency of GES (i.e., the search returns the true equivalence class with high probability for sufficiently large n) using an assumption called strong faithfulness. Though it is known that strong faithfulness is very restrictive [59], such conditions appear to be necessary for proving high-dimensional consistency results for many algorithms [28]. We refer readers to Drton and Maathuis [13] and Scutari, Graafland and Gutiérrez [51] for other scored-based structure learning methods.

1.3. *Overview of the paper.* While many MCMC methods for structure learning have been proposed, to our knowledge, no theoretical result on the mixing time is available. This is probably because the structure of the state space is highly complicated. The primary goal of this paper is to fill the gap by deriving nonasymptotic mixing time bounds. We find that structure MCMC and order MCMC methods are, unfortunately, hard to analyze due to the technical difficulty in bounding sizes of equivalence classes. The equivalence class sampler of Castellettiet al. [8] uses six graph operators to move between equivalence classes [23], but we can explicitly construct slow mixing examples for this sampler with fixed p. This motivates us to propose our own equivalence class sampler, RW-GES, which uses a random walk proposal scheme that mimics and generalizes the local moves employed by GES. We prove a high-dimensional rapid mixing result for the RW-GES sampler, which essentially says that, under some conditions, the number of iterations needed to find the true equivalence class grows only polynomially in n and p with high probability. The proof consists of three steps, which we now explain separately.

In Section 2, we first develop a general theory on the complexity of local MCMC algorithms for model selection. This section is self-contained and of considerable independent interest. We build a weighted path argument and use a Poincaré-type inequality [27] to obtain a novel, generally applicable mixing time bound under a unimodal assumption; see Condition 1 and Theorem 2. This result can be applied to other model selection problems such as variable selection and stochastic block models. It sharpens the existing mixing time bounds for random walk MH algorithms in the literature [65, 69] and can be utilized to derive theoretical guarantees for locally informed MH algorithms [66]. Our theory also reveals a link between optimization and sampling: if for some model selection problem, there is a greedy local search with consistency guarantee, it is hopeful that, with some modifications, we may convert the greedy algorithm to a local MH sampler that has provable rapid mixing property. In general, rapid mixing is more difficult to prove and more informative than the consistency of a greedy search, since the former characterizes the overall complexity of the algorithm and requires an analysis of the local posterior landscape in the whole state space.

The RW-GES sampler for structure learning is formally introduced in Section 3. To impose sparsity, we define the state space to be the set of all equivalence classes that satisfy some node degree constraints. We do not explicitly define the target posterior distribution in this section (which will be done in Section 4); instead, we assume the posterior has some consistency property that typically holds for sufficiently large sample sizes. To use Theorem 2, we need to verify its assumption for the structure learning problem, which requires us to bound the neighborhood size (see Lemma 1) and construct "canonical paths" for the RW-GES sampler. We show by examples (see Examples 2 and 3) that a major and unique challenge in the path construction is to verify that the equivalence classes located on the "boundary" of the restricted search space cannot be local modes. To overcome this, we introduce a "swap" proposal move to RW-GES and prove a key combinatorial property of DAGs in Lemma 2. Combining it with the well-known Chickering algorithm [11], we obtain the canonical paths of RW-GES.

In Section 4, we propose an empirical Bayes model for structure learning and prove that it has the desired high-dimensional consistency property assumed in Section 3. Our model generalizes the DAG selection model of Lee, Lee and Lin [33], and we show that it yields the same marginal fractional likelihood for Markov equivalent DAGs. The main result in this section is Theorem 5, which gives the strong selection consistency of our structure learning model. For Bayesian methods, such consistency results have only been established lately for the DAG selection problem with known ordering [7, 33]. Roughly speaking, in our consistency result, the maximum degree of searched DAGs is allowed to grow at rate $\sqrt{\log p}$; see Remark 10. The analogous high-dimensional consistency results for both variable selection and DAG selection are obtained as intermediate steps of our proof of Theorem 5.

The rapid mixing of RW-GES now follows from the mixing time bound given in Theorem 2 and the results of Sections 3 and 4. It is formally stated in Section 5. For comparison, we provide two slow mixing examples in the same section. The first one (see Example 4) shows why it is difficult to relax a key assumption used in our analysis, which is called the "strong beta-min condition" and is similar to the strong faithfulness assumption. The second (see Example 5) illustrates that the equivalence class sampler of Castellettiet al. [8] may mix slowly when p is small and n is large. We conduct simulation studies in Section 6 to show that our theoretical results hold "approximately" for moderately large sample sizes and provide useful guidance on the use of RW-GES in practice. Section 7 concludes the paper with discussions on why the structure MCMC is difficult to analyze and potential extensions of RW-GES. All proofs are relegated to the Supplementary Material [67]. For readers' convenience, a notation table is given in Supplementary Material Section A.

2. Mixing time bounds for model selection problems.

2.1. *A general setup.* In this section, we use $\Theta = \Theta_p$ to denote a finite model space for a general model selection problem with p variables; for example, for the structure learning problem, each $\theta \in \Theta$ can be a unique equivalence class. Let $\mathcal{N}\colon \Theta \to 2^{\Theta}$ be given such that $\theta \notin \mathcal{N}(\theta)$ for each $\theta \in \Theta$; $\mathcal{N}$ is called a neighborhood function. We say θ' is a neighbor of θ if and only if $\theta' \in \mathcal{N}(\theta)$. We say $\mathcal{N}$ is "symmetric" if $\theta \in \mathcal{N}(\theta')$ always implies $\theta' \in \mathcal{N}(\theta)$. When we need to emphasize Θ is equipped with $\mathcal{N}$, we denote the space by $(\Theta, \mathcal{N})$. Let π denote a posterior distribution on Θ for a Bayesian procedure; assume it is known up to a normalizing constant and $\pi(\theta) > 0$ for each θ. Given a function $h\colon (0,\infty) \to (0,\infty)$, define a Markov chain $\mathbf{K}^h$ on Θ by

$$\mathbf{K}^h(\theta,\theta') = \frac{h(\pi(\theta')/\pi(\theta))}{\sum_{\tilde{\theta}\in\mathcal{N}(\theta)} h(\pi(\tilde{\theta})/\pi(\theta))} \mathbb{1}_{\mathcal{N}(\theta)}(\theta'), \tag{1}$$

where $\mathbb{1}$ is the indicator function. That is, given current state θ, $\mathbf{K}^h$ moves to some $\theta' \in \mathcal{N}(\theta)$ with probability $\propto h(\pi(\theta')/\pi(\theta))$. Given $\mathbf{K}^h$, define another Markov chain $\mathbf{P}^h$ by

$$\mathbf{P}^h(\theta,\theta') = \begin{cases} \mathbf{K}^h(\theta,\theta') \min\left\{1, \dfrac{\pi(\theta')\mathbf{K}^h(\theta',\theta)}{\pi(\theta)\mathbf{K}^h(\theta,\theta')}\right\}, & \text{if } \theta' \neq \theta, \\ 1 - \sum_{\tilde{\theta}\neq\theta} \mathbf{P}^h(\theta,\tilde{\theta}), & \text{if } \theta' = \theta. \end{cases} \tag{2}$$

If $\mathbf{P}^h$ is irreducible, then π is the unique stationary distribution of $\mathbf{P}^h$. To avoid periodicity, we will often work with the lazy version $\mathbf{P}^h_{\text{lazy}} = (\mathbf{P}^h + \mathbf{I})/2$, where $\mathbf{I}$ is the identity matrix.

DEFINITION 1 (Local Metropolis–Hastings algorithms). We say $\mathbf{P}^h$ defined by (2) is a local MH algorithm with local proposal $\mathbf{K}^h$. If $h \equiv 1$, we say $\mathbf{P}^h$ is the random walk MH algorithm. If h is nonconstant and nondecreasing, we say $\mathbf{K}^h$ is a (locally) informed proposal and $\mathbf{P}^h$ is a (locally) informed MH algorithm.

The locally informed MH algorithm was proposed by Zanella [66]. The main idea is to assign larger proposal probabilities to those neighboring states with larger posterior so that the chain can quickly move to high-posterior regions. Let h in (1) be $h(u) = u^a$ for some $a \geq 0$. Observe that when $a = 0$, $\mathbf{K}^h$ is reduced to the random walk proposal, and when $a \to \infty$ we obtain the greedy search (see the definition below). So, informed proposals are generally more aggressive than random walk but less aggressive than greedy search.

DEFINITION 2 (Greedy local search). A greedy (local) search on $(\Theta, \mathcal{N})$ with initial state $\theta^{(0)}$ generates $\theta^{(1)}, \theta^{(2)}, \ldots,$ sequentially by letting $\theta^{(i)} = \arg\max_{\theta' \in \mathcal{N}(\theta^{(i-1)}) \cup \{\theta^{(i-1)}\}} \pi(\theta')$ for each $i \geq 1$. The search stops and returns $\theta^{(j)}$ if $\theta^{(j)} = \theta^{(j-1)}$.

The efficiency of both greedy search and MH algorithms largely depends on the choice of $\mathcal{N}$. For model selection problems, $\mathcal{N}(\cdot)$ is usually much smaller than Θ so that the algorithm is computationally affordable. But $\mathcal{N}$ should also provide enough connectivity so that the algorithm cannot get trapped at suboptimal local modes (θ is a local mode if $\pi(\theta) > \pi(\theta')$ for any $\theta' \in \mathcal{N}(\theta)$). We measure the convergence rate of MH algorithms using mixing time.

DEFINITION 3 (Mixing time). Let $\mathbf{P}$ be an irreducible and aperiodic transition matrix defined on a finite state space Θ, with stationary distribution π. Define its mixing time by

$$T_{\text{mix}}(\mathbf{P}) = \max_{\theta \in \Theta} \min\{t \geq 0 \colon \|\mathbf{P}^t(\theta, \cdot) - \pi(\cdot)\|_{\text{TV}} \leq 1/4\},$$

where $\|\cdot\|_{\text{TV}}$ denotes the total variation distance which takes value in [0, 1].

REMARK 1. We say an MCMC algorithm is rapidly mixing if its mixing time grows at most polynomially in the complexity parameters n (sample size) and p (number of variables). For most high-dimensional model selection problems, the size of Θ grows at least superpolynomially with p. For variable selection, which is probably the best-studied problem, Yang, Wainwright and Jordan [65] proved the rapid mixing of a random walk MH algorithm, and Zhou et al. [68] showed that an informed MH algorithm can converge much faster and obtain a mixing time independent of p.

2.2. *A multipurpose path method.* We propose a general method for bounding the mixing time of $\mathbf{P}^h$ defined in (2) and proving consistency properties of the posterior distribution π. The bounds to be derived in this section are nonasymptotic, and p is treated as a fixed constant. We begin by assuming that the triple $(\Theta, \mathcal{N}, \pi)$ satisfies the following condition, where $|\cdot|$ denotes the cardinality of a set.

CONDITION 1. $|\Theta| < \infty$, $\mathcal{N}$ is symmetric, and $\pi > 0$. There exists a function $g \colon \Theta \to \Theta$, a state $\theta^* \in \Theta$ and constants $t_1, t_2 > 0$, $p > 1$ such that (i) $|\mathcal{N}(\theta)| \leq p^{t_1}$ for each $\theta \in \Theta$, and (ii) $g(\theta) \in \mathcal{N}(\theta)$ and $\pi(g(\theta))/\pi(\theta) \geq p^{t_2}$ for each $\theta \neq \theta^*$.

REMARK 2. Part (ii) is equivalent to either of the following statements:

(a) For any $\theta \neq \theta^*$, $\max_{\theta' \in \mathcal{N}(\theta)} \pi(\theta') \geq p^{t_2} \pi(\theta)$.
(b) For any $\theta \neq \theta^*$, there exists $k < \infty$ and a sequence $(\theta_0 = \theta, \theta_1, \theta_2, \ldots, \theta_k = \theta^*)$ such that $\theta_i \in \mathcal{N}(\theta_{i-1})$ and $\pi(\theta_i)/\pi(\theta_{i-1}) \geq p^{t_2}$ for each $i = 1, \ldots, k$.

We introduce the function g because, for model selection problems, one often verifies Condition 1 by explicitly identifying some $g(\theta)$ for each θ. There may exist many choices of g so that Condition 1 holds. Without loss of generality, we always define $g(\theta^*) = \theta^*$. Then part (ii) implies that for any θ there exists $k \leq |\Theta|$ such that $g^k(\theta) = \theta^*$, and θ^* is the only attracting fixed point of g. We will call g a canonical transition function and a sequence of the form $(\theta, g(\theta), \ldots, g^k(\theta) = \theta^*)$ a canonical path. We can think of a canonical path as a candidate "greedy search path" since the posterior keeps increasing along the path, but note that a greedy search does not necessarily follow a canonical path since $g(\theta)$ may not be the maximizer of π in $\mathcal{N}(\theta)$.

Roughly speaking, in the model selection context, θ^* can be thought of as the "true" data-generating model, and Condition 1 can be interpreted as an algorithmic consistency property since it implies that θ^* is the unique mode of π and the greedy search always returns θ^*; see part (i) of Theorem 1. For variable selection, Yang, Wainwright and Jordan [65] proved that Condition 1 holds with high probability under some mild high-dimensional assumptions and then used the canonical path method of Sinclair [52] to bound the mixing time of the random walk MH algorithm. We generalize their result to our setup.

THEOREM 1. *Let Θ, $\mathcal{N}$, π, g, θ^*, t_1, t_2, p be as given in Condition 1. Let $\mathbf{P}^h$ be given by (2) and $\mathbf{P}^h_{\text{lazy}} = (\mathbf{P}^h + \mathbf{I})/2$ be its lazy version. The following statements hold*:

(i) *The greedy search always returns θ^* regardless of the initial state.*
(ii) *If $t_2 > t_1$, then $\pi(\theta^*) \geq 1 - p^{-(t_2 - t_1)}$.*
(iii) *If $t_2 > t_1$, then*

$$T_{\text{mix}}(\mathbf{P}^h_{\text{lazy}}) \leq \frac{4\ell_{\max}}{\{1 - p^{-(t_2-t_1)}\} \min_{\theta \neq \theta^*} \mathbf{P}^h(\theta, g(\theta))} \log\left(\frac{4}{\pi_{\min}}\right),$$

where $\ell_{\max} = \max_{\theta \neq \theta^*} \min\{k \geq 1 \colon g^k(\theta) = \theta^*\}$ and $\pi_{\min} = \min_{\theta \in \Theta} \pi(\theta)$.
(iv) *If $t_2 > t_1$ and $h \equiv 1$, then $\mathbf{P}^h(\theta, g(\theta)) \geq p^{-t_1}$.*

PROOF. See Supplementary Material Section B.4. □

REMARK 3. Part (ii) of Theorem 1 shows that π concentrates on θ^*, which can be further used to show the strong selection consistency of a Bayesian model selection procedure (see Section 4.3). This is very useful since we only require polynomial (in p) bounds for the ratio $\pi(g(\theta))/\pi(\theta)$ in Condition 1, while $|\Theta|$ may be (super)exponential in p. For a random walk MH algorithm, by parts (iii) and (iv), the order of mixing time is given by $p^{t_1} \ell_{\max} \log \pi_{\min}^{-1}$. For the greedy search, note that $\ell_{\max}$ is an upper bound for the steps needed to find θ^*, and in each step the search needs to evaluate π for at most p^{t_1} states. Hence, the greedy search and random walk MH algorithm have very similar complexity.

We now show that the mixing time bound in Theorem 1 can be improved. The new bound given in Theorem 2 has two major advantages. First, it does not involve $\ell_{\max}$, which can be large. Second, it replaces $\min_{\theta \neq \theta^*} \mathbf{P}^h(\theta, g(\theta))$ in Theorem 1 with $\min_{\theta \neq \theta^*} \mathbf{P}^h(\theta, \mathcal{N}^*(\theta))$, where $\mathcal{N}^*(\theta)$ is the set of all "desirable moves" for $\mathbf{P}^h$ at θ including $g(\theta)$. This is key to bounding the mixing times of informed MH algorithms. To prove Theorem 2, we use a novel path argument that may be of independent interest. For each $\theta \neq \theta^*$, we construct a set of paths from θ to θ^* using all desirable moves. By properly weighting these paths, we are able to bound the mixing time using a Poincaré-type inequality [27], which significantly generalizes the canonical path method. See Supplementary Material Section B for details.

THEOREM 2. *Let Θ, $\mathcal{N}$, π, g, θ^*, t_1, t_2, p be as given in Condition 1, and $\mathbf{P}^h$ be given by (2). For each $\theta \neq \theta^*$, define $\mathcal{N}^*(\theta) = \{\theta' \in \mathcal{N}(\theta) \colon \pi(\theta') \geq p^{t_2}\pi(\theta)\}$. Let $\pi_{\min} = \min_{\theta \in \Theta} \pi(\theta)$. If $t_2 > t_1$, then*

$$T_{\text{mix}}(\mathbf{P}^h_{\text{lazy}}) \leq \frac{2C(p, t_1, t_2) \log(\frac{4}{\pi_{\min}})}{\min_{\theta \neq \theta^*} \mathbf{P}^h(\theta, \mathcal{N}^*(\theta))} \quad \text{where } C(p, t_1, t_2) = \frac{1 + (1 - p^{t_1 - t_2})^{-1}}{[1 - p^{(t_1 - t_2)/2}]^2}.$$

PROOF. See Supplementary Material Section B.6. □

REMARK 4. Theorem 2 can be used to immediately improve some existing mixing time bounds in the literature. Both Yang, Wainwright and Jordan [65] and Zhuo and Gao [69] proved the rapid mixing of a random-walk MH algorithm for some high-dimensional discrete-state-space problem by showing Condition 1 holds for some g and using the canonical path method underlying Theorem 1. Theorem 2 shows that $\ell_{\max}$ can be dropped (in an asymptotic setting where $p \to \infty$ and $t_1 < t_2$ are fixed, $C(p, t_1, t_2) \to 2$).

REMARK 5. Another important application of Theorem 2 is the mixing time analysis of informed MH algorithms. Define $\mathcal{N}^t(\theta) = \{\theta' \in \mathcal{N}(\theta) \colon \pi(\theta') \geq p^t \pi(\theta)\}$. If t_2 is sufficiently large and the function h in (1) is chosen properly, it is often possible to show that $\min_{\theta \neq \theta^*} \mathbf{P}^h(\theta, \mathcal{N}^t(\theta)) \geq c$ for some $t > t_1$ and fixed constant $c > 0$. Indeed, for the LIT-MH algorithm for variable selection considered in Zhou et al. [68], one can follow their calculations to verify that this holds for $c = 1/4$, and then by Theorem 2, the order of the mixing time is only $\log \pi_{\min}^{-1}$. This cannot be achieved by using Theorem 1, since $\mathbf{P}^h(\theta, g(\theta))$ can be as small as $O(p^{-t_1})$ (e.g., when all neighboring states have the same posterior probabilities).

The theory developed in this section relies on Condition 1, which is a property of the triple $(\Theta, \mathcal{N}, \pi)$. If Condition 1 holds, one can use Theorem 2 to study the mixing times of any local MH algorithm. For simplicity, for the structure learning problem to be studied in the rest of this paper, we will only consider the random walk proposal, and our main task is to construct a triple $(\Theta, \mathcal{N}, \pi)$ that satisfies Condition 1. We will often define $\mathcal{N}$ on Θ and then use $\mathcal{N}$ to refer to a neighborhood relation on a restricted space $\Theta_0 \subset \Theta$; this means that the neighborhood of $\theta \in \Theta_0$ is given by $\mathcal{N}(\theta) \cap \Theta_0$. Note that even if $(\Theta, \mathcal{N}, \pi)$ satisfies Condition 1, $(\Theta_0, \mathcal{N}, \pi)$ may not, which is one challenge in the sparse structure learning problem to be considered.

3. The RW-GES sampler and its canonical paths.

3.1. *Notation and terminology.* We set up the notation to be used for the structure learning problem. Let $[p] = \{1, \ldots, p\}$ and $| \cdot |$ denote the cardinality of a set. A subset of $[p]$ is typically denoted by S. The Hamming distance between two sets S, S' is denoted by $d_{\mathrm{H}}(S, S') = |S \setminus S'| + |S' \setminus S|$.

A DAG G is a pair (V, E) where V is the vertex set and $E \subset V \times V$ is the set of directed edges. Throughout the paper, we assume $V = [p]$ for DAG models, representing random variables $\mathsf{X}_1, \ldots, \mathsf{X}_p$. Note that $(i, i) \notin E$ for any $i \in [p]$. Let $|G|$ denote the number of edges in the DAG G; thus, $|G| = |E|$. We use the notation $i \to j \in G$ to mean that $(i, j) \in E$ and $(j, i) \notin E$. The notation $i \to j \notin G$ means that $(i, j) \notin E$. For two DAGs $G = (V, E)$ and $G' = (V, E')$, we write $G' = G \cup \{i \to j\}$ if $E' = E \cup (i, j)$, and $G' = G \setminus \{i \to j\}$ if $E' = E \setminus (i, j)$. We write $G = G'$ if and only if G and G' have the same vertex set and edge set. Given a DAG G, we say node i is a parent of node j (and node j is a child of node i) if $i \to j \in G$. Let $\mathrm{Pa}_j(G) = \{i \in [p] \colon i \to j \in G\}$ denote the set of parents of node j; the in-degree of node j is $|\mathrm{Pa}_j(G)|$. The maximum in-degree of G is $\max_j |\mathrm{Pa}_j(G)|$. Similarly, let $\mathrm{Ch}_j(G) = \{i \in [p] \colon j \to i \in G\}$, and $|\mathrm{Ch}_j(G)|$ is called the out-degree of node j. The degree of a node is the sum of its in-degree and out-degree, and the maximum degree of G is $\max_j |\mathrm{Pa}_j(G) \cup \mathrm{Ch}_j(G)|$. We may simply write Pa_j if we are not referring to a specific DAG or the underlying DAG is clear from context. The Hamming distance between two DAGs G, G' is defined by $d_{\mathrm{H}}(G, G') = \sum_{j \in [p]} d_{\mathrm{H}}(\mathrm{Pa}_j(G), \mathrm{Pa}_j(G'))$.

An equivalence class of DAGs is typically denoted by $\mathcal{E}$. We always interpret $\mathcal{E}$ as a set of DAGs, and use $|\mathcal{E}|$ to denote the number of member DAGs in $\mathcal{E}$. The equivalence class of a DAG G is also denoted by $[G]$; thus, $\mathcal{E} = [G]$ if and only if $G \in \mathcal{E}$. The set of CI relations encoded by a DAG G or an equivalence class $\mathcal{E}$ is denoted by $\mathcal{CI}(G)$ or $\mathcal{CI}(\mathcal{E})$, respectively. Note that we always have $\mathcal{CI}(G) = \mathcal{CI}([G])$.

We say a p-variate distribution μ is Markovian w.r.t. a p-vertex DAG G and G is an independence map (I-map) of μ if all CI relations encoded by G hold for μ. If the converse is also true, we say μ is faithful or perfectly Markovian w.r.t. G, and G is a perfect map of μ [54, 55]. We say μ is DAG-perfect if there exists some DAG that is a perfect map of μ. A DAG G is an I-map of a DAG G' and its equivalence class $[G']$ if $\mathcal{CI}(G) \subseteq \mathcal{CI}(G')$, and G is a minimal I-map (of G') if any sub-DAG of G (different from G) is not an I-map of G'. Given the set $\mathcal{CI}(G)$, a minimal I-map of G with ordering σ, which we denote by G_σ, can be uniquely defined as follows: for any $i < j$, $\sigma(i) \to \sigma(j) \in G_\sigma$ if and only if nodes $\sigma(i), \sigma(j)$ are not conditionally independent given nodes $\{\sigma(1), \dots, \sigma(j-1)\} \setminus \{\sigma(i)\}$ [53]. An example for $p = 3$ is given below. If μ is a p-variate positive measure, a unique minimal I-map of $\mathcal{CI}(\mu)$ with ordering σ can be constructed in an analogous manner [29].

EXAMPLE 1. Let $p = 3$ and G be the DAG $1 \to 3 \leftarrow 2$. Let G_σ denote the minimal I-map of G with ordering σ. If $\sigma = (1, 2, 3)$ or $(2, 1, 3)$, then $G_\sigma = G$ since σ is an ordering of G. If σ is any other ordering, then G_σ is the complete DAG (i.e., a DAG without missing edge). For example, if $\sigma = (1, 3, 2)$, G_σ has three edges $1 \to 3$, $1 \to 2$ and $3 \to 2$; in particular, the edge $1 \to 2$ is included since $1 \not\perp\!\!\!\perp 2 | 3$ in G.

3.2. *Search spaces and posterior distributions.* To apply the general theory developed in Section 2, it suffices to construct a triple $(\Theta, \mathcal{N}, \pi)$ that satisfies Condition 1. We will do this for both high-dimensional DAG selection and structure learning. Recall that for DAG selection, our goal is to estimate an underlying DAG model from the data when we know it has some ordering σ, and for structure learning, our goal is to estimate the equivalence class of the DAG model. We first define the search spaces (i.e., model spaces) for the two problems. Let $\mathcal{G}_p$ denote the space of all p-vertex DAGs, which grows superexponentially in p. We consider two sparsity constraints for DAGs: one for the maximum in-degree and the other for the maximum out-degree. For $d_{\text{in}}, d_{\text{out}} \in [p]$, define

$$\mathcal{G}_p(d_{\text{in}}, d_{\text{out}}) = \Big\{ G \in \mathcal{G}_p : \max_j \big|\text{Pa}_j(G)\big| \le d_{\text{in}}, \text{ and } \max_j \big|\text{Ch}_j(G)\big| \le d_{\text{out}} \Big\}.$$

Since all Markov equivalent DAGs have the same skeleton, the two constraints ensure that the degree of any DAG $G' \in [G]$ for some $G \in \mathcal{G}_p(d_{\text{in}}, d_{\text{out}})$ is at most $d_{\text{in}} + d_{\text{out}}$. One may also use a single constraint for the maximum degree, but for the theoretical analysis to be carried out in this paper, it is more convenient to specify d_{in}, d_{out} separately. This setup is appealing to practitioners, since a DAG model with bounded degree is easier to visualize and interpret. Let $\mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$ denote the space of "sparse equivalence classes" defined by

$$\mathcal{C}_p(d_{\text{in}}, d_{\text{out}}) = \big\{ [G] : G \in \mathcal{G}_p(d_{\text{in}}, d_{\text{out}}) \big\}.$$

Hence, $\mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$ is the set of all equivalence classes that contain at least one member in $\mathcal{G}_p(d_{\text{in}}, d_{\text{out}})$. We will use $\mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$ as the model space for the sparse structure learning problem. The unrestricted space is denoted by $\mathcal{C}_p = \mathcal{C}_p(p, p)$.

Recall that $\mathbb{S}^p$ is the space of all permutations of $[p]$. For each $\sigma \in \mathbb{S}^p$, let

$$\begin{aligned} \mathcal{G}_p^\sigma &= \{ G \in \mathcal{G}_p : \sigma \text{ is a topological ordering of } G \} \\ &= \big\{ G \in \mathcal{G}_p : \sigma(j) \to \sigma(i) \notin G \text{ for any } i < j \big\}. \end{aligned}$$

Note a DAG may have multiple orderings; in particular, the empty DAG belongs to $\mathcal{G}_p^\sigma$ for any $\sigma \in \mathbb{S}^p$. Let $\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}}) = \mathcal{G}_p^\sigma \cap \mathcal{G}_p(d_{\text{in}}, d_{\text{out}})$ denote the space of sparse DAG models with ordering σ, which is the space we consider for the sparse DAG selection problem.

For our target posterior probability distributions, we assume they can be expressed by using a Bayesian scoring criterion $\psi : \mathcal{G}_p \to \mathbb{R}$ such that $\psi(G) = \psi(G')$ for any Markov equivalent G and G', a property known as "score equivalence" [11]. For an equivalence class $\mathcal{E}$, define $\psi(\mathcal{E}) = \psi(G)$ using any $G \in \mathcal{E}$. Let the unnormalized posterior probability of a DAG G be given by $e^{\psi(G)}$, and that of an equivalence class $\mathcal{E}$ be given by $e^{\psi(\mathcal{E})}$ (see Section 4.1 for more details). We further assume that ψ is decomposable; for each G,

$$\psi(G) = \sum_{j \in [p]} \psi_j(\mathrm{Pa}_j(G)),$$

where for each j, $\psi_j : 2^{[p]} \to \mathbb{R}$ gives the local score at node j.

3.3. *Neighborhood functions and the RW-GES sampler.* We define our neighborhood function on $\mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$ by considering operations on all member DAGs of each equivalence class. To this end, we first define three neighborhoods on the unrestricted space $\mathcal{G}_p$ for each DAG G, which correspond to three types of edge modification: addition, deletion and swap:

$$\begin{aligned}
\mathcal{N}_{\text{add}}(G) &= \{G' \in \mathcal{G}_p : G' = G \cup \{i \to j\} \text{ for some } i \to j \notin G\},\\
\mathcal{N}_{\text{del}}(G) &= \{G' \in \mathcal{G}_p : G' = G \setminus \{i \to j\} \text{ for some } i \to j \in G\},\\
\mathcal{N}_{\text{swap}}(G) &= \{G' \in \mathcal{G}_p : G' = (G \cup \{k \to j\}) \setminus \{\ell \to j\} \text{ for some } k \to j \notin G, \ell \to j \in G\}.
\end{aligned}$$

Note that a swap move consists of adding an incoming edge and deleting one at the same node, which is a straightforward extension of the swap proposal used in variable selection problems. Define the "add-delete-swap neighborhood" of G by

$$\mathcal{N}_{\text{ads}}(G) = \mathcal{N}_{\text{add}}(G) \cup \mathcal{N}_{\text{del}}(G) \cup \mathcal{N}_{\text{swap}}(G). \tag{3}$$

For each equivalence class $\mathcal{E} \in \mathcal{C}_p$, define

$$\mathcal{N}_{\text{ads}}(\mathcal{E}) = \{[G'] : G' \in \mathcal{N}_{\text{ads}}(G) \text{ for some } G \in \mathcal{E}\}, \tag{4}$$

and define the sets $\mathcal{N}_{\text{add}}(\mathcal{E})$, $\mathcal{N}_{\text{del}}(\mathcal{E})$ and $\mathcal{N}_{\text{swap}}(\mathcal{E})$ analogously; for example, $\mathcal{E}' \in \mathcal{N}_{\text{add}}(\mathcal{E})$ if and only if there exist $G \in \mathcal{E}$ and $G' \in \mathcal{E}'$ such that $G' \in \mathcal{N}_{\text{add}}(G)$. (The neighborhood notation is overloaded here, but the meaning should be clear from the argument.) Clearly, $\mathcal{N}_{\text{ads}}(\mathcal{E}) = \mathcal{N}_{\text{add}}(\mathcal{E}) \cup \mathcal{N}_{\text{del}}(\mathcal{E}) \cup \mathcal{N}_{\text{swap}}(\mathcal{E})$, and $\mathcal{N}_{\text{ads}}$ is symmetric on both $\mathcal{G}_p$ and $\mathcal{C}_p$. The following lemma gives a bound on the size of $\mathcal{N}_{\text{ads}}(\mathcal{E})$, which is needed later when we verify part (i) of Condition 1.

LEMMA 1. *For any $\mathcal{E} \in \mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$,*

$$|\mathcal{N}_{\text{ads}}(\mathcal{E}) \cap \mathcal{C}_p(d_{\text{in}}, d_{\text{out}})| \le 3p(p-1)(d_{\text{in}} + d_{\text{out}})2^{d_{\text{in}}+d_{\text{out}}}.$$

PROOF. See Supplementary Material Section D.1. □

As explained in Section 2.1, we can construct a random walk MH algorithm on the restricted space $\mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$ using $\mathcal{N}_{\text{ads}}$. The proposal distribution is given by $\mathbf{K}(\mathcal{E}, \mathcal{E}') = 1/|\mathcal{N}_{\text{ads}}(\mathcal{E})|$ for each $\mathcal{E}' \in \mathcal{N}_{\text{ads}}(\mathcal{E})$, where $\mathcal{N}_{\text{ads}}(\mathcal{E})$ denotes the neighborhood on the restricted space. It should be noted that, in practice, there is no need to calculate the size of $\mathcal{N}_{\text{ads}}(\mathcal{E})$ or enumerate member DAGs in $\mathcal{E}$. States in $\mathcal{N}_{\text{ads}}(\mathcal{E})$ can be proposed very efficiently by using some local graph operators, which is explained in detail in Supplementary Material Section H.1. We call this sampler random walk GES (RW-GES), since it uses a neighborhood function similar to that of the GES algorithm [11], which is a two-stage greedy search on the space $\mathcal{C}_p$ that uses $\mathcal{N}_{\text{add}}$ in the first stage and $\mathcal{N}_{\text{del}}$ in the second. Swap moves are not used in GES, and we will use $\mathcal{N}_{\text{ges}}(\cdot) = \mathcal{N}_{\text{add}}(\cdot) \cup \mathcal{N}_{\text{del}}(\cdot)$ to denote the neighborhood relation used by GES.

3.4. *Motivating examples.* Assume the data-generating distribution is perfectly Markovian w.r.t. some DAG G^* (which henceforth is called the "true DAG") and let $\mathcal{E}^* = [G^*]$ be the true equivalence class. In the classical asymptotic regime where p is fixed and sample size n tends to infinity, Chickering [11] proved that for a large class of Bayesian scoring criteria, GES and the greedy search on $(\mathcal{C}_p, \mathcal{N}_{\text{ges}})$ are consistent. According to our discussion following Condition 1, if we fix p and let $n \to \infty$, we can mimic the consistency proof of GES and use Theorem 1 to bound the mixing time of the random walk MH algorithm on $(\mathcal{C}_p, \mathcal{N}_{\text{ges}})$. The purpose of this subsection is to use examples to illustrate the technical challenges we encounter as we try to extend this argument to the space $\mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$.

To simplify the discussion, we assume the score ψ (i.e., log-posterior) satisfies the following condition, known as local consistency [11] (which is only used for making heuristic arguments in this section). It essentially says that all CI relations encoded by G^* can be correctly identified, which we expect to happen when $n = \infty$.

CONDITION 2. If distinct DAGs G, G' satisfy $G' = G \cup \{i \to j\}$, then (i) $\psi(G) > \psi(G')$ if $i \perp\!\!\!\perp j | \text{Pa}_j(G)$ in G^*, and (ii) $\psi(G') > \psi(G)$ if $i \not\!\perp\!\!\!\perp j | \text{Pa}_j(G)$ in G^*.

Under Condition 2, GES is consistent [11] and no equivalence class other than $[G^*]$ can be a local mode on $(\mathcal{C}_p, \mathcal{N}_{\text{ges}})$ (the reason will become clear in the next subsection). However, once we introduce the degree constraint (which is necessary for proving high-dimensional consistency results), local modes can arise on the boundary of the restricted space. To illustrate this, we construct two examples below. Example 2 explains why swap moves are useful and why in the consistency proof of GES we only consider edge removals when the current equivalence class is an I-map of $\mathcal{E}^*$. Example 3 shows that for the sparse DAG selection problem with degree constraints, local modes can also arise unexpectedly.

EXAMPLE 2. Let $p = 3$ and DAGs G^*, G be given by

$$G^*: 1 \to 2 \to 3, \qquad G: 2 \leftarrow 1 \to 3.$$

Consider how to increase the score of G by single-edge addition or deletion under Condition 2. Since $1 \not\!\perp\!\!\!\perp 2$ and $1 \not\!\perp\!\!\!\perp 3$ in G^*, both edges cannot be removed. However, since $2 \not\!\perp\!\!\!\perp 3 | 1$ in G^*, we can add the edge $2 \to 3$ to G to increase the score. The complete DAG $G \cup \{2 \to 3\}$ is an I-map of G^*, from which we should be able to remove the edge $1 \to 3$ since $1 \perp\!\!\!\perp 3 | 2$. One can apply the same argument to any other DAG in $\mathcal{E} = [G]$ and conclude that $\psi(\mathcal{E}) > \psi(\mathcal{E}')$ for any $\mathcal{E}' \in \mathcal{N}_{\text{del}}(\mathcal{E})$. In particular, we cannot remove the edge between nodes 1, 3 from any $G \in \mathcal{E}$, though the two nodes are not connected in G^*.

Next, we impose the constraint $d_{\text{in}} = 1$. Since G has two edges, we have $\mathcal{N}_{\text{add}}(\mathcal{E}) = \{\tilde{\mathcal{E}}\}$, where $\tilde{\mathcal{E}}$ is the equivalence class of all complete DAGs. But any complete DAG has maximum in-degree 2, which means that moving from $\mathcal{E}$ to $\tilde{\mathcal{E}}$ is forbidden and $\mathcal{E}$ is a local mode on $(\mathcal{C}_p(d_{\text{in}} = 1, d_{\text{out}} = p), \mathcal{N}_{\text{ges}})$. However, a swap move allows us to directly move from G to G^* by removing $1 \to 3$ and adding $2 \to 3$ simultaneously; that is, $\mathcal{E}$ is not a local mode on $(\mathcal{C}_p(1, p), \mathcal{N}_{\text{ads}})$ where $\mathcal{N}_{\text{ads}}$ is given by (4).

EXAMPLE 3. Consider the DAG selection problem with $p = 5$ and $\sigma = (1, 2, 3, 4, 5)$. Let G^*, G be DAGs in $\mathcal{G}_p^\sigma$ with edge sets

$$G^*: \{(1,2), (1,3), (2,4), (2,5)\}, \qquad G: \{(1,2), (1,4), (2,3), (2,5)\}.$$

Under Condition 2, we can increase the score of G by adding $1 \to 3$ or $2 \to 4$, but deleting $1 \to 4$ or $2 \to 3$ will lower the score since $1 \not\!\perp\!\!\!\perp 4$ and $2 \not\!\perp\!\!\!\perp 3$ in G^*. Now let $d_{\text{in}} = d_{\text{out}} = 2$. Though $G^*, G \in \mathcal{G}_p^\sigma(2, 2)$, G is a local mode on $(\mathcal{G}_p^\sigma(2, 2), \mathcal{N}_{\text{ads}})$ because adding either $1 \to 3$ or $2 \to 4$ violates the out-degree constraint (note swap moves may not be helpful either).

3.5. *Overview of the canonical path construction.* Let the true DAG model $G^* \in \mathcal{G}_p(d_{\mathrm{in}}, d_{\mathrm{out}})$ and let $\mathcal{E}^* = [G^*]$. To verify Condition 1 for the triple $(\mathcal{C}_p(d_{\mathrm{in}}, d_{\mathrm{out}}), \mathcal{N}_{\mathrm{ads}}, e^{\psi})$, we need to show that for any $\mathcal{E} \neq \mathcal{E}^*$, we can identify some $g(\mathcal{E}) \in \mathcal{N}_{\mathrm{ads}}(\mathcal{E})$ such that $\psi(g(\mathcal{E})) > \psi(\mathcal{E})$. By Remark 2, this is equivalent to constructing a canonical path from any $\mathcal{E}$ to $\mathcal{E}^*$. We briefly discuss the main idea behind our construction in this subsection. It will be helpful to think of the space $\mathcal{C}_p(d_{\mathrm{in}}, d_{\mathrm{out}})$ as the union of $\{\mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}}) : \sigma \in \mathbb{S}^p\}$ and think of structure learning as simultaneous DAG selection for all $p!$ orderings.

Suppose RW-GES starts at some $\mathcal{E}$, which contains a member DAG $G \in \mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}})$ for an arbitrary $\sigma \in \mathbb{S}^p$. We will first construct a canonical path on $\mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}})$, denoted by $(G_0 = G, G_1, G_2, \ldots, G_k)$, where the terminal state G_k (if possible) is given by

$$\hat{G}(\sigma) = \underset{G \in \mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}})}{\arg\max} \psi(G). \tag{5}$$

If Condition 2 holds and $G_{\sigma}^* \in \mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}})$, we claim $\hat{G}(\sigma) = G_{\sigma}^*$, where we recall G_{σ}^* is the minimal I-map of G^* with ordering σ. To show this, without loss of generality, assume $\sigma = (1, 2, \ldots, p)$, and note that the following CI relations hold in G^* for each $j \in [p]$ by the definition of minimal I-maps (see Section 3.1):

$$j \perp\!\!\!\perp [j-1] \setminus \mathrm{Pa}_j(G_{\sigma}^*) | \mathrm{Pa}_j(G_{\sigma}^*), \quad \text{and} \quad j \not\perp\!\!\!\perp i | [j-1] \setminus \{i\} \quad \text{for each } i \in \mathrm{Pa}_j(G_{\sigma}^*). \tag{6}$$

Under Condition 2, the first property in (6) implies that if G is a DAG such that $\mathrm{Pa}_j(G_{\sigma}^*) \subsetneq \mathrm{Pa}_j(G)$, we can increase the score of G by removing some edge $\ell \to j$, and the second implies that if $\mathrm{Pa}_j(G_{\sigma}^*) \not\subseteq \mathrm{Pa}_j(G)$, we can add some edge $k \to j$ or perform a swap. This shows $\hat{G}(\sigma) = G_{\sigma}^*$ and suggests how we can construct the path from G to G_{σ}^*. However, as discussed in the previous subsection, the main challenge is to deal with the degree constraints.

Now suppose that RW-GES can move from $\mathcal{E}$ to $[G_{\sigma}^*]$ following the path $(\mathcal{E}, \mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k)$ where $\mathcal{E}_i = [G_i]$. If $[G_{\sigma}^*] = \mathcal{E}^*$ (i.e., $G_{\sigma}^* = G^*$ or G_{σ}^* is Markov equivalent to G^*), we have obtained the path from $\mathcal{E}$ from $\mathcal{E}^*$. If $[G_{\sigma}^*] \neq \mathcal{E}^*$, then one can use the famous Chickering algorithm [11, 39] to construct a path from $[G_{\sigma}^*]$ to $\mathcal{E}^*$ (see Lemma D3 in Supplementary Material Section D.4). Intuitively, since G_{σ}^* is an I-map of G^*, the skeleton of G^* must be a subset of the skeleton of G_{σ}^* (see Lemma C3), and we can remove edges from some other member DAG of $[G_{\sigma}^*]$.

Unfortunately, to rigorously prove that $\hat{G}(\sigma) = G_{\sigma}^*$ for all $\sigma \in \mathbb{S}^p$ in high-dimensional settings, one often needs to impose restrictive assumptions on the true data-generating mechanism, such as strong faithfulness [42]. To our knowledge, there is no fully satisfactory solution to this issue, and we will make a similar assumption in our theoretical analysis in Section 4 and assume $\hat{G}(\sigma) = G_{\sigma}^*$ in this section. Nevertheless, we will construct canonical paths of RW-GES using a flexible and finer argument, which in some cases, can be used to show the rapid mixing of RW-GES under weaker assumptions; see Supplementary Material Section I.

3.6. *Canonical add-delete-swap paths of RW-GES.* The discussion above suggests that we can construct the canonical paths of RW-GES by first constructing the canonical paths for all DAG selection problems. To this end, fix an arbitrary $\sigma \in \mathbb{S}^p$ first, and consider the sparse DAG selection problem with state space $\mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}})$, neighborhood function $\mathcal{N}_{\mathrm{ads}}$ and posterior e^{ψ}. We treat G_{σ}^* as the true model, and we need to construct a candidate canonical transition function for this problem, $g^{\sigma} : \mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}}) \to \mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}})$, such that for any $G \in \mathcal{G}_p^{\sigma}(d_{\mathrm{in}}, d_{\mathrm{out}})$,

$$g^{\sigma}(G) \in \mathcal{N}_{\mathrm{ads}}(G), \quad \text{and} \quad (g^{\sigma})^k(G) = G_{\sigma}^* \quad \text{for some } k < \infty.$$

For Condition 1 to hold, we also need $\psi(g^\sigma(G)) > \psi(G)$. To overcome the out-degree constraint issue illustrated by Example 3, we will construct $g^\sigma(G)$ by first analyzing each node separately. Observe that if there is no out-degree constraint, the DAG selection problem is equivalent to p variable selection problems: for each j, we need to estimate the set Pa_j, which takes value in the space $\mathcal{M}_p^\sigma(j, d_{\mathrm{in}})$ defined by

$$\mathcal{M}_p^\sigma(j, d_{\mathrm{in}}) = \{S \subseteq \mathcal{A}_p^\sigma(j) \colon |S| \le d_{\mathrm{in}}\}, \qquad \mathcal{A}_p^\sigma(j) = \{k \in [p] \colon \sigma^{-1}(k) < \sigma^{-1}(j)\}, \tag{7}$$

where $\mathcal{A}_p^\sigma(j)$ is the set of variables that precede X_j in the ordering σ. Motivated by the discussion following (6), we construct a transition function on the space $\mathcal{M}_p^\sigma(j, d_{\mathrm{in}})$ in Definition 4, which gives the "optimal" add-delete-swap move for Pa_j. Recall that we assume $\psi(G) = \sum_j \psi_j(\mathrm{Pa}_j(G))$ for each G.

DEFINITION 4. Assume $G_\sigma^* \in \mathcal{G}_p^\sigma(d_{\mathrm{in}}, d_{\mathrm{out}})$ and let $S_{\sigma,j}^* = \mathrm{Pa}_j(G_\sigma^*)$. For each j, we construct $g_j^\sigma \colon \mathcal{M}_p^\sigma(j, d_{\mathrm{in}}) \to \mathcal{M}_p^\sigma(j, d_{\mathrm{in}})$ as follows. Fix an arbitrary $S \in \mathcal{M}_p^\sigma(j, d_{\mathrm{in}})$, and let $T = S_{\sigma,j}^* \setminus S$ and $R = S \setminus S_{\sigma,j}^*$.

(i) If $S = S_{\sigma,j}^*$, let $g_j^\sigma(S) = S_{\sigma,j}^*$.
(ii) If $S_{\sigma,j}^* \subset S$, let $g_j^\sigma(S) = S \setminus \{\tilde{\ell}\}$ where $\tilde{\ell} = \arg\max_{\ell \in R} \psi_j(S \setminus \{\ell\})$.
(iii) If $S_{\sigma,j}^* \nsubseteq S$ and $|S| < d_{\mathrm{in}}$, let $g_j^\sigma(S) = S \cup \{\tilde{k}\}$ where $\tilde{k} = \arg\max_{k \in T} \psi_j(S \cup \{k\})$.
(iv) If $S_{\sigma,j}^* \nsubseteq S$ and $|S| = d_{\mathrm{in}}$, let $g_j^\sigma(S) = (S \cup \{\tilde{k}\}) \setminus \{\tilde{\ell}\}$ where $(\tilde{k}, \tilde{\ell}) = \arg\max_{(k,\ell) \in T \times R} \psi_j((S \cup \{k\}) \setminus \{\ell\})$.

In case (ii), we say node j is (strictly) overfitted; in cases (iii) and (iv), we say it is underfitted. We use $g_j^\sigma(G)$ to denote the DAG obtained by replacing the parent set of j in G with $g_j^\sigma(\mathrm{Pa}_j(G))$; that is, $\mathrm{Pa}_j(g_j^\sigma(G)) = g_j^\sigma(\mathrm{Pa}_j(G))$, and for any $i \neq j$, $\mathrm{Pa}_i(g_j^\sigma(G)) = \mathrm{Pa}_i(G)$.

REMARK 6. It is clear from definition that $d_{\mathrm{H}}(g_j^\sigma(S), S_{\sigma,j}^*) < d_{\mathrm{H}}(S, S_{\sigma,j}^*)$ if $S \neq S_{\sigma,j}^*$. Further, $g_j^\sigma(G) \in \mathcal{N}_{\mathrm{ads}}(G)$ and $d_{\mathrm{H}}(g_j^\sigma(G), G_\sigma^*) < d_{\mathrm{H}}(G, G_\sigma^*)$ if $\mathrm{Pa}_j(G) \neq \mathrm{Pa}_j(G_\sigma^*)$. In words, if node j is overfitted in G, $g_j^\sigma(G)$ is obtained by removing an incoming edge of node j. If node j is underfitted, $g_j^\sigma(G)$ is obtained by adding an incoming edge of node j (if the in-degree constraint is violated, remove another incoming edge of node j). An example is provided in Figure 1. Note that this rationale is similar to that for GES and forward–backward stepwise regression. We always first transform an underfitted model to overfitted and then remove redundant variables or edges (recall Example 2).

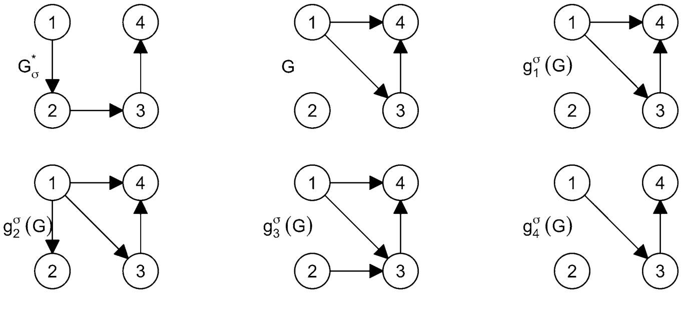

FIG. 1. *An example for the operator g_j^σ. We consider four nodes with ordering $\sigma = (1, 2, 3, 4)$; assume $d_{\mathrm{in}} = 3$. G_σ^* has three edges, $1 \to 2$, $2 \to 3$ and $3 \to 4$. Consider another DAG G with edges $1 \to 3$, $1 \to 4$ and $3 \to 4$. The DAGs $g_1^\sigma(G)$, $g_2^\sigma(G)$, $g_3^\sigma(G)$, $g_4^\sigma(G)$ are shown above. For example, since $\mathrm{Pa}_4(G_\sigma^*) = \{3\} \subset \mathrm{Pa}_4(G) = \{1, 3\}$, node 4 is overfitted in G, and by part (ii) of Definition 4, $g_4^\sigma(G)$ is obtained by removing the edge $1 \to 4$ from G.*

REMARK 7. Consider the variable selection problem with model space $\mathcal{M}_p^\sigma(j, d_{\text{in}})$ and true model $S^*_{\sigma,j}$. Yang, Wainwright and Jordan [65] proved that, under very mild high-dimensional assumptions, g_j^σ satisfies Condition 1 with high probability; that is,

$$\psi_j\big(g_j^\sigma(S)\big) - \psi_j(S) \geq t \log p \quad \forall S \in \mathcal{M}_p^\sigma(j, d_{\text{in}}) \setminus \big\{\text{Pa}_j\big(G^*_\sigma\big)\big\}, \tag{8}$$

for some $t > 0$ (in their conclusion t is a universal constant).

Suppose that (8) holds for each j. Then, to show that the triple $(\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}}), \mathcal{N}_{\text{ads}}, e^\psi)$ satisfies part (ii) of Condition 1, we only need to use the operators $\{g_j^\sigma : j \in [p]\}$ to construct a path from any $G \in \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$ to G^*_σ. At first glance, this seems trivial since we can use g_j^σ repeatedly to convert any $\text{Pa}_j(G)$ to $\text{Pa}_j(G^*_\sigma)$. However, the definition of g_j^σ only guarantees that $g_j^\sigma(G) \in \mathcal{G}_p^\sigma(d_{\text{in}}, p)$, but the maximum out-degree of $g_j^\sigma(G)$ can be larger than that of G. Indeed, Example 3 in Section 3.4 shows that, in extreme cases, none of the operators $g_1^\sigma, \dots, g_p^\sigma$ yields a DAG that is different from G and belongs to $\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$. Fortunately, we are able to prove that, as long as d_{out} is chosen sufficiently large, there always exists some j such that g_j^σ yields a different DAG in $\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$. We define

$$d^*_\sigma = \max_{j \in [p]} \big|\text{Pa}_j\big(G^*_\sigma\big) \cup \text{Ch}_j\big(G^*_\sigma\big)\big|, \qquad d^* = \max_{\sigma \in \mathbb{S}^p} d^*_\sigma, \tag{9}$$

where d^* will be used later in Theorem 3.

LEMMA 2. *Assume $d^*_\sigma \leq d_{\text{in}}$ and $\min\{d^*_\sigma d_{\text{in}} + 1, p\} \leq d_{\text{out}}$. For any $G \in \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$ such that $G \neq G^*_\sigma$, there exists some $j \in [p]$ such that $g_j^\sigma(G) \in \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$ and $g_j^\sigma(G) \neq G$.*

PROOF. The key idea of the proof is to use the pigeonhole principle multiple times to derive the contradiction. See Supplementary Material Section D.2. □

COROLLARY 1. *Let $\sigma \in \mathbb{S}^p$. Assume that $d^*_\sigma \leq d_{\text{in}}$ and $\min\{d^*_\sigma d_{\text{in}} + 1, p\} \leq d_{\text{out}}$.*

(i) *There exists a function $g^\sigma : \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}}) \to \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$ such that for any $G \neq G^*_\sigma$, $g^\sigma(G) = g_j^\sigma(G) \neq G$ for some $j \in [p]$ and $(g^\sigma)^k(G) = G^*_\sigma$ for some $k \leq (d^*_\sigma + d_{\text{in}})p$.*

(ii) *If (8) holds for each $j \in [p]$, Condition 1 holds for the triple $(\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}}), \mathcal{N}_{\text{ads}}, e^\psi)$ with $t_1 = 3$ and $t_2 = t$.*

PROOF. See Supplementary Material Section D.3. □

We are now ready to construct a canonical transition function $g \colon \mathcal{C}_p(d_{\text{in}}, d_{\text{out}}) \to \mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$ for the structure learning problem using operators $\{g_j^\sigma : j \in [p], \sigma \in \mathbb{S}^p\}$. If $\mathcal{E}$ contains a minimal I-map of G^*, we define $g(\mathcal{E})$ using Chickering algorithm [11]; see Lemma D3 in the Supplementary Material. If not, by the definition of $\mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$, there exists $G \in \mathcal{E} \cap \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$ for some $\sigma \in \mathbb{S}^p$, and we can define $g(\mathcal{E})$ using the function g^σ constructed for the DAG selection problem. But note that we need to fix the DAG representation of each $\mathcal{E}$ so that $g(\mathcal{E})$ can be defined uniquely. We give an explicit construction of g in the proof of Theorem 3, the main result for this section.

THEOREM 3. *Assume that $d^* \leq d_{\text{in}}$ and $\min\{d^* d_{\text{in}} + 1, p\} \leq d_{\text{out}}$. Then $G^*_\sigma \in \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$ for each $\sigma \in \mathbb{S}^p$. Further, there exists a function $g \colon \mathcal{C}_p(d_{\text{in}}, d_{\text{out}}) \to \mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$ such that $g(\mathcal{E}^*) = \mathcal{E}^*$ and the following hold for any $\mathcal{E} \in \mathcal{C}_p(d_{\text{in}}, d_{\text{out}}) \setminus \{\mathcal{E}^*\}$:*

(i) *$g(\mathcal{E}) = [g_j^\sigma(G)]$ for some $j \in [p]$, $\sigma \in \mathbb{S}^p$ and $G \in \mathcal{E} \cap \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$ such that $g_j^\sigma(G) \neq G$.*

(ii) *There exist $k \leq (d^* + d_{\text{in}})p$ and $k \leq \ell \leq (2d^* + d_{\text{in}})p$ such that $g^k(\mathcal{E}) = G^*_\sigma$ for some $\sigma \in \mathbb{S}^p$ and $g^\ell(\mathcal{E}) = \mathcal{E}^*$.*

PROOF. See Supplementary Material Section D.4. □

We conclude this section with the following corollary, which shows that to establish part (ii) of Condition 1 for the sparse structure learning problem, it only remains to prove that (8) holds for all j and σ simultaneously. This will be done rigorously in the next section.

COROLLARY 2. *Assume* $d^* \leq d_{\text{in}}$, $\min\{d^* d_{\text{in}} + 1, p\} \leq d_{\text{out}}$ *and* ψ *is score equivalent so that we can define* $\psi(\mathcal{E}) = \psi(G)$ *using any* $G \in \mathcal{E}$. *If* (8) *holds for each* $\sigma \in \mathbb{S}^p$ *and each* $j \in [p]$, *part* (*ii*) *of Condition* 1 *holds for the triple* $(\mathcal{C}_p(d_{\text{in}}, d_{\text{out}}), \mathcal{N}_{\text{ads}}, e^{\psi})$ *with* $t_2 = t$.

PROOF. See Supplementary Material Section D.5. □

4. High-dimensional consistency of an empirical Bayes model for structure learning.

4.1. *Model, prior and posterior distributions.* Let X be an $n \times p$ data matrix where each row is an i.i.d. copy of a normal random vector $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_p)$. (The font for the random vector X and that for the data matrix X are different.) Assume that, given a DAG G, the distribution of X can be described by the structural equation model (SEM),

$$\mathsf{X} = B^\top \mathsf{X} + \mathsf{e}, \quad \mathsf{e} \sim N_p(0, \Omega), \tag{10}$$

for some $(B, \Omega) \in \mathcal{D}_p(G)$, where

$$\begin{aligned} \mathcal{D}_p(G) = \big\{(B, \Omega) \colon\ & B \in \mathbb{R}^{p \times p}, B_{ij} = 0 \text{ if } i \to j \notin G, \text{for any } i, j \in [p]; \\ & \Omega = \operatorname{diag}(\omega_1, \ldots, \omega_p), \omega_i > 0 \text{ for any } i \in [p]\big\}. \end{aligned} \tag{11}$$

That is, each X_j follows a linear regression model where explanatory variables with nonzero regression coefficients must be parents of node j in G. The matrix B is often called the weighted adjacency matrix. We can equivalently express (10) as

$$\mathsf{X} \sim N_p\big(0, \Sigma(B, \Omega)\big) \quad \text{where } \Sigma(B, \Omega) = \big(I - B^\top\big)^{-1} \Omega (I - B)^{-1} \tag{12}$$

is called the modified Cholesky decomposition (I denotes the identity matrix). The SEM representation of the Gaussian DAG model is used frequently in the literature [3, 12, 60].

Let $\pi_0(B, \Omega | G)$ denote the conditional prior distribution with support $\mathcal{D}_p(G)$. It suffices to specify it for $\{(\beta_j(G), \omega_j) \colon j = 1, \ldots, p\}$, where $\beta_j(G)$ is the subvector of the jth column of B with entries indexed by $\text{Pa}_j(G)$, and ω_j is the jth diagonal element of Ω. We use the empirical prior proposed by Lee, Lee and Lin [33], which is an extension of the empirical variable selection model of Martin, Mess and Walker [38]. Our prior assumes that, given G, $(\beta_1(G), \omega_1), \ldots, (\beta_p(G), \omega_p)$ are independently distributed according to

$$\begin{gathered} \pi_0(\omega_j | G) \propto \omega_j^{-\kappa/2 - 1}, \\ \beta_j(G) | \text{Pa}_j(G) = S_j, \omega_j \sim N_{|S_j|}\bigg(\big(X_{S_j}^\top X_{S_j}\big)^{-1} X_{S_j}^\top X_j, \frac{\omega_j}{\gamma} \big(X_{S_j}^\top X_{S_j}\big)^{-1}\bigg), \end{gathered}$$

where $\gamma > 0$, $\kappa \geq 0$ are hyperparameters, X_j denotes the jth column of the data matrix X and X_S is the submatrix containing columns indexed by S. Next, we compute the marginal likelihood of G by integrating out (B, Ω) and using a fractional exponent $\alpha \in (0, 1)$ to offset the overuse of data caused by the empirical prior. The resulting fractional marginal likelihood is given by $f_\alpha(G) = \prod_{j=1}^p f_{\alpha, j}(\text{Pa}_j(G))$, where

$$f_{\alpha, j}(S) = \big(1 + \alpha \gamma^{-1}\big)^{-|S|/2} \big\{X_j^\top \big(I - X_S \big(X_S^\top X_S\big)^{-1} X_S^\top\big) X_j\big\}^{-(\alpha n + \kappa)/2}.$$

More details about this empirical prior are given in Supplementary Material Section F.1.

For sparse DAG selection with ordering σ, the state space is $\mathcal{G}_p^\sigma(d_{\mathrm{in}}, d_{\mathrm{out}})$. For each G on this space, we specify its prior probability by

$$\pi_0^\sigma(G) \propto \left(c_1 p^{c_2}\right)^{-|G|}, \tag{13}$$

where $c_1 > 0$, $c_2 \geq 0$ are hyperparameters. We can then calculate the posterior distribution by $\pi_n^\sigma(G) \propto \pi_0^\sigma(G) f_\alpha(G)$. Using the fractional marginal likelihood f_α, we get

$$\pi_n^\sigma(G) \propto e^{\psi(G)} \mathbb{1}_{\mathcal{G}_p^\sigma(d_{\mathrm{in}}, d_{\mathrm{out}})}(G) \quad \text{where} \tag{14}$$

$$\psi(G) = \sum_{j=1}^{p} \psi_j\big(\mathrm{Pa}_j(G)\big), \quad \text{and} \quad e^{\psi_j(S)} = \left(c_1 p^{c_2}\right)^{-|S|} f_{\alpha,j}(S). \tag{15}$$

For the sparse structure learning problem, we use the prior

$$\pi_0(\mathcal{E}, G) \propto \left(c_1 p^{c_2}\right)^{-|G|} \pi_0(G|\mathcal{E}) \mathbb{1}_{\mathcal{E}}(G), \tag{16}$$

where $\pi_0(G|\mathcal{E})$ satisfies $\sum_{G \in \mathcal{E}} \pi_0(G|\mathcal{E}) = 1$. Denote the corresponding posterior distribution by π_n. Marginalizing out G from $\pi_n(\mathcal{E}, G)$, we get

$$\pi_n(\mathcal{E}) \propto \sum_{G \in \mathcal{E}} \pi_0(G|\mathcal{E}) e^{\psi(G)}.$$

In Lemma 3 below, we prove that ψ yields the same value for any Markov equivalent DAGs. Hence, we can define $\psi(\mathcal{E}) = \psi(G)$ using any $G \in \mathcal{E}$, and $\pi_n(\mathcal{E})$ can be expressed by

$$\pi_n(\mathcal{E}) \propto e^{\psi(\mathcal{E})} \mathbb{1}_{\mathcal{C}_p(d_{\mathrm{in}}, d_{\mathrm{out}})}(\mathcal{E}). \tag{17}$$

The indicator function in (17) serves to remind us of the restricted search space. We do not consider estimating the DAG or ordering from π_n. Indeed, for $G \in \mathcal{E}$, $\pi_n(G)$ depends on the conditional prior probability $\pi_0(G|\mathcal{E})$, which we leave unspecified.

LEMMA 3. *The function ψ defined by* (15) *satisfies that $\psi(G) = \psi(G')$ whenever G and G' are Markov equivalent DAGs.*

PROOF. See Supplementary Material Section F.2. □

We will refer to $\psi_j(\mathrm{Pa}_j)$, $\psi(G)$, $\psi(\mathcal{E})$ as the scores of Pa_j, G and $\mathcal{E}$, respectively. Note that a scoring criterion derived from a nodewise normal-inverse-gamma prior for $(B, \Omega)|G$ does not necessarily have the property given in Lemma 3. For nonempirical prior distributions, see Geiger and Heckerman [19] and Peluso and Consonni [47] for related results.

4.2. *High-dimensional setup.* Let G^* denote the true DAG model and $\mathcal{E}^* = [G^*]$ be the true equivalence class that we want to recover from the data. Assume that each row of X is drawn independently from $N_p(0, \Sigma^*)$, a normal distribution perfectly Markovian w.r.t. G^*. We will show π_n defined in (17) concentrates on $[G^*]$ by first proving that for each σ, π_n^σ defined in (14) concentrates on the minimal I-map G_σ^*. Due to normality, G_σ^* can be equivalently defined by using the modified Cholesky decomposition.

DEFINITION 5. Let Σ^* be positive definite and $N_p(0, \Sigma^*)$ be perfectly Markovian w.r.t. some DAG G^*. For each $\sigma \in \mathbb{S}^p$, let $\mathcal{D}_p(\sigma) = \bigcup_{G \in \mathcal{G}_p^\sigma} \mathcal{D}_p(G)$. By Lemma C6, we can define $(B_\sigma^*, \Omega_\sigma^*)$ to be the unique pair in $\mathcal{D}_p(\sigma)$ such that

$$\left(I - (B_\sigma^*)^\top\right)^{-1} \Omega_\sigma^* \left(I - B_\sigma^*\right)^{-1} = \Sigma^*.$$

Define G_σ^* to be the DAG such that $i \to j \in G_\sigma^*$ if and only if $(B_\sigma^*)_{ij} \neq 0$, which by Lemma C5, is the minimal I-map of G^* with ordering σ.

Consider a high-dimensional setting with $p = p(n)$ tending to infinity. The true DAG model G^*, true covariance matrix Σ^* and prior parameters c_1, c_2, α, γ, d_{in}, d_{out} are all implicitly indexed by n. We say a constant is universal if it does not depend on n. To derive our consistency results, we need to make a few assumptions on the parameters and Σ^*.

(A1) There exist $\underline{\nu} = \underline{\nu}(n)$, $\overline{\nu} = \overline{\nu}(n) > 0$ and a universal constant $\delta_0 > 0$ such that

$$0 < \frac{\underline{\nu}}{(1-\delta_0)^2} \le \lambda_{\min}(\Sigma^*) \le \lambda_{\max}(\Sigma^*) \le \frac{\overline{\nu}}{(1+\delta_0)^2},$$

where $\lambda_{\min}$, $\lambda_{\max}$ denote the smallest and largest eigenvalues, respectively.

(A2) The sparsity parameter d_{in} and n, p satisfy that $d_{\text{in}} \log p = o(n)$.

(A3) Prior parameters satisfy that $\kappa \le n$, $1 \le c_1\sqrt{1+\alpha/\gamma} \le p$ and

$$c_2 \ge (\alpha+1)(4d_{\text{in}}+6)+t$$

for some universal constant $t > 0$.

(A4) Assumption on the maximum in-degree of G^*_σ.

(A4.1) Let $\nu_0 = 4\overline{\nu}^2\underline{\nu}^{-4}(\overline{\nu}-\underline{\nu})^2$. For some $\sigma \in \mathbb{S}^p$, $(\nu_0+1)\max_{j\in[p]} |\mathrm{Pa}_j(G^*_\sigma)| \le d_{\text{in}}$.

(A4.2) Assumption (A4.1) holds for every $\sigma \in \mathbb{S}^p$.

(A5) Assumption on B^*_σ, B^*.

(A5.1) There exists a universal constant $C_\beta > 0$ such that for some $\sigma \in \mathbb{S}^p$,

$$\min\{|(B^*_\sigma)_{ij}|^2 : (B^*_\sigma)_{ij} \ne 0\} \ge 5(C_\beta + 4c_2)\frac{\overline{\nu}^2 \log p}{\alpha\underline{\nu}^2 n}, \tag{18}$$

where B^*_σ is given by Definition 5.

(A5.2) There exists a universal constant $C_\beta > 0$ such that (18) holds for every $\sigma \in \mathbb{S}^p$.

The first three assumptions are standard and commonly used in high-dimensional statistical theory. Assumption (A1) is the standard restricted eigenvalue condition [6]. Assumption (A2) controls the growth rates of p and d_{in} (which determines the maximum model size for nodewise variable selection), and together with Assumption (A3), ensures that we cannot overfit the data; recall from (16) that the hyperparameter c_2 controls the penalty on the model size, so it plays the same role as the tuning parameter in the penalized likelihood methods. Such assumptions (especially a condition similar to $d_{\text{in}} \log p = o(n)$) are required for most high-dimensional problems including variable selection [25, 63–65], stochastic block model [18], covariance matrix estimation [31, 45, 57], undirected Gaussian graphical models [5, 34, 50] and DAG selection [7, 33]; see Banerjee, Castillo and Ghosal [4] for a recent review. Note that the numerical constants in our assumptions are very conservative. For example, Assumption (A3) suggests that c_2 should grow linearly with d_{in}, but in practice, one can use some c_2 much smaller than $4d_{\text{in}}$, which we will illustrate using a simulation study in Section 6.3.

Assumption (A4.1) requires that the maximum in-degree of the "true model" for DAG selection with ordering σ is sufficiently small compared with d_{in}. It is similar to Assumption D of Yang, Wainwright and Jordan [65] and is technically needed to show that an MH sampler using add-delete-swap moves cannot get stuck at DAG models with maximum in-degree equal to d_{in}. But unlike their setup, we assume both lower and upper restricted eigenvalues are available, which enables us to avoid imposing an irrepresentability condition as in Yang, Wainwright and Jordan [65] (see their Assumption D). Assumption (A4.2) restricts the maximum in-degree of all minimal I-maps of G^*, which is allowed to have the same order as d_{in}, if $\overline{\nu}$, $\underline{\nu}$ defined in Assumption (A1) can be bounded by universal constants.

Assumption (A5.1) is the well-known beta-min condition for DAG selection with ordering σ [7, 33]. According to Definition 5, the SEM representation (10) holds for $(B, \Omega) =$

$(B_\sigma^*, \Omega_\sigma^*)$. Hence, Assumption (A5.1) just means that all nonzero regression coefficients (i.e., signal sizes) of the true SEM with ordering σ are sufficiently large. Assumption (A5.2) is for structure learning and assumes the beta-min condition holds uniformly over all $\sigma \in \mathbb{S}^p$; this is often known as the strong beta-min or permutation beta-min condition [59] and was used in Van de Geer and Bühlmann [60] and Aragam, Amini and Zhou [3]. If p and Σ^* are fixed, which implies B_σ^* is fixed for all $\sigma \in \mathbb{S}^p$, then Assumption (A5.2) can always be satisfied by choosing some large n. We need the strong beta-min condition (or some similar assumption) since we want to first establish that with high probability, for every $\sigma \in \mathbb{S}^p$, the minimal I-map G_σ^* has the highest score among all DAGs in $\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$, which is needed for proving consistency results for structure learning. For methods based on CI tests, a similar assumption, known as "strong faithfulness," is commonly used [42] (strong beta-min condition essentially replaces partial correlations in strong faithfulness with partial regression coefficients). Uhler et al. [59] showed that the volume of normal distributions that are strongly faithful is very small. Though strong faithfulness and strong beta-min condition are not directly comparable, both seem to be fairly restrictive [60]. Unfortunately, without them, we cannot preclude the possibility that GES or local MH algorithms get trapped at local modes; see Example 4 in Section 5.1. A discussion on how to overcome such limitations is given in Supplementary Material Section I. We end this subsection with one more remark on Assumption (A1).

REMARK 8. The restricted eigenvalue condition can be used to obtain some useful bounds related to B_σ^* and Ω_σ^*. Write $\Omega_\sigma^* = \operatorname{diag}(\omega_{\sigma,1}^*, \dots, \omega_{\sigma,p}^*)$. The decomposition (12) implies that $\omega_{\sigma,k}^* \in (\underline{\nu}, \overline{\nu})$ for any $\sigma \in \mathbb{S}^p$ and $k \in [p]$ since the diagonal elements of Σ^* and $(\Sigma^*)^{-1}$ can be bounded by the extreme eigenvalues of Σ^*. Further, we can bound the ℓ^2-norm of the true regression coefficients for node j by $\sum_{i \in [p]} (B_\sigma^*)_{ij}^2 \le \omega_{\sigma,j}^* / \underline{\nu} - 1$, using the fact that the operator norm is no less than the ℓ^2-norm of any column.

4.3. *Strong selection consistency results.*

For a general model selection problem, we say a Bayesian procedure has strong selection consistency if the posterior probability of the true model converges to 1 in probability with respect to the true data-generating probability measure [7, 26, 43]. By part (ii) of Theorem 1, to prove the strong selection consistency, we only need to show that Condition 1 is satisfied for some universal $t_2 > t_1$.

We begin with the strong selection consistency for nodewise variable selection and DAG selection problems. It turns out that we only need (8) holds for any $j \in [p]$ and $\sigma \in \mathbb{S}^p$. By Corollary 2, this consistency property of $\{g_j^\sigma : j \in [p], \sigma \in \mathbb{S}^p\}$ is also key to the verification of Condition 1 for structure learning. The complete proof for Theorem 4 is highly technical, and the most involved step is to establish an analogous consistency result for a single variable selection problem using our empirical prior, which is treated in detail in Supplementary Material Section E and may be of independent interest.

THEOREM 4. *Let $X \in \mathbb{R}^{n \times p}$ have i.i.d. rows drawn from $N_p(0, \Sigma^*)$, which is perfectly Markovian w.r.t. G^*. Suppose Assumptions (A1), (A2), (A3), (A4.2) and (A5.2) hold. Let $t > 0$ be the universal constant given in Assumption (A3) and assume $C_\beta \ge 8t/3$. For sufficiently large n, with probability at least $1 - 3p^{-1}$, the following statements hold:*

(i) *Consistency of the operators $\{g_j^\sigma : j \in [p], \sigma \in \mathbb{S}^p\}$ given in Definition 4:*

$$\min\{\psi_j(g_j^\sigma(S)) - \psi_j(S) : \sigma \in \mathbb{S}^p, j \in [p], S \in \mathcal{M}_p^\sigma(j, d_{\text{in}}) \setminus \{S_{\sigma,j}^*\}\} \ge t \log p,$$

where ψ_j is given in (15) and $S_{\sigma,j}^* = \operatorname{Pa}_j(G_\sigma^*)$.

(ii) *If* $t > 2$, *we have the strong selection consistency of nodewise variable selection*,

$$\min_{\sigma \in \mathbb{S}^p} \min_{j \in [p]} \frac{\exp(\psi_j(S^*_{\sigma,j}))}{\sum_{S \in \mathcal{M}^\sigma_p(j,d_{\text{in}})} \exp(\psi_j(S))} \geq 1 - p^{-(t-2)},$$

where $\mathcal{M}^\sigma_p(j, d_{\text{in}})$ *is defined in* (7).

(iii) *If* $t > 3$, *we have the strong selection consistency of sparse DAG selection*,

$$\min_{\sigma \in \mathbb{S}^p} \frac{\exp(\psi(G^*_\sigma))}{\sum_{G \in \mathcal{G}^\sigma_p(d_{\text{in}},d_{\text{out}})} \exp(\psi(G))} \geq 1 - p^{-(t-3)},$$

where $\psi(G)$ *is defined in* (15).

PROOF. See Supplementary Material Section F.3. □

REMARK 9. The universal constant t can be chosen arbitrarily large. Given any $t > 0$, in order that Theorem 4 holds, we can always choose some c_2 that has same order as d_{in} and assume that the universal constant C_β in Assumption (A5.2) is sufficiently large.

As a corollary, the strong selection consistency for a single DAG selection problem with ordering σ can be obtained by replacing Assumptions (A4.2) and (A5.2) with Assumptions (A4.1) and (A5.1). This result was also proved in Lee, Lee and Lin [33] under similar assumptions, but the method we use is different (the primary goal of Lee, Lee and Lin [33] was to derive minimax posterior convergence rates for the weighted adjacency matrix). Note that if σ is an ordering of G^*, then $G^*_\sigma = G^*$.

COROLLARY 3. *Let* $X \in \mathbb{R}^{n \times p}$ *have i.i.d. rows drawn from the distribution* $N_p(0, \Sigma^*)$, *which is perfectly Markovian w.r.t.* G^*. *Suppose Assumptions* (A1), (A2), (A3), (A4.1) *and* (A5.1) *hold for some* $t > 3$ *and* $C_\beta \geq 8t/3$. *Let* σ *be as given in Assumptions* (A4.1) *and* (A5.1). *For sufficiently large* n, *with probability at least* $1 - 3p^{-1}$,

$$\frac{\exp(\psi(G^*_\sigma))}{\sum_{G \in \mathcal{G}^\sigma_p(d_{\text{in}},d_{\text{out}})} \exp(\psi(G))} \geq 1 - p^{-(t-3)}.$$

PROOF. The proof is wholly analogous to that for Theorem 4. □

In order to show Condition 1 holds and use Theorem 1 to prove the strong selection consistency of sparse structure learning, it only remains to invoke Lemma 1 to bound the size of $\mathcal{N}_{\text{ads}}(\cdot)$ and then apply Corollary 2. Recall the definition of d^*_σ and d^* given in (9).

THEOREM 5. *Let* $X \in \mathbb{R}^{n \times p}$ *have i.i.d. rows drawn from* $N_p(0, \Sigma^*)$, *which is perfectly Markovian w.r.t.* G^*. *Suppose* $d^* \leq d_{\text{in}}$, $d^* d_{\text{in}} + 1 \leq d_{\text{out}}$ *and* $d_{\text{in}} + d_{\text{out}} \leq t_0 \log_2 p$ *for some universal constant* $t_0 > 0$, *and Assumptions* (A1), (A2), (A3), (A4.2) *and* (A5.2) *hold with* $C_\beta \geq 8t/3$ *and* $t > t_0 + 3$. *For sufficiently large* n, *with probability at least* $1 - 3p^{-1}$,

$$\frac{\exp(\psi(\mathcal{E}^*))}{\sum_{\mathcal{E} \in \mathcal{C}_p(d_{\text{in}},d_{\text{out}})} \exp(\psi(\mathcal{E}))} \geq 1 - p^{-(t-t_0-3)},$$

where $\psi(\mathcal{E}) = \psi(G)$ *for any* $G \in \mathcal{E}$ *and* $\psi(G)$ *is defined in* (15). *Further, the greedy search on* $(\mathcal{C}_p(d_{\text{in}}, d_{\text{out}}), \mathcal{N}_{\text{ads}}, e^\psi)$ *returns* $\mathcal{E}^*$ *regardless of the initial state.*

PROOF. See Supplementary Material Section F.4. □

REMARK 10. The assumption $d_{\text{in}} + d_{\text{out}} = O(\log p)$ is mild, since the total number of edges in the DAG may have order p even if $d_{\text{in}} + d_{\text{out}} = O(1)$. In light of Assumption (A4.2), we may assume d^*, d_{in} have approximately the same order. Thus, roughly speaking, the assumptions of Theorem 5 imply that d^*, d_{in} cannot grow faster than $\sqrt{\log p}$.

4.4. *Consistency results for sub-Gaussian random matrices.* The normality assumption on the true distribution of X can be relaxed. We can extend the consistency result obtained in Theorem 4 to the case where X is a sub-Gaussian random matrix (we still consider the posterior distributions defined in Section 4.1). Let each row of X be an i.i.d. copy of a random vector X, which has mean zero, covariance matrix Σ^* and distribution μ. Assume that μ is sub-Gaussian with sub-Gaussian parameter bounded by a universal constant, and $N_p(0, \Sigma^*)$ is perfectly Markovian w.r.t. a DAG G^* (i.e., μ is not necessarily perfectly Markovian w.r.t. G^*). This includes the case where some node variables are Gaussian and some are discrete and bounded [32]. Then, under a set of similar assumptions, we can prove a consistency result analogous to Theorem 4(i); see Theorem F1 in Supplementary Material Section F.5. By Corollary 2, this proves part (ii) of Condition 1, and other strong selection consistency results in Theorem 4 follow.

The main idea of the proof of Theorem F1 is similar to the Gaussian case. We first generalize the variable selection results of Yang, Wainwright and Jordan [65] to random matrices, which is performed in Supplementary Material Section E.3. However, the proof techniques are very different from the Gaussian case in that we need to use random matrix theory [62] and error propagation results to show that all minimal I-maps of G^* can be recovered from the empirical covariance matrix. The key distinction between the two scenarios is that in the sub-Gaussian case uncorrelatedness does not imply independence. Consequently, some calculations are more involved, and we need to require a slightly stronger assumption on c_2: in the sub-Gaussian case, we require $d_{\mathrm{in}}\overline{\nu}^4/\underline{\nu}^6 = O(c_2)$, while in the Gaussian case we only need $d_{\mathrm{in}} = O(c_2)$.

5. Mixing time results for Bayesian structure learning.

5.1. *Rapid mixing of the RW-GES sampler.* Recall that RW-GES is simply the random walk MH algorithm defined by (2) with $h \equiv 1$ and the triple $(\mathcal{C}_p(d_{\mathrm{in}}, d_{\mathrm{out}}), \mathcal{N}_{\mathrm{ads}}, \pi_n)$ where π_n is given by (17). In the proof of Theorem 5, we have verified that Condition 1 holds, and thus we can apply the mixing time bounds in Section 2.2 to obtain the main result of this work, rapid mixing of RW-GES.

THEOREM 6. *Consider the setting of Theorem 5, and let $\pi_{\min} = \min_{\mathcal{E}\in\mathcal{C}_p(d_{\mathrm{in}},d_{\mathrm{out}})} \pi_n(\mathcal{E})$. Let $\mathbf{P}$ denote the transition matrix of the RW-GES sampler and $\mathbf{P}_{\mathrm{lazy}}$ denote its lazy version. For sufficiently large n, with probability at least $1 - 3p^{-1}$, we have*

$$T_{\mathrm{mix}}(\mathbf{P}_{\mathrm{lazy}}) \le C t_0 p^{t_0+2} (\log p) \log\left(\frac{4}{\pi_{\min}}\right),$$

for some universal constant C, where t_0 is as given in Theorem 5.

PROOF. See Supplementary Material Section G.1. □

COROLLARY 4. *Suppose Assumptions (A1) and (A2) hold. We have*

$$\min_{\mathcal{E}\in\mathcal{C}_p(d_{\mathrm{in}},d_{\mathrm{out}})} \frac{\pi_n(\mathcal{E})}{\pi_n(\mathcal{E}^*)} \ge \left(c_1 p^{c_2}\sqrt{1+\alpha/\gamma}\right)^{-p(d_{\mathrm{in}}+d^*)} \left(\frac{2\overline{\nu}}{\underline{\nu}}\right)^{-p(\alpha n+\kappa)/2}.$$

Hence, under the setting of Theorem 6, the mixing time of the RW-GES sampler can be bounded by a polynomial of n and p.

PROOF. See Supplementary Material Section G.2. □

REMARK 11. Corollary 4 implies that RW-GES is rapidly mixing with high probability. The term $\log \pi_{\min}$ in the mixing time bound is only used to handle the worst scenario where the chain starts from the state with minimum posterior probability. If the chain starts from some "good" estimate, the actual mixing rate of the chain can be much faster; see Proposition 1 of Sinclair [52].

If in the beta-min condition, we only assume that the minimum edge weight of B^* (the weighted adjacency matrix of the true DAG G^*) is sufficiently large, the rapid mixing of RW-GES does not hold. It is not difficult to construct an explicit example where RW-GES is slowly mixing. In the following example, we let $p = 3$ be fixed and show that the mixing time grows exponentially in n. One can extend our example to the case $p = n$ by adding variables $\mathsf{X}_4, \dots, \mathsf{X}_n$ such that, for any $j = 4, \dots, n$, the observed vector X_j is exactly orthogonal to all the other column vectors of the data matrix.

EXAMPLE 4. Assume $p = 3$ and the true SEM is given by

$$X_1 = z_1, \qquad X_2 = b_1 X_1 + z_2, \qquad X_3 = b_2 X_2 + z_3,$$

where z_1, z_2, z_3 are vectors orthogonal to each other and $\|z_j\|_2^2 = n$ for each j. Thus, we can let the true DAG G^* be $1 \to 2 \to 3$. Suppose the prior parameters satisfy that $d_{\text{in}} = d_{\text{out}} = 2$, $c_2 = \sqrt{n}$, $\kappa = 0$ and c_1, α, γ are fixed constants such that $c_1\sqrt{1 + \alpha/\gamma} = 1$. Assume the true regression coefficients $b_1, b_2 > 0$ are given by

$$b_1^2 = b_2^2 = \frac{K c_2 \log p}{\alpha n} = o(1),$$

where K is some large universal constant. So, b_1, b_2 satisfy the bound in (18). Consider the DAG $\tilde{G}$ given by $1 \to 2 \leftarrow 3$, which has $[\tilde{G}] = \{\tilde{G}\}$. The topological ordering of $\tilde{G}$ can be chosen to be $\sigma = (1, 3, 2)$, and the minimal I-map G_σ^* is a complete DAG. One can show that the edge weight of $1 \to 3$ in G_σ^* is $b_1 b_2$. It is easy to verify that $b_1^2 b_2^2 = o(c_2 n^{-1} \log p)$, so the true model fails to satisfy the strong beta-min condition. Indeed, we can prove that RW-GES is slowly mixing. See Supplementary Material Section G.4.

5.2. *Rapid mixing results for sparse DAG selection.* Suppose the ordering is given and the search is restricted to $\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$. We can construct a random walk MH sampler using neighborhood function $\mathcal{N}_{\text{ads}}$ defined in (3) and posterior distribution π_n^σ defined in (14), which is just the standard add-delete-swap MH sampler. Denote its transition matrix by $\mathbf{P}^\sigma$. If there is no out-degree constraint, by posterior modularity, one can perform sampling for the parent set of each node separately; thus, there is no need to directly draw DAG samples. However, when $d_{\text{out}} < p$, the posterior distributions of $\text{Pa}_1, \dots, \text{Pa}_p$ are not independent, and this add-delete-swap sampler provides a convenient solution. Since by Theorem 4(i) and Corollary 1, the triple $(\mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}}), \mathcal{N}_{\text{ads}}, \pi_n^\sigma)$ satisfies Condition 1; the mixing time bound for $\mathbf{P}^\sigma$ immediately follows from Theorem 2.

THEOREM 7. *Suppose Assumptions* (A1), (A2), (A3), (A4.1) *and* (A5.1) *hold for some* $\sigma \in \mathbb{S}^p$, $t > 3$ *and* $C_\beta \geq 8t/3$. *Further, assume that* $\min\{d_\sigma^* d_{\text{in}} + 1, p\} \leq d_{\text{out}}$. *For sufficiently large* n, *with probability at least* $1 - 3p^{-1}$, *we have*

$$T_{\text{mix}}(\mathbf{P}_{\text{lazy}}^\sigma) \leq C d_{\text{in}} p^2 \log\left(\frac{4}{\pi_{\min}^\sigma}\right),$$

for some universal constant C, *where* $\pi_{\min}^\sigma = \min_{G \in \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})} \pi_n^\sigma(G)$.

PROOF. See Supplementary Material Section G.3. □

REMARK 12. The assumptions are much weaker than those used in Theorem 6. In particular, we can allow a much larger model size for each nodewise variable selection problem. This is mainly because for any $G \in \mathcal{G}_p^\sigma(d_{\text{in}}, d_{\text{out}})$, we have $|\mathcal{N}_{\text{ads}}(G)| = O(d_{\text{in}} p^2)$. But for an equivalence class $\mathcal{E} \in \mathcal{C}_p(d_{\text{in}}, d_{\text{out}})$, the size of $\mathcal{N}_{\text{ads}}(\mathcal{E})$ may grow exponentially in $d_{\text{in}} + d_{\text{out}}$.

5.3. *Slow mixing examples for a CPDAG sampler.* The neighborhood $\mathcal{N}_{\text{ads}}(\mathcal{E})$ used in RW-GES can be very large for some $\mathcal{E}$, which seems to be undesirable. However, other choices of the neighborhood relation on $\mathcal{C}_p$ (which may seem very reasonable) can cause the search algorithm to be trapped in suboptimal local modes.

A popular approach to constructing sampling algorithms on $\mathcal{C}_p$ is to use the CPDAG (completed partially directed acyclic graph) representations of equivalence classes. Any equivalence class $\mathcal{E}$ can be uniquely represented by a CPDAG, a partially directed acyclic graph that satisfies two conditions: (i) it has the same skeleton as any $G \in \mathcal{E}$; (ii) an edge is directed if and only if the edge is directed in the same orientation in every $G \in \mathcal{E}$. A CPDAG is also called an essential graph [2]. One can define local proposal moves on $\mathcal{C}_p$ by modifying CPDAGs. However, one can easily end up with a CPDAG sampler that is slowly mixing even when p is fixed and n goes to infinity.

EXAMPLE 5. Let $p = 3$ and the true data-generating DAG G^* be $1 \to 3 \leftarrow 2$. Since G^* is the only member in $\mathcal{E}^* = [G^*]$, the CPDAG of $\mathcal{E}^*$ is the same as G^*. Let $\tilde{\mathcal{E}}$ be the equivalence class that contains all complete DAGs. It is easy to verify that the CPDAG of $\tilde{\mathcal{E}}$ is a complete undirected graph. If we define the neighborhood of $\tilde{\mathcal{E}}$ as all the CPDAGs that can be obtained by adding or removing a directed or undirected edge from $\tilde{\mathcal{E}}$, then the only CPDAGs we can move to from $\tilde{\mathcal{E}}$ are $1 - 2 - 3$, $1 - 3 - 2$ and $2 - 1 - 3$. However, given sufficiently large sample size, all these three CPDAGs should have much smaller score than $\tilde{\mathcal{E}}$. For example, the CPDAG $1 - 3 - 2$ encodes the CI relation $1 \perp\!\!\!\perp 2|3$, which does not exist in G^*, and thus connecting nodes 1 and 2 should increase the score. See Supplementary Material Section G.5 for an explicit construction of this example and another 5-node example, where we further prove that the CPDAG sampler proposed by Castellettiet al. [8] is slowly mixing.

6. Simulation studies on the RW-GES sampler.

6.1. *A rapid mixing example.* In this section, we present three simulation studies, which illustrate the theoretical results we have proved. We first construct a rapid mixing example for $p = 100$ and $n = 800$. In order to approximately satisfy the strong beta-min condition, we randomly generate the true DAG G^* such that its maximum node degree is 2 and its largest connected sub-DAG only has 10 nodes, and then for each edge (i, j) in G^*, we sample B^*_{ij} from the uniform distribution on $(0.5, 1.5) \cup (-1.5, -0.5)$. The DAG G^* we obtain has 66 edges, among which 24 are directed in the CPDAG representation of $[G^*]$; see Supplementary Material Section H.2 for the visualization. Each row of the data matrix X is drawn independently from $N_p(0, \Sigma^*)$ where $\Sigma^* = (I - (B^*)^\top)^{-1}(I - B^*)^{-1}$. We use $\alpha = 0.99$, $\gamma = 0.01$, $\kappa = 0$, $c_1 = 1$, $c_2 = 2$ and run 20 RW-GES chains, all initialized at the null model, for 5×10^4 iterations. All 20 runs are able to find the true equivalence class in about 10^5 iterations, which indicates a fast mixing rate; see the left panel of Figure 2. This example illustrates that though the strong beta-min condition is restrictive, if the true DAG is sufficiently sparse and has a "simple" structure, RW-GES can be rapidly mixing for a moderately large sample size (in Supplementary Material Section H.2, we use this idea to explicitly construct toy examples with $p \gg n$ that satisfy all assumptions of Theorem 6). For comparison, we repeat the analysis by only using the first 200 observations, and we find that 11 chains

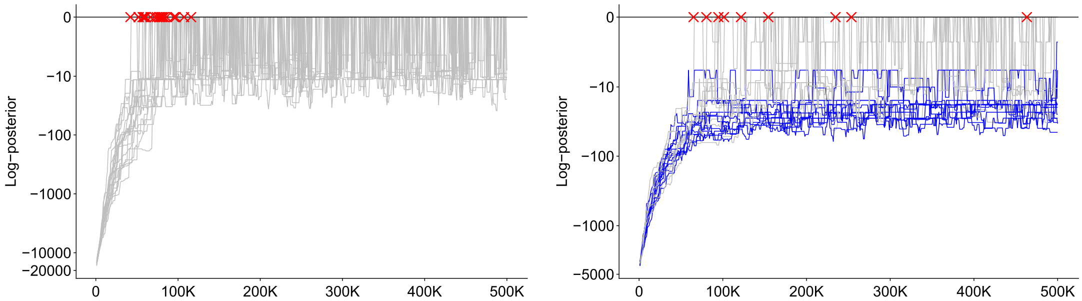

FIG. 2. *Trajectories of* 20 *independent RW-GES runs for a simulated data set with* $p = 100$. *Left*: $n = 800$; *right*: $n = 200$. *The posterior probabilities are unnormalized and the log-posterior of* $[G^*]$ *is set to zero. Red crosses mark the times that RW-GES first collects* $[G^*]$. *Runs that never sample* $[G^*]$ *are shown in blue.*

fail to sample $[G^*]$. The right panel of Figure 2 suggests that these 11 chains get stuck at different local modes. Notice that $[G^*]$ still seems to have the largest posterior probability in this case, which implies that the beta-min condition at least holds for the true ordering (i.e., if the true ordering is known, we can recover the true DAG). However, since n is small, the strong beta-min condition is significantly violated, which makes the posterior distribution on the space of equivalence classes highly multimodal.

6.2. *Performance in a high-dimensional scenario.* The complexity of the structure learning problem largely depends on p and the sparsity level of the true DAG G^*. We can roughly measure the sparsity using the maximum degree of G^*, denoted by $\deg(G^*)$. Assumption (A2) and Remark 10 suggest that we consider $\deg(G^*) \log p = O(n)$ and $\deg(G^*) = O(\sqrt{\log p})$. In the second simulation study, we examine these asymptotic orders by using 7 simulation settings where n grows linearly and p grows exponentially. Given p, we generate G^* by first sampling a random ordering and then including each edge with probability $D/(p-1)$, where the parameter D gives the expected number of neighbors of each node. We let D grow at rate $\sqrt{n}$ (so we actually let $\deg(G^*) \log p$ grow slightly faster than n). We generate X using the normal SEM associated with G^* and choose the hyperpamareters in the same way as in Section 6.1. The number of RW-GES iterations is set to grow polynomially with p but slightly slower than p^2. We always initiate the sampler at the null model and discard the first 80% iterations as burn-in. For each setting, we generate 20 replicates (G^* and X are resampled each time), and the results are shown in Table 1 (see Supplementary Material Section H.2 for the definition of true/false positive rates). Observe that the true positive rate

TABLE 1
Performance of RW-GES in 7 *settings. For the kth setting,* $p = 7 \cdot 2^{k-1}$, $n = 30(k+1)$, $D = 0.2\sqrt{n}$ *and the number of RW-GES iterations* $N_{\text{mcmc}} \approx 300 \cdot (4p/7)^{1.66}$. *TPR* (*skeleton*): *true positive rate with edge directions ignored*; *TPR*: *true positive rate* (*edge directions determined by the CPDAG*); *FPR*: *false positive rate. Results are averaged over* 20 *replicates, and the number in parentheses is the standard error*

p	n	D	$N_{\text{mcmc}}/1000$	TPR (skeleton)	TPR	FPR
7	60	1.549	3	0.854 (0.03)	0.721 (0.06)	0.047 (0.02)
14	90	1.897	10	0.89 (0.02)	0.668 (0.06)	0.03 (0.006)
28	120	2.191	30	0.91 (0.01)	0.73 (0.03)	0.017 (0.003)
56	150	2.449	100	0.871 (0.01)	0.629 (0.03)	0.015 (0.002)
112	180	2.683	300	0.866 (0.01)	0.634 (0.02)	0.0091 (0.0005)
224	210	2.898	1000	0.86 (0.008)	0.634 (0.01)	0.0049 (0.0002)
448	240	3.098	3000	0.869 (0.004)	0.648 (0.008)	0.0027 (0.00007)

TABLE 2

Simulation study with $p = 20$, $n = 100$ and expected node degree $D = 4$. Results are averaged over 50 data sets

	$c_2 = 1.3$			$c_2 = 1.1 + 0.1d$		
d	TPR (skeleton)	TPR	FPR	TPR (skeleton)	TPR	FPR
4	0.542 (0.02)	0.307 (0.02)	0.0824 (0.004)	0.539 (0.01)	0.316 (0.02)	0.0788 (0.004)
5	0.61 (0.01)	0.339 (0.02)	0.101 (0.004)	0.601 (0.01)	0.325 (0.02)	0.0977 (0.004)
6	0.665 (0.01)	0.383 (0.02)	0.115 (0.005)	0.657 (0.01)	0.393 (0.02)	0.101 (0.005)
7	0.706 (0.01)	0.412 (0.02)	0.123 (0.006)	0.699 (0.01)	0.419 (0.02)	0.096 (0.005)
8	0.72 (0.01)	0.413 (0.02)	0.132 (0.006)	0.694 (0.01)	0.421 (0.02)	0.0993 (0.006)
9	0.718 (0.01)	0.401 (0.02)	0.138 (0.007)	0.695 (0.01)	0.437 (0.02)	0.0932 (0.005)

(for both skeleton and CPDAG estimation) becomes stable as p grows, while the false positive rate even decreases. Though for most real-world problems, the strong beta-min condition is unlikely to be satisfied and $[G^*]$ may not be correctly identified, this study shows that the theoretical insights on the MCMC complexity is useful. In particular, the performance of RW-GES seems stable under the asymptotic regime $\deg(G^*) \log p = O(n)$.

6.3. *On the choice of* c_2. The third simulation study aims to investigate the optimal choice of c_2, which is the most important prior hyperparameter of our model since it determines the order of the penalty on the graph size. We fix $p = 20$ and $n = 100$ and generate 50 true DAGs and data sets using the method described in Section 6.2 with $D = 4$ (recall this gives the expected degree of a single node). When implementing RW-GES, we impose the maximum degree constraint, denoted by d (i.e., the sampler only searches equivalence classes with maximum degree bounded by d); see Supplementary Material Section H.1 for details. RW-GES is run for 40,000 iterations for each simulated data set. We first fix $c_2 = 1.3$ and try $d = 4, 5, \dots, 9$. The results are shown in the left column of Table 2. True positive rates increase with d, since some nodes in the true DAG may have large degrees and their incoming edges cannot all be detected if d is small. However, the false positive rate also increases because the search space quickly grows with d. Next, we repeat the experiment by setting $c_2 = 1.1 + 0.1d$, which according to our tests, appears to yield close-to-optimal performance in this simulation setting. As can be seen from the right column of Table 2, the false positive rate remains roughly a constant and the true positive rates are comparable or even better than those for $c_2 = 1.3$. Recall that to prove posterior consistency, we assume c_2 is greater than $4(\alpha + 1)d_{\text{in}}$ plus some constant in Assumption (A3). This simulation study shows that, though the coefficient in Assumption (A3) is quite pessimistic, the linear growth rate (w.r.t. the maximum degree constraint) is a useful rule of thumb for tuning c_2 in practice.

7. Discussion.

7.1. *Mixing of structure MCMC and order MCMC methods.* In this work, we have only analyzed the mixing times of MCMC algorithms defined on the space of equivalence classes, but the same strategy can be pursued to study samplers defined on the DAG space and order space. Observe that the canonical paths we constructed in Section 3.6 for the RW-GES sampler can also be thought of as paths on the DAG space. Given an equivalence class $\mathcal{E}$, we first pick arbitrarily some $G \in \mathcal{E}$. If G has ordering σ, we move from G to the minimal I-map G^*_σ by only add-delete-swap modifications of the DAG. To move from G^*_σ to G^*, we have to change the ordering. For RW-GES, the neighborhood function defined in (4) allows us to "switch" from G^*_σ to a Markov equivalent DAG $\tilde{G}$, which is still an I-map of G^* but no longer minimal, and then we can remove edges from $\tilde{G}$ (the existence of such $\tilde{G}$ is guaranteed by Chickering algorithm). Repeating this procedure, we obtain a path from G^*_σ to G^*.

Consider the classical structure MCMC sampler, a random walk MH algorithm defined on $\mathcal{G}_p$ that uses single-edge addition, deletion and reversal to propose local moves [9, 36]. Since any two Markov equivalent DAGs G, G' are connected by a sequence of covered edge reversals (see Supplementary Material Section C.1), structure MCMC is able to traverse equivalence classes and move from G^*_σ to any other Markov equivalent DAG. Therefore, the canonical paths of RW-GES are also paths of structure MCMC (introduce swap moves if a restricted space is considered). The same argument can be applied to order MCMC samplers, since a covered edge reversal can be seen as an adjacent transposition on $\mathbb{S}^p$ [53]. Unfortunately, the size of an equivalence class can easily be very large, and it is unclear whether structure MCMC can always quickly leave any equivalence class even if the maximum degree is bounded. In Supplementary Material Section G.6, we construct an interesting example where G_0 is Markov equivalent to $G^* \cup \{2 \to 1\}$ but it is quite difficult for structure MCMC to remove the edge between nodes 1 and 2 from G_0. Indeed, we show that on average it takes structure MCMC $O(p^4)$ iterations to move from G_0 to G^*, while it only takes RW-GES $O(p^2)$ iterations to move from $[G_0]$ to $[G^*]$. Nevertheless, we conjecture that structure MCMC is still rapidly mixing under the assumptions we used in Section 4 (recall "rapid mixing" only requires the mixing time to be polynomial in n and p), though the proof would probably require a skillful analysis of how the size of an equivalence class changes with single-edge modifications of its member DAGs.

One caveat is that the target posterior distributions on DAG and order spaces are typically different from our target π_n defined in (17). For example, for DAG MCMC methods, it is convenient to use the prior $\pi_0^{\mathrm{dag}}(G) \propto (c_1 p^{c_2})^{-|G|}$ for $G \in \mathcal{G}_p(d_{\mathrm{in}}, d_{\mathrm{out}})$, which yields the posterior $\pi_n^{\mathrm{dag}}(G) \propto e^{\psi(G)} \mathbb{1}_{\mathcal{G}_p(d_{\mathrm{in}}, d_{\mathrm{out}})}(G)$. Comparing them with (16) and (17), we see that $\pi_0^{\mathrm{dag}}(\mathcal{E}) = \sum_{G \in \mathcal{E}} \pi_0^{\mathrm{dag}}(G) \propto |\mathcal{E}| \pi_0(\mathcal{E})$ and $\pi_n^{\mathrm{dag}}(\mathcal{E}) \propto |\mathcal{E}| \pi_n(\mathcal{E})$. Note that we do not use π_0^{dag} for equivalence class samplers [8] since calculating the size of $\mathcal{E}$ can be extremely time-consuming. On the order space, the situation is more subtle since one DAG can be compatible with multiple orderings [14, 15]. If rapid mixing of structure MCMC can be established, we expect that the same argument can be used to show the strong selection consistency of π_n^{dag}.

7.2. *Advantages and extensions of RW-GES.*

One advantage of RW-GES over GES is that RW-GES considers a restricted search space and is equipped with the swap proposal. This is particularly important to theoretical analysis. Nonsparse models can easily overfit the data (e.g., if node j has more than n parents, then X_j can be perfectly explained leading to an infinite score), which is why the sparsity constraint is necessary for proving high-dimensional consistency results. For GES, even if the maximum degree of G^* is bounded, there is still a possibility that GES visits nonsparse equivalence classes along its search path and then its behavior becomes completely unpredictable. In the proof of Nandy, Hauser and Maathuis [42] on the high-dimensional consistency of GES, the authors directly assumed that the output of the first stage is not too large; see Assumption (A5) therein.

The main methodological difference between the two algorithms is that GES is essentially an optimization algorithm, while RW-GES is used for sampling. The general theory on the relation between optimization and sampling suggests that each has its own unique advantages [58]. In particular, when the sample size is not large, the posterior tends to be multimodal and MCMC sampling (if it converges) can yield better estimates via model averaging [24]. One can also use the output of GES as the initial state for RW-GES, which to some extent, may achieve the benefits of both methods. In our theoretical analysis, we choose to focus on RW-GES just for its simplicity. One can generalize it in many ways to improve its performance in practice, for example, by using an informed proposal scheme or combining it with tempering techniques (i.e., running multiple RW-GES samplers at different temperatures). One simple modification that may significantly improve the sampler's performance is to first estimate a large conditional independence graph [40, 50] and then use it to tune the proposal probabilities. This can be seen as a randomized extension of the method of Nandy, Hauser and Maathuis [42]. The canonical paths we construct in Section 3 can always be applied as long as the sampler proposes states from $\mathcal{N}_{\text{ads}}(\cdot)$ (or a superset of it). But one important takeaway from our theory is that using a neighborhood smaller than $\mathcal{N}_{\text{ads}}(\cdot)$ may lead to slow mixing even when the sample size is sufficiently large. A detailed investigation into more sophisticated local MCMC schemes using $\mathcal{N}_{\text{ads}}(\cdot)$ is left to future research.

Acknowledgments. The authors would like to thank all anonymous reviewers whose comments have helped improve the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplementary material for "Complexity analysis of Bayesian learning of high-dimensional DAG models and equivalence classes." (DOI: 10.1214/23-AOS2280SUPP; .pdf). Part A: a notation table. Part B: more results for mixing times of finite Markov chains and proofs for Section 2. Part C: preliminaries for graphical models. Part D: proofs for Section 3. Part E: auxiliary results for high-dimensional empirical variable selection. Part F: proofs for Section 4. Part G: proofs and examples for Sections 5 and 7. Part H: further details about RW-GES implementation and simulation studies. Part I: discussion on the case where the strong beta-min or faithfulness condition fails.

REFERENCES

[1] AGRAWAL, R. and UHLER, C. (2018). Minimal I-MAP MCMC for scalable structure discovery in causal DAG models. In *International Conference on Machine Learning* 89–98.

[2] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312 https://doi.org/10.1214/aos/1031833662

[3] ARAGAM, B., AMINI, A. and ZHOU, Q. (2019). Globally optimal score-based learning of directed acyclic graphs in high-dimensions. In *Advances in Neural Information Processing Systems* 4450–4462.

[4] BANERJEE, S., CASTILLO, I. and GHOSAL, S. (2021). Bayesian inference in high-dimensional models. arXiv preprint. Available at arXiv:2101.04491.

[5] BANERJEE, S. and GHOSAL, S. (2015). Bayesian structure learning in graphical models. *J. Multivariate Anal.* **136** 147–162. MR3321485 https://doi.org/10.1016/j.jmva.2015.01.015

[6] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

[7] CAO, X., KHARE, K. and GHOSH, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47** 319–348. MR3909935 https://doi.org/10.1214/18-AOS1689

[8] CASTELLETTI, F., CONSONNI, G., DELLA VEDOVA, M. L. and PELUSO, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. *Bayesian Anal.* **13** 1235–1260. MR3855370 https://doi.org/10.1214/18-BA1101

[9] CASTELO, R. and KOČKA, T. (2003). On inclusion-driven learning of Bayesian networks. *J. Mach. Learn. Res.* **4** 527–574. MR2072261 https://doi.org/10.1162/153244304773936045

[10] CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498. MR1929415 https://doi.org/10.1162/153244302760200696

[11] CHICKERING, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554. MR1991085 https://doi.org/10.1162/153244303321897717

[12] DRTON, M., FOYGEL, R. and SULLIVANT, S. (2011). Global identifiability of linear structural equation models. *Ann. Statist.* **39** 865–886. MR2816341 https://doi.org/10.1214/10-AOS859

[13] DRTON, M. and MAATHUIS, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* **4** 365–393.

[14] EATON, D. and MURPHY, K. (2007). Bayesian structure learning using dynamic programming and MCMC. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* 101–108.

[15] ELLIS, B. and WONG, W. H. (2008). Learning causal Bayesian network structures from experimental data. *J. Amer. Statist. Assoc.* **103** 778–789. MR2524009 https://doi.org/10.1198/016214508000000193

[16] FRIEDMAN, N. and KOLLER, D. (2003). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Mach. Learn.* **50** 95–125.

[17] GAO, B. and CUI, Y. (2015). Learning directed acyclic graphical structures with genetical genomics data. *Bioinformatics* **31** 3953–3960. https://doi.org/10.1093/bioinformatics/btv513

[18] GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909

[19] GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. MR1936324 https://doi.org/10.1214/aos/1035844981

[20] GIUDICI, P. and CASTELO, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Mach. Learn.* **50** 127–158.

[21] GOUDIE, R. J. B. and MUKHERJEE, S. (2016). A Gibbs sampler for learning DAGs. *J. Mach. Learn. Res.* **17** Paper No. 30, 39. MR3491124

[22] GRZEGORCZYK, M. and HUSMEIER, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Mach. Learn.* **71** 265.

[23] HE, Y., JIA, J. and YU, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Ann. Statist.* **41** 1742–1779. MR3127848 https://doi.org/10.1214/13-AOS1125

[24] HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176 https://doi.org/10.1214/ss/1009212519

[25] JEONG, S. and GHOSAL, S. (2021). Posterior contraction in sparse generalized linear models. *Biometrika* **108** 367–379. MR4259137 https://doi.org/10.1093/biomet/asaa074

[26] JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074 https://doi.org/10.1080/01621459.2012.682536

[27] KAHALE, N. (1997). A semidefinite bound for mixing rates of Markov chains. *Random Structures Algorithms* **11** 299–313. MR1608821 https://doi.org/10.1002/(SICI)1098-2418(199712)11:4<299::AID-RSA2>3.0.CO;2-U

[28] KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

[29] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models: Principles and Techniques. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2778120

[30] KUIPERS, J. and MOFFA, G. (2017). Partition MCMC for inference on acyclic digraphs. *J. Amer. Statist. Assoc.* **112** 282–299. MR3646571 https://doi.org/10.1080/01621459.2015.1133426

[31] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 https://doi.org/10.1214/09-AOS720

[32] LAURITZEN, S. L. (1992). Propagation of probabilities, means, and variances in mixed graphical association models. *J. Amer. Statist. Assoc.* **87** 1098–1108. MR1209568

[33] LEE, K., LEE, J. and LIN, L. (2019). Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors. *Ann. Statist.* **47** 3413–3437. MR4025747 https://doi.org/10.1214/18-AOS1783

[34] LIU, S., SUZUKI, T., RELATOR, R., SESE, J., SUGIYAMA, M. and FUKUMIZU, K. (2017). Support consistency of direct sparse-change learning in Markov networks. *Ann. Statist.* **45** 959–990. MR3662445 https://doi.org/10.1214/16-AOS1470

[35] MAATHUIS, M. H., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248. https://doi.org/10.1038/nmeth0410-247

[36] MADIGAN, D., YORK, J. and ALLARD, D. (1995). Bayesian graphical models for discrete data. *Int. Stat. Rev.* 215–232.

[37] MADIGAN, D., ANDERSSON, S. A., PERLMAN, M. D. and VOLINSKY, C. T. (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm. Statist. Theory Methods* **25** 2493–2519.

[38] MARTIN, R., MESS, R. and WALKER, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879 https://doi.org/10.3150/15-BEJ797

[39] MEEK, C. (1997). Graphical models: Selecting causal and statistical models. Ph.D. thesis, Carnegie Mellon Univ., Pittsburgh, PA.

[40] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[41] MUNTEANU, P. and BENDOU, M. (2001). The EQ framework for learning equivalence classes of Bayesian networks. In *Proceedings* 2001 *IEEE International Conference on Data Mining* 417–424. IEEE, San Jose, CA.

[42] NANDY, P., HAUSER, A. and MAATHUIS, M. H. (2018). High-dimensional consistency in score-based and hybrid structure learning. *Ann. Statist.* **46** 3151–3183. MR3851768 https://doi.org/10.1214/17-AOS1654

[43] NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987 https://doi.org/10.1214/14-AOS1207

[44] NIINIMÄKI, T. M., PARVIAINEN, P. and KOIVISTO, M. (2011). Partial order MCMC for structure discovery in Bayesian networks. In *Proceedings of the Twenty-Seventh Conference Conference on Uncertainty in Artificial Intelligence* (*UAI*-11) 557–564. AUAI Press, Barcelona, Spain.

[45] PATI, D., BHATTACHARYA, A., PILLAI, N. S. and DUNSON, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.* **42** 1102–1130. MR3210997 https://doi.org/10.1214/14-AOS1215

[46] PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. *The Morgan Kaufmann Series in Representation and Reasoning*. Morgan Kaufmann, San Mateo, CA. MR0965765

[47] PELUSO, S. and CONSONNI, G. (2020). Compatible priors for model selection of high-dimensional Gaussian DAGs. *Electron. J. Stat.* **14** 4110–4132. MR4170698 https://doi.org/10.1214/20-EJS1768

[48] PENA, J. M. (2007). Approximate counting of graphical models via MCMC. In *AISTATS* 355–362.

[49] PERLMAN, M. D. (2001). Graphical model search via essential graphs. In *Algebraic Methods in Statistics and Probability* (*Notre Dame*, *IN*, 2000). *Contemp. Math.* **287** 255–265. Amer. Math. Soc., Providence, RI. MR1873680 https://doi.org/10.1090/conm/287/04790

[50] RASKUTTI, G., YU, B. and WAINWRIGHT, M. J. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. *Adv. Neural Inf. Process. Syst.* **21**.

[51] SCUTARI, M., GRAAFLAND, C. E. and GUTIÉRREZ, J. M. (2019). Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *Internat. J. Approx. Reason.* **115** 235–253. MR4018631 https://doi.org/10.1016/j.ijar.2019.10.003

[52] SINCLAIR, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combin. Probab. Comput.* **1** 351–370. MR1211324 https://doi.org/10.1017/S0963548300000390

[53] SOLUS, L., WANG, Y. and UHLER, C. (2021). Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika* **108** 795–814. MR4341352 https://doi.org/10.1093/biomet/asaa104

[54] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

[55] STUDENÝ, M. (2005). *Probabilistic Conditional Independence Structures*. *Information Science and Statistics*. Springer, London. MR3183760

[56] SU, C. and BORSUK, M. E. (2016). Improving structure MCMC for Bayesian networks through Markov blanket resampling. *J. Mach. Learn. Res.* **17** Paper No. 118, 20. MR3543524

[57] SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. MR3144466

[58] TALWAR, K. (2019). Computational separations between sampling and optimization. *Adv. Neural Inf. Process. Syst.* **32**.

[59] UHLER, C., RASKUTTI, G., BÜHLMANN, P. and YU, B. (2013). Geometry of the faithfulness assumption in causal inference. *Ann. Statist.* **41** 436–463. MR3099109 https://doi.org/10.1214/12-AOS1080

[60] VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. MR3099113 https://doi.org/10.1214/13-AOS1085

[61] VERMA, T. and PEARL, J. (1991). Equivalence and synthesis of causal models. Technical report, UCLA, Computer Science Department.

[62] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[63] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

[64] YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139 https://doi.org/10.1214/14-AOS1289

[65] YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.* **44** 2497–2532. MR3576552 https://doi.org/10.1214/15-AOS1417

[66] ZANELLA, G. (2020). Informed proposals for local MCMC in discrete spaces. *J. Amer. Statist. Assoc.* **115** 852–865. MR4107684 https://doi.org/10.1080/01621459.2019.1585255

[67] ZHOU, Q. and CHANG, H. (2023). Supplement to "Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes." https://doi.org/10.1214/23-AOS2280SUPP

[68] ZHOU, Q., YANG, J., VATS, D., ROBERTS, G. O. and ROSENTHAL, J. S. (2022). Dimension-free mixing for high-dimensional Bayesian variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1751–1784. MR4515557 https://doi.org/10.1111/rssb.12546

[69] ZHUO, B. and GAO, C. (2021). Mixing time of Metropolis–Hastings for Bayesian community detection. *J. Mach. Learn. Res.* **22** Paper No. 10, 89. MR4253703