Supplement to "Dimension-free Mixing for High-dimensional Bayesian Variable Selection"

Quan Zhou, Jun Yang, Dootika Vats, Gareth O. Roberts and Jeffrey S. Rosenthal

Section S1 is a brief review on some known results for the drift condition, and the proof of Theorem 2 is provided in Section S2. In Section S3, we state the main result of Yang et al. [2016] and explain how to establish Condition 1 for any fixed constants $c_1, c_2 \geq 0$, using essentially the same assumptions. Section S4 provides the proofs for all results given in Section 3. Section S5 gives the information about the code and real data used in this work.

S1. Preliminary results for the drift condition

We use the notation introduced in Section 4. Given a drift condition on the set $\mathsf{X} \setminus \mathsf{C}$, it is well known that the entry time of the chain into C has a "thin-tailed" distribution.

Lemma S1. *Let $(\mathsf{X}_t)_{t \in \mathbb{N}}, \mathcal{X}, \boldsymbol{P}$ be as given in Assumption A. Suppose that there exist a function $V \colon \mathcal{X} \mapsto [1, \infty)$, a constant $\lambda \in (0, 1)$, and a non-empty set $\mathsf{C} \in \mathcal{E}$ such that $(\boldsymbol{P}V)(x) \leq \lambda V(x)$ for every $x \notin \mathsf{C}$. Let $\tau_{\mathsf{C}} = \min\{t \geq 0 \colon \mathsf{X}_t \in \mathsf{C}\}$. Then, for any $x \in \mathcal{X}$,*

$$\mathbb{E}_x[\lambda^{-\tau_{\mathsf{C}}}] \leq V(x), \quad \text{and } \mathbb{P}_x(\tau_{\mathsf{C}} \geq t) \leq \lambda^t V(x), \quad \forall\, t \in \mathbb{N}.$$

Proof. Let $\mathsf{Y}_t = \lambda^{-t} V(\mathsf{X}_t)$. The drift condition implies that $\mathsf{Y}_{t \wedge \tau}$ is a supermartingale. The results then follow from optional sampling theorem and Markov's inequality. □

The following theorem due to Jerison [2016] gives a very useful bound on the mixing time when we have the generating function of the hitting time of some state x^*. In the original version [Jerison, 2016, Theorem 4.5], it is assumed that the drift condition holds on $\mathcal{X} \setminus \{x^*\}$. An inspection of their proof reveals that we only need $(\boldsymbol{P}V)(x^*) < \infty$ and $\mathbb{E}_x[\lambda^{-\tau^*}] \leq V(x)$ (if the single element drift condition holds, then this follows from Lemma S1). For more general results on the relationship between hitting time and mixing time, see Aldous [1982], Griffiths et al. [2014], Peres and Sousi [2015], Anderson et al. [2019] among many others. These results are mostly developed for finite state spaces.

Theorem S1. *Let $(\mathsf{X}_t)_{t \in \mathbb{N}}, \mathcal{X}, \boldsymbol{P}, \pi$ be as given in Assumption A. Suppose there exist a function $V \colon \mathcal{X} \mapsto [1, \infty)$, a constant $\lambda \in (0, 1)$ and a point $x^* \in \mathcal{X}$ such that $(\boldsymbol{P}V)(x^*) < \infty$ and $\mathbb{E}_x[\lambda^{-\tau^*}] \leq V(x)$ where $\tau^* = \min\{t \geq 0 \colon \mathsf{X}_t = x^*\}$. Then, for every $t \in \mathbb{N}$ and $x \in \mathcal{X}$,*

$$\|\boldsymbol{P}^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq 2V(x)\lambda^{t+1}.$$

Further, $\|\boldsymbol{P}^t(x^*, \cdot) - \pi\|_{\mathrm{TV}} \leq \lambda^{t+1}$ for every $t \in \mathbb{N}$.

Proof. See Jerison [2016, Chapter 4.6] for the proof. □

Remark S1. As shown in Jerison [2016, Chapter 4.6], the assumptions of Theorem S1 imply that π is unique and $\pi(x^*) \geq 1-\lambda$. For high-dimensional model selection problems where x^* is the true model, this yields the rate of strong model selection consistency.

Remark S2. In Jerison's proof of Theorem S1, a key intermediate step is to show that $\|\boldsymbol{P}^t(x^*, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbb{P}_\pi(\tau^* > t)$. Let X_t, $\tilde{\mathsf{X}}_t$ be two Markov chains with transition kernel $\boldsymbol{P}$, $\mathsf{X}_0 = x^*$ and $\tilde{\mathsf{X}}_0 \sim \pi$. By the famous coupling inequality [Pitman, 1976, Lindvall, 2002], we have $\|\boldsymbol{P}^t(x^*, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbb{P}(T > t)$ for $T = \min\{t \geq 0\colon \mathsf{X}_t = \tilde{\mathsf{X}}_t = x^*\}$. So it only remains to couple $\mathsf{X}_t, \tilde{\mathsf{X}}_t$ in such a way that $\tilde{\mathsf{X}}_t = x^*$ implies $\mathsf{X}_t = x^*$. Jerison [2016] finds this coupling (though not explicitly) by using a duality technique, known as "intertwining of Markov chains" [Yor, 1988, Diaconis and Fill, 1990], and a monotonicity result of Lund et al. [2006]. The latter requires that $\boldsymbol{P}$ be reversible and have non-negative spectrum.

S2. Proof of Theorem 2

The outline of the proof of Theorem 2 is as follows. Let $\tau^* = \min\{t \geq 0\colon \mathsf{X}_t = x^*\}$ denote the hitting time of the state x^*. By Theorem S1 in the supplement, all we need is to bound the generating function for τ^*, $\mathbb{E}_x[\alpha^{-\tau^*}]$ for $\alpha \in (0, 1)$. For our problem, directly bounding the generating function seems difficult. So we first find a tail bound instead. To this end, we split the path of (X_t) into disjoint "excursions" in A and A^c (the length of excursion in A^c may be zero). This splitting scheme is the most important step of our proof (see Section S2.1). For each excursion in A, there is some positive probability that the chain can hit x^*, and then we can use a union bound to handle the tail probability of τ^* in the same way as in Theorem 1 of Rosenthal [1995] (see Sections S2.2 and S2.3). Finally, by carefully tuning the parameters in the tail bound for τ^*, we are able to compute its generating function (see Section S2.4).

Remark S3. Roberts and Tweedie [1999, Theorem 2.1] give a bound on the generating function for the regeneration times in the drift-and-minorization setting. For that problem whether the chain regenerates depends on the outcome of an independent coin flip, and thus it is possible to condition on the number of coin flips needed to regenerate. But in our setting, we cannot bound the generating function of τ^* by conditioning on the number of excursions in A needed for the chain to hit x^*, since such conditioning distorts the distribution of (X_t).

S2.1. Path splitting for (X_t)

We first find a decomposition of $\boldsymbol{P}$. Define a transition kernel $\boldsymbol{Q}$ by

$$\boldsymbol{Q}(x, \mathsf{C}) = \frac{\boldsymbol{P}(x, \mathsf{C} \cap \mathsf{A})}{\boldsymbol{P}(x, \mathsf{A})}, \quad \forall\, x \in \mathsf{A},\ \mathsf{C} \in \mathcal{E}.$$

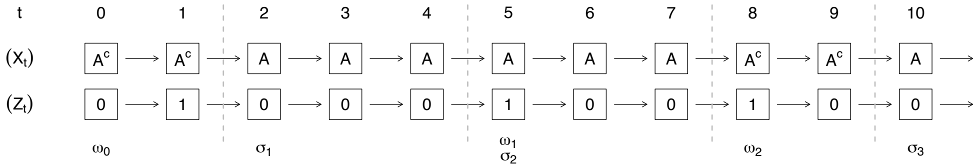

Fig. 2. An example of the evolution of (X_t, Z_t). For (X_t), we only indicate whether it is in A or not. For $t = 5$ and $t = 8$, we generate X_t from $\boldsymbol{R}(X_{t-1}, \cdot)$. Since $X_5 \in \mathsf{A}$, $\omega_1 = \sigma_2 = 5$. For this example, we have $S_0 = 2$, $N_1 = 3$, $S_1 = 0$, $N_2 = 3$ and $S_2 = 2$.

The case $x \notin \mathsf{A}$ is irrelevant to our proof, and one can simply let $\boldsymbol{Q}(x, \cdot) = \boldsymbol{P}(x, \cdot)$ if $x \notin \mathsf{A}$. For $x \in \mathsf{A}$, the distribution $\boldsymbol{Q}(x, \cdot)$ is just $\boldsymbol{P}(x, \cdot)$ conditioned on the chain staying in A. Further, condition (v) implies that $(1-q)\boldsymbol{Q}(x, \cdot) \leq \boldsymbol{P}(x, \cdot)$. Hence, there always exists a "complementary" transition kernel $\boldsymbol{R}$ such that

$$\boldsymbol{P}(x, \cdot) = q\boldsymbol{R}(x, \cdot) + (1-q)\boldsymbol{Q}(x, \cdot), \quad \forall x \in \mathcal{X}. \tag{19}$$

Now we re-construct the Markov chain $(\mathsf{X}_t)_{t \in \mathbb{N}}$. First, we generate a sequence of i.i.d. Bernoulli random variables, $(Z_0, Z_1, \dots)$, such that Z_i is equal to 1 with probability q. Starting with $\mathsf{X}_0 = x \in \mathcal{X}$, we update the chain as follows.

$$\begin{aligned} \text{If } \mathsf{X}_t \in \mathsf{A}^c, \quad & \text{generate } \mathsf{X}_{t+1} \sim \boldsymbol{P}(\mathsf{X}_t, \cdot). \\ \text{If } \mathsf{X}_t \in \mathsf{A},\ Z_{t+1} = 0, \quad & \text{generate } \mathsf{X}_{t+1} \sim \boldsymbol{Q}(\mathsf{X}_t, \cdot). \\ \text{If } \mathsf{X}_t \in \mathsf{A},\ Z_{t+1} = 1, \quad & \text{generate } \mathsf{X}_{t+1} \sim \boldsymbol{R}(\mathsf{X}_t, \cdot). \end{aligned}$$

It follows from (19) that marginally, $(\mathsf{X}_t)_{t \in \mathbb{N}}$ is a Markov chain with transition kernel $\boldsymbol{P}$, and we will use $\mathbb{P}_x$ to denote the corresponding probability measure. Let $(\mathcal{F}_t)_{t \in \mathbb{N}} = \sigma(\mathsf{X}_0, \dots, \mathsf{X}_t, Z_0, \dots, Z_t)$ denote the filtration generated by $(\mathsf{X}_t, Z_t)_{t \in \mathbb{N}}$. Set $\omega_0 = 0$ and then define the following hitting times with respect to $(\mathcal{F}_t)$ recursively.

$$\sigma_k = \min\{t \geq \omega_{k-1} \colon \mathsf{X}_t \in \mathsf{A}\}, \qquad \omega_k = \min\{t > \sigma_k \colon Z_t = 1\}, \qquad k = 1, 2, \dots.$$

Observe that $\mathsf{X}_t \in \mathsf{A}$ if $\sigma_k \leq t \leq \omega_k - 1$. For $k \geq 1$, ω_k marks the k-th time we update the chain using $\boldsymbol{R}$, and then we return to the set A at time σ_{k+1}. Note that $\sigma_{k+1} = \omega_k$ if $\mathsf{X}_{\omega_k} \in \mathsf{A}$ (i.e., the "return" happens immediately.) For $k \geq 1$, let $S_{k-1} = \sigma_k - \omega_{k-1} \geq 0$ and $N_k = \omega_k - \sigma_k \geq 1$. We have

$$\omega_k = S_0 + (N_1 + S_1) + \cdots + (N_{k-1} + S_{k-1}) + N_k. \tag{20}$$

N_k (resp. S_{k-1}) can be seen as the length of the k-th stay of (X_t) in A (resp. A^c). See Figure 2 for a graphical illustration of our path splitting scheme.

S2.2. Tail bound for τ^*

Let $\mathsf{B}_k = (\mathsf{X}_t \neq x^*, \ \text{ for } t = \sigma_k, \dots, \omega_k - 1)$ be the event that (X_t) does not hit x^* during its k-th stay in A. Then, for every t,

$$\mathbb{P}_x(\tau^* > t \geq \omega_k) \leq \mathbb{P}_x(\mathsf{B}_1 \cap \cdots \cap \mathsf{B}_k),$$

since by ω_k, (X_t) has finished k "excursions" in A. As in Rosenthal [1995, Theorem 1], we apply the union bound to get

$$\mathbb{P}_x(\tau^* > t) \leq \mathbb{P}_x(\omega_j > t) + \mathbb{P}_x\left(\mathsf{B}_1 \cap \cdots \cap \mathsf{B}_j\right), \tag{21}$$

which holds for any positive integer j.

The calculation of $\mathbb{P}_x\left(\mathsf{B}_1 \cap \cdots \cap \mathsf{B}_j\right)$ is deferred to the next section. Here we show that $S_0, N_1, S_1, \ldots$ all have geometrically decreasing tails, and thus so does ω_j. First, note that $N_k = \min\{i \geq 1 \colon Z_{\sigma_k+i} = 1\}$, which is just a geometric random variable since (Z_t) is an i.i.d. sequence. Therefore, $\mathbb{E}[u^{N_k}]$ exists for any $u < (1-q)^{-1}$. We choose

$$u = \frac{1}{1 - q/2},$$

which is less than $\min\{\lambda_1^{-1}, \lambda_2^{-1}\}$ by condition (v). It is also evident by construction that N_k is independent of $\mathcal{F}_{\sigma_k}$, which yields

$$\mathbb{E}_x[u^{N_k} \mid \mathcal{F}_{\sigma_k}] = \frac{uq}{1 - u(1-q)} = 2, \quad \text{a.s.} \tag{22}$$

Next, consider the random variables $\{S_k \colon k = 0, 1, \ldots\}$. We have

$$\mathbb{E}_x[u^{S_0}] \leq \mathbb{E}_x[\lambda_1^{-S_0}] \leq V_1(x), \tag{23}$$

by Lemma S1 and drift condition (i). Similarly, for S_k with $k \geq 1$,

$$\mathbb{E}_x[u^{S_k} \mid \mathcal{F}_{\omega_k - 1}] \leq \mathbb{E}_x[V_1(\mathsf{X}_{\omega_k}) \mid \mathcal{F}_{\omega_k-1}] \leq M/2, \quad \text{a.s.} \tag{24}$$

The second inequality follows from condition (iii) and the observation that X_{ω_k} is generated from $\boldsymbol{R}(\mathsf{X}_{\omega_k-1}, \cdot)$ and $\mathsf{X}_{\omega_k-1} \in \mathsf{A}$. Using (20), (22), (23), (24) and conditioning on $\mathcal{F}_{\omega_j-1}, \mathcal{F}_{\sigma_j}, \ldots, \mathcal{F}_{\sigma_1}$ recursively (which is allowed since $\omega_k - \sigma_k \geq 1$), we find that

$$\mathbb{E}_x[u^{\omega_j}] = \mathbb{E}_x[u^{S_0+\cdots+N_j}] \leq 2V_1(x)M^{j-1}. \tag{25}$$

The tail probability $\mathbb{P}_x(\omega_j > t)$ then can be bounded by Markov's inequality.

S2.3. Upper bound for $\mathbb{P}_x(\mathsf{B}_1 \cap \cdots \cap \mathsf{B}_j)$

It is clear that (X_t, Z_t) forms a bivariate Markov chain and thus

$$\mathbb{P}_x\left(\mathsf{B}_k \mid \mathsf{X}_{\sigma_k} = y,\, Z_{\sigma_k} = z\right) = \mathbb{P}_y(\tau^* \geq \omega_1 \mid Z_0 = z) = \mathbb{P}_y(\tau^* \geq \omega_1), \quad \text{a.s.}$$

where $\omega_1 = \min\{t \geq 1 \colon Z_t = 1\}$ because $y = \mathsf{X}_{\sigma_k} \in \mathsf{A}$. Moreover, $\omega_1 = N_1$ is a geometric random variable independent of $\mathcal{F}_0$. Conditioning on N_1, we find

$$\begin{aligned}\mathbb{P}_y(\tau^* \geq N_1) &= \sum_{t=1}^{\infty} \mathbb{P}_y(\tau^* \geq t \mid N_1 = t)(1-q)^{t-1} q \\ &= \sum_{t=1}^{\infty} \mathbb{P}_y(\tau^* \geq t \mid Z_0 = \cdots = Z_{t-1} = 0)(1-q)^{t-1} q,\end{aligned} \tag{26}$$

where in the last step we have used that Z_t is independent of $(\mathsf{X}_0, \dots, \mathsf{X}_{t-1})$ and thus the event $\{\tau^* \geq t\}$. On the event $\{Z_0 = \cdots = Z_{t-1} = 0\}$, $(\mathsf{X}_0, \dots, \mathsf{X}_{t-1})$ is a Markov chain with transition kernel $\boldsymbol{Q}$. For $x \in \mathsf{A} \setminus \{x^*\}$, let $q_x = \mathbb{P}(x, \mathsf{A}^c)$ and write

$$(\boldsymbol{P}V_2)(x) = (1 - q_x)(\boldsymbol{Q}V_2)(x) + q_x \, \mathbb{E}_x[V_2(\mathsf{X}_1) \mid \mathsf{X}_1 \in \mathsf{A}^c].$$

Since $(\boldsymbol{P}V_2)(x) \leq \lambda_2 V_2(x)$ for some $\lambda_2 < 1$, we have $(\boldsymbol{Q}V_2)(x) \leq \lambda_2 V_2(x)$ by condition (iv). This drift condition enables us to apply Lemma S1 and obtain from (26) that

$$\mathbb{P}_y(\tau^* \geq N_1) \leq \sum_{t=1}^{\infty} \lambda_2^t V_2(y)(1-q)^{t-1} q = \frac{V_2(y)\lambda_2 q}{1 - \lambda_2(1-q)},$$

for $y \in \mathsf{A}$. Since $\lambda_2, q \in (0,1)$ and $K \geq \sup_{y \in \mathsf{A}} V_2(y)$,

$$\sup_{y \in \mathsf{A}} \mathbb{P}_y(\tau^* \geq N_1) \leq \frac{qK}{1 - \lambda_2} := \rho. \tag{27}$$

Note that $\rho < 1$ by condition (v). By conditioning on $\mathcal{F}_{\sigma_k}$, we find that $\mathbb{P}_x(\mathsf{B}_k \mid \mathsf{B}_1, \dots, \mathsf{B}_{k-1})$ is also bounded by ρ, and thus $\mathbb{P}_x(\mathsf{B}_1 \cap \cdots \cap \mathsf{B}_j) \leq \rho^j$.

S2.4. Generating function for τ^*

Returning to the bound given in (21), we have

$$\begin{aligned} \mathbb{E}_x[\alpha^{-\tau^*}] &= 1 + (\alpha^{-1} - 1) \sum_{k=0}^{\infty} \mathbb{P}(\tau^* > k)\alpha^{-k} \\ &\leq 1 + (\alpha^{-1} - 1) \sum_{k=0}^{\infty} \alpha^{-k} \left[\rho^{j_k} + \mathbb{P}_x(\omega_{j_k} > k)\right], \end{aligned}$$

for any $\alpha \in (0,1)$, where ρ is as given in (27). We choose $j_k = \lfloor rk + 1 \rfloor \geq rk$ for some $r > 0$. It then follows from (25) and (27) that

$$\mathbb{E}_x[\alpha^{-\tau^*}] \leq 1 + (\alpha^{-1} - 1) \sum_{k=0}^{\infty} \alpha^{-k} \left[\rho^{rk} + \frac{2V_1(x)}{M} \frac{M^{rk}}{u^k}\right].$$

The right-hand side is diverging if either ρ^r or M^r/u is too large. Hence, to obtain the optimal convergence rate, we set

$$r = \frac{\log u}{\log(M/\rho)}, \qquad \text{which yields } \rho^r = \frac{M^r}{u} < 1.$$

Since $M/2 \geq 1$, $0 < \rho < 1$ and $1 < u < 2$, we always have $r \in (0,1)$.

Finally, we choose $\alpha = (1 + \rho^r)/2 < 1$ and find that

$$\mathbb{E}_x[\alpha^{-\tau^*}] \leq 1 + \left(1 + \frac{2V_1(x)}{M}\right) \frac{1 - \alpha}{\alpha - \rho^r} = 2 + \frac{2V_1(x)}{M}.$$

The proof is then completed by applying Theorem S1.

S3. Justification for Condition 1

In this section, we review the high-dimensional assumptions used in Yang et al. [2016] to prove Condition 1.

Theorem S2 (Yang et al. [2016]). *Consider the Bayesian variable selection problem described in Section 2. Suppose the true error variance $\sigma_z^2 = 1$ and the following conditions hold for some finite constants $\mathfrak{C}_0 \geq 0$, $\mathfrak{C}_1 > 0$, $\zeta \in (0, 1]$ such that $\mathfrak{C}_1 \zeta \geq 4$.*

(A) $\|X\beta^*\|_2^2 \leq g \log p$, *and* $\|X_{u^*}\beta^*_{u^*}\|_2^2 \leq \mathfrak{C}_0 \log p$ *where* $u^* = [p] \setminus \gamma^*$.

(B) $X_j^\top X_j = n$ *for each* j, *and*

$$\min_{\gamma \in \mathcal{M}(s_0)} \Lambda_{\min}(X_\gamma^\top X_\gamma) \geq n\zeta,$$

where $\Lambda_{\min}$ *denotes the smallest eigenvalue.*

(C) *For* $z \sim \mathrm{MN}(0, I_n)$,

$$\mathbb{E}\left[\max_{\gamma \in \mathcal{M}(s_0)} \max_{k \notin \gamma} |X_k^\top P_\gamma^\perp z|\right] \leq \frac{1}{2}\sqrt{\mathfrak{C}_1 \zeta n \log p}.$$

(D) $\kappa_0 \geq 2$, $\kappa_1 \geq 1/2$ *and* $\kappa = \kappa_0 + \kappa_1 \geq 4(\mathfrak{C}_0 + \mathfrak{C}_1) + 2$.

(E) *Let* $\Psi(X) = \max_{\gamma \in \mathcal{M}(s_0)} \|(X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top X_{\gamma^* \setminus \gamma}\|_{\mathrm{op}}^2$. *Then,*

$$\max\left\{1,\ (2\zeta^{-2}\Psi(X) + 1)s^*\right\} \leq s_0 \leq \frac{n}{32 \log p} - \frac{\mathfrak{C}_0}{4}.$$

(F) *The threshold* $\beta_{\min}$ *given in* (5) *satisfies*

$$\beta_{\min}^2 \geq \frac{128(\kappa + \mathfrak{C}_0 + \mathfrak{C}_1) \log p}{\zeta^2 n}.$$

Then, with probability $1 - O(p^{-a})$ *for some universal constant* $a > 0$, *Condition 1 holds for* $c_0 = 2$ *and* $c_1 = 4$.

Proof. See Yang et al. [2016, Lemma 4]. Though the original result was stated for $c_0 = 2$ and $c_1 = 3$, replacing $c_1 = 3$ with $c_1 = 4$ does not require any change of their proof. □

Remark S4. As explained in Yang et al. [2016], the assumptions made in Theorem S2 are mild. In particular, Condition (C) holds for $\mathfrak{C}_1 = O(s_0/\zeta)$. A similar result for the empirical normal-inverse-gamma prior is proved in Zhou and Chang [2021, Supplement D].

We note that Theorem S2 holds for any other fixed values of c_0, c_1 under essentially the same assumptions. To explain the reason, we briefly describe below the main idea of the proof of Yang et al. [2016].

Sketch of the proof for Theorem S2. To simplify the discussion, we assume the restricted eigenvalue ζ is a universal constant and $\mathfrak{C}_0 = O(1)$. For two positive sequences a_n, b_n, we write $a_n = \Omega(b_n)$ if $b_n = O(a_n)$, and $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Using $s_0 \log p = O(n)$ and concentration inequalities, one can show that models in $\mathcal{M}(s_0)$ cannot overfit, and thus $y^\top P_\gamma^\perp y = \Omega(n)$ due to the normal noise. Then Condition (A) guarantees that the term $g^{-1} y^\top y = O(\log p)$ in (6) is negligible, and we write

$$B(\gamma, \gamma') = p^{\kappa(|\gamma| - |\gamma'|)} \left\{ 1 + \frac{y^\top (P_{\gamma'} - P_\gamma) y}{y^\top P_{\gamma'}^\perp y + O(\log p)} \right\}^{n/2}.$$

Consider $\gamma' = \gamma \cup \{j\}$ for some overfitted γ and $j \notin \gamma$. Since $\gamma^* \subseteq \gamma$, Condition (A) implies that $y^\top P_{\gamma'}^\perp y = \Theta(n)$, and Condition (C) yields that $y^\top (P_{\gamma'} - P_\gamma) y = O(\mathfrak{C}_1 \log p)$. Hence,

$$B(\gamma, \gamma') = p^{-\kappa} \left\{ 1 + \frac{O(\mathfrak{C}_1 \log p)}{\Theta(n)} \right\}^{n/2}.$$

To prove $B(\gamma, \gamma') \leq p^{-c_0}$ for some $c_0 > 0$, we only need to assume $\kappa \geq a_1 \mathfrak{C}_1 + a_2$ for some sufficiently large constants a_1 and a_2 (and then apply the inequality $1 + x \leq e^x$).

Next, consider $\gamma' = \gamma \cup \{j\}$ for some overfitted γ and $j \in \gamma^* \setminus \gamma$. By Yang et al. [2016, Lemma 8], j can be chosen such that $y^\top (P_{\gamma'} - P_\gamma) y = \Omega(n \beta_{\mathrm{min}}^2)$. (Note that this may not be true for every $j \in \gamma^* \setminus \gamma$.) Therefore, we can write

$$B(\gamma', \gamma) = p^{\kappa} \left\{ 1 - \frac{\Omega(n \beta_{\mathrm{min}}^2)}{y^\top P_\gamma^\perp y} \right\}^{n/2}.$$

Here one needs to consider two possible subcases. If $y^\top P_{\gamma'}^\perp y = \Theta(n)$, then to prove $B(\gamma', \gamma) \leq p^{-c_1}$ for some $c_1 > 0$, we just need to assume $\beta_{\mathrm{min}}^2 \geq a_3 (\kappa + 1) \log p / n$ for some sufficiently large a_3. If $y^\top P_{\gamma'}^\perp y$ has a larger order than n, we need a slightly different argument. By Yang et al. [2016, Lemma 8], we can pick j such that $\gamma' = \gamma \cup \{j\}$ satisfies

$$\frac{y^\top (P_{\gamma'} - P_\gamma) y}{y_{\mathrm{s}}^\top P_\gamma^\perp y_{\mathrm{s}}} = \Omega\left(\frac{1}{s^*}\right),$$

where $y_{\mathrm{s}} = X_{\gamma^*} \beta_{\gamma^*}^*$ denotes the signal part of y. Then, $B(\gamma', \gamma) \leq p^{-c_1}$ would hold if $\kappa s^* \log p \leq n / a_4$ for some sufficiently large a_4. □

S4. Proofs for Section 3

S4.1. Proof of Lemma 1

Proof. Consider part (i) first. Since $y^\top P_\gamma^\perp y \in [0, y^\top y]$, we have

$$1 \leq \frac{g^{-1} y^\top y + y^\top P_\gamma^\perp y}{g^{-1} y^\top y} \leq 1 + g,$$

and thus $V_1(\gamma) \in [1, e]$. The bounds for V_2 are evident since $\gamma \in \mathcal{M}(s_0)$. Part (ii) follows from the fact that $y^\top P_{\gamma \cup \{j\}}^\perp y \leq y^\top P_\gamma^\perp y$ since $P_\gamma^\perp$ projects a vector onto the space orthogonal to the column space of X_γ. For part (iii), we use the following two inequalities,

$$e^{-x} \leq 1 - \frac{x}{2}, \qquad e^x \leq 1 + 2x, \qquad \forall\, x \in [0, 1]. \tag{28}$$

Then, for $k \in \gamma \setminus \gamma^*$, the bound for $R_2(\gamma, \gamma \setminus \{k\})$ follows from the first inequality above, and similarly, for $j \in (\gamma \cup \gamma^*)^c$, the bound for $R_2(\gamma, \gamma \cup \{j\})$ follows from the second. □

S4.2. Drift condition for overfitted models

Recall the weighting functions defined in (12). The corresponding normalizing constants can be expressed by

$$\begin{aligned}
Z_{\mathrm{a}}(\gamma) &= \sum_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)} \left(B(\gamma, \gamma') \wedge p^{c_1} \right), \\
Z_{\mathrm{d}}(\gamma) &= \sum_{\gamma' \in \mathcal{N}_{\mathrm{d}}(\gamma)} \left(1 \vee B(\gamma, \gamma') \wedge p^{c_0} \right), \\
Z_{\mathrm{s}}(\gamma) &= \sum_{\gamma' \in \mathcal{N}_{\mathrm{s}}(\gamma)} \left(p s_0 \vee B(\gamma, \gamma') \wedge p^{c_1} \right).
\end{aligned}$$

Under Condition 1, we can bound $Z_\star(\gamma)$ for an overfitted γ as follows.

Lemma S2. *Suppose Condition 1 holds and $\gamma \in \mathcal{M}(s_0)$ is an overfitted model.*

(i) $Z_{\mathrm{a}}(\gamma) \leq p^{1-c_0}$.

(ii) For any $k \in \gamma$, $w_{\mathrm{d}}(\gamma \setminus \{k\} \mid \gamma) = p^{c_0}$ if $k \in \gamma \setminus \gamma^*$, and 1 if $k \in \gamma^*$.

(iii) $Z_{\mathrm{d}}(\gamma) = (|\gamma| - s^*) p^{c_0} + s^*$.

(iv) For any $j \notin \gamma$ and $k \in \gamma^*$, $B(\gamma, (\gamma \cup \{j\}) \setminus \{k\}) \leq p^{-(c_0 + c_1)}$.

Proof. For any overfitted model γ with $|\gamma| \leq s_0$, by Condition (1a),

$$Z_{\mathrm{a}}(\gamma) = \sum_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)} B(\gamma, \gamma') \wedge p^{c_1} \leq |\mathcal{N}_{\mathrm{a}}(\gamma)| p^{-c_0} \leq p^{1-c_0},$$

since any γ' in $\mathcal{N}_{\mathrm{a}}(\gamma)$ is obtained by adding a non-influential covariate to γ.

To prove part (ii), note that for any $k \in \gamma \setminus \gamma^*$, we have

$$B(\gamma, \gamma \setminus \{k\}) = B(\gamma \setminus \{k\}, \gamma)^{-1} \geq p^{c_0},$$

since $\gamma \setminus \{k\}$ is still overfitted. If $k \in \gamma^*$, then $\gamma' = \gamma \setminus \{k\}$ is underfitted and $\gamma^* \setminus \gamma = \{k\}$. Hence, by Condition (1b), $B(\gamma, \gamma') \leq p^{-c_1} < 1$. Part (iii) follows from (ii).

Consider a swap move. For any $j \notin \gamma$ and $k \in \gamma^* \subset \gamma$, we have

$$B(\gamma, (\gamma \cup \{j\}) \setminus \{k\}) = B(\gamma, \gamma \cup \{j\}) B(\gamma \cup \{j\}, (\gamma \cup \{j\}) \setminus \{k\}) \leq p^{-(c_0 + c_1)},$$

since $B(\gamma, \gamma \cup \{j\}) \leq p^{-c_0}$ by Condition (1a) and $B(\gamma \cup \{j\}, (\gamma \cup \{j\}) \setminus \{k\}) \leq p^{-c_1}$ by Condition (1b). Part (iv) then follows. □

The drift condition provided in Proposition 1 follows from the following lemma.

Lemma S3. *Suppose that Condition 1 holds for some $c_0 \geq 2$ and $c_1 \geq 1$. For any overfitted model γ such that $\gamma \neq \gamma^*$ and $|\gamma| \leq s_0$,*

$$\sum_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)} R_2(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq \frac{1}{s_0 p^{c_0 - 1}},$$
$$\sum_{\gamma' \in \mathcal{N}_{\mathrm{d}}(\gamma)} R_2(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq -\frac{1}{4 s_0} + O\left(\frac{1}{s_0 p^{c_0 - 1}}\right),$$
$$\sum_{\gamma' \in \mathcal{N}_{\mathrm{s}}(\gamma)} R_2(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq p^{-c_0}.$$

Proof. Consider addition first. Since γ is overfitted, we can only add non-influential covariates. For any $j \notin \gamma$, it follows from (17) and Condition (1a) that

$$\boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma \cup \{j\}) \leq \frac{B(\gamma, \gamma \cup \{j\})}{2} \leq \frac{1}{2 p^{c_0}}.$$

Thus, using Lemma 1(iii) we obtain that

$$\sum_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)} R_2(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq \frac{(p - |\gamma|)}{s_0 p^{c_0}} \leq \frac{1}{s_0 p^{c_0 - 1}}.$$

Consider deletion moves. Observe that V_2 only changes if we remove a non-influential covariate. For any $k \in \gamma \setminus \gamma^*$, by Condition (1a), and Lemma S2(i),

$$B(\gamma, \gamma \setminus \{k\}) \boldsymbol{K}_{\mathrm{lit}}(\gamma \setminus \{k\}, \gamma) = \frac{1}{2 Z_{\mathrm{a}}(\gamma \setminus \{k\})} \geq \frac{p^{c_0 - 1}}{2} > 1.$$

Thus, we find by applying Lemma 1(ii) and (17) that

$$-R_2(\gamma, \gamma \setminus \{k\}) \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma \setminus \{k\}) \geq \frac{\boldsymbol{K}_{\mathrm{lit}}(\gamma, \gamma \setminus \{k\})}{2 s_0} = \frac{p^{c_0}}{4 s_0 \left[(|\gamma| - s^*) p^{c_0} + s^*\right]}.$$

Since $c_0 \geq 2$ and there are $(|\gamma| - s^*)$ non-influential covariates that we may remove,

$$-\sum_{\gamma' \in \mathcal{N}_{\rm d}(\gamma)} R_2(\gamma, \gamma') \boldsymbol{P}_{\rm lit}(\gamma, \gamma') \geq \frac{(|\gamma| - s^*)(4s_0)^{-1}}{|\gamma| - s^* + s^* p^{-c_0}} = \frac{1}{4s_0} + O\left(\frac{1}{s_0 p^{c_0 - 1}}\right).$$

In the last step we have used that $|\gamma| - s^* \geq 1$, $s^* < p$, and $(1+x)^{-1} \sim 1 - x$ for $x = o(1)$.

For the swap moves, note that V_2 only changes if we swap an influential covariate $k \in \gamma^*$ with a non-influential covariate $j \notin \gamma$. The total number of such pairs is $(p - s_0)s^*$. Let $\gamma' = (\gamma \cup \{j\}) \setminus \{k\}$ denote the resulting model. By Lemma 1(iii), Lemma S2(iv) and (17),

$$R_2(\gamma, \gamma') \boldsymbol{P}_{\rm lit}(\gamma, \gamma') \leq \frac{1}{s_0} B(\gamma, \gamma') \leq \frac{1}{s_0 p^{c_0 + c_1}}.$$

Since $(p - s_0)s^* \leq p s_0$,

$$\sum_{\gamma' \in \mathcal{N}_{\rm s}(\gamma)} R_2(\gamma, \gamma') \boldsymbol{P}_{\rm lit}(\gamma, \gamma') \leq \frac{1}{p^{c_0 + c_1 - 1}} \leq \frac{1}{p^{c_0}},$$

which concludes the proof. □

S4.3. Drift condition for underfitted models

We first prove two auxiliary results.

Lemma S4. *Suppose Condition 1 holds and $\gamma \in \mathcal{M}(s_0)$ is an underfitted model.*

(i) $Z_{\rm a}(\gamma) \geq p^{c_1}$.

(ii) $s_0 \leq Z_{\rm d}(\gamma) \leq s_0 p^{c_0}$.

(iii) If $|\gamma| = s_0$, *then* $p^{c_1} \leq Z_{\rm s}(\gamma) \leq s_0 p^{c_1 + 1}$.

Proof. By Condition (1b), there exists some j^* such that $B(\gamma, \gamma \cup \{j^*\}) \geq p^{c_1}$, which proves part (i). Part (ii) follows from the definition of $w_{\rm d}$ and that $|\mathcal{N}_{\rm d}(\gamma)| = |\gamma| \leq s_0$. A similar argument using Condition (1c) and $|\mathcal{N}_{\rm s}(\gamma)| \leq p s_0$ proves part (iii). □

Lemma S5. *Suppose that $B(\gamma, \gamma') \geq p^a$ for some $a \in \mathbb{R}$ and define*

$$b = \frac{a + \kappa(|\gamma'| - |\gamma|)}{n \kappa_1}.$$

If $b \in [0, 1]$, then $-R_1(\gamma, \gamma') \geq b/2$. If $b \in [-1, 0]$, then $R_1(\gamma, \gamma') \leq -2b$.

Proof. First, by (4) and the definition of V_1, we have

$$\log\left\{p^{\kappa(|\gamma'| - |\gamma|)} B(\gamma, \gamma')\right\} = -\frac{n \log(1 + g)}{2} \log \frac{V_1(\gamma')}{V_1(\gamma)}.$$

Using $1 + g = p^{2\kappa_1}$ and $R_1(\gamma, \gamma') = V_1(\gamma')/V_1(\gamma) - 1$, we obtain that

$$\frac{\log B(\gamma, \gamma')}{\log p} = \kappa(|\gamma| - |\gamma'|) - n\kappa_1 \log[1 + R_1(\gamma, \gamma')].$$

Then, $B(\gamma, \gamma') \geq p^a$ implies that

$$-R_1(\gamma, \gamma') \geq 1 - \exp\left\{-\frac{a + \kappa(|\gamma'| - |\gamma|)}{n\kappa_1}\right\} = 1 - e^{-b}.$$

Applying the two inequalities in (28) yields the result. □

The drift condition provided in Proposition 2 follows from the following lemma.

Lemma S6. *Suppose Condition 1 holds and $\gamma \in \mathcal{M}(s_0)$ is underfitted.*

(i) We always have

$$0 \leq \sum_{\gamma' \in \mathcal{N}_{\mathrm{d}}(\gamma)} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq \frac{|\gamma|(e-1)}{2p^{c_1}}.$$

(ii) If $c_1 \geq (c_0 + 1) \vee 2$ and $\kappa + c_1 \leq n\kappa_1$,

$$-\sum_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \geq \frac{\kappa + c_1}{8n\kappa_1}.$$

(iii) If $|\gamma| = s_0$, $4 \leq c_1 \leq n\kappa_1$, $n = O(p)$, $\kappa = O(s_0)$ and $s_0 \log p = O(n)$,

$$-\sum_{\gamma' \in \mathcal{N}_{\mathrm{s}}(\gamma)} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \geq \frac{c_1}{8n\kappa_1} + o\left(\frac{1}{n\kappa_1}\right).$$

Proof of part (i) (deletion). By Lemma S4(i), we have

$$B(\gamma, \gamma \setminus \{k\}) \boldsymbol{K}_{\mathrm{lit}}(\gamma \setminus \{k\}, \gamma) = \frac{B(\gamma, \gamma \setminus \{k\}) w_{\mathrm{a}}(\gamma \mid \gamma \setminus \{k\})}{2Z_{\mathrm{a}}(\gamma \setminus \{k\})} \leq \frac{1}{2p^{c_1}},$$

since $B(\gamma, \gamma') w_{\mathrm{a}}(\gamma' \mid \gamma) \leq 1$ for any $\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)$. It then follows from (17) that

$$R_1(\gamma, \gamma \setminus \{k\}) \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma \cup \{k\}) \leq \frac{R_1(\gamma, \gamma \setminus \{k\})}{2p^{c_1}}.$$

By Lemma 1(i), $R_1(\gamma, \gamma \setminus \{k\}) \leq e - 1$, from which the asserted bound follows. □

Proof of part (ii) (addition). Define a set of "good" addition moves as

$$\mathcal{G} = \mathcal{G}(\gamma) = \{\gamma \cup \{j\} \colon j \notin \gamma,\, B(\gamma, \gamma \cup \{j\}) \geq p^{c_1 - 1}\}.$$

By Condition (1b), $\mathcal{G}$ contains at least one element, which we denote by $\mathcal{T}(\gamma)$, such that $B(\gamma, \mathcal{T}(\gamma)) \geq p^{c_1}$. By Lemma S5 and the assumption on c_1,

$$-R_1(\gamma, \mathcal{T}(\gamma)) \geq \frac{\kappa + c_1}{2n\kappa_1} =: A. \tag{29}$$

Using Lemma S5 again and the assumption that $c_1 \geq 2$,

$$- R_1(\gamma, \gamma') \geq \frac{\kappa + c_1 - 1}{2n\kappa_1} \geq \frac{A}{2}, \quad \forall \gamma' \in \mathcal{G}. \tag{30}$$

Further, for any $\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)$, $\boldsymbol{K}_{\mathrm{lit}}(\gamma', \gamma) \geq (2s_0 p^{c_0})^{-1}$ by Lemma S4(ii). It then follows from (17) that

$$\boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \geq \min\left\{ \frac{w_{\mathrm{a}}(\gamma' \mid \gamma)}{2Z_{\mathrm{a}}(\gamma)}, \frac{B(\gamma, \gamma')}{2s_0 p^{c_0}} \right\} \geq \frac{w_{\mathrm{a}}(\gamma' \mid \gamma)}{2} \min\left\{ \frac{1}{Z_{\mathrm{a}}(\gamma)}, \frac{1}{s_0 p^{c_0}} \right\}.$$

By Lemma S4(i) and the assumption $c_1 \geq c_0 + 1$, we have $Z_{\mathrm{a}}(\gamma) \geq p^{c_1} \geq s_0 p^{c_0}$. Using the above displayed inequality, we obtain that

$$\boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \geq \frac{w_{\mathrm{a}}(\gamma' \mid \gamma)}{2Z_{\mathrm{a}}(\gamma)}, \quad \forall \gamma' \in \mathcal{G}. \tag{31}$$

Define $\mathcal{G}' = \mathcal{G} \setminus \{\mathcal{T}(\gamma)\}$, which may be empty. Let $W = \sum_{\gamma' \in \mathcal{G}'} w_{\mathrm{a}}(\gamma' \mid \gamma)$. Then,

$$\begin{aligned} Z_{\mathrm{a}}(\gamma) &= \sum\nolimits_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)} w_{\mathrm{a}}(\gamma' \mid \gamma) \\ &= p^{c_1} + W + \sum\nolimits_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma) \setminus \mathcal{G}(\gamma)} w_{\mathrm{a}}(\gamma' \mid \gamma) \\ &\leq W + 2p^{c_1}, \end{aligned} \tag{32}$$

since for any $\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma) \setminus \mathcal{G}(\gamma)$, we have $B(\gamma, \gamma') < p^{c_1 - 1}$. By Lemma 1(ii), $R_1(\gamma, \gamma \cup \{j\}) \leq 0$ for any $j \notin \gamma$, which implies that

$$\sum_{\gamma' \in \mathcal{N}_{\mathrm{a}}(\gamma)} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq \sum_{\gamma' \in \mathcal{G}} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma').$$

Some algebra using (29), (30), (31) and (32) yields that

$$- \sum_{\gamma' \in \mathcal{G}} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \geq \frac{A p^{c_1}}{2Z_{\mathrm{a}}(\gamma)} + \sum_{\gamma' \in \mathcal{G}'} \frac{A w_{\mathrm{a}}(\gamma' \mid \gamma)}{4Z_{\mathrm{a}}(\gamma)} \geq \frac{A}{4}, \tag{33}$$

which concludes the proof. □

Proof of part (iii) (swap). First, we use an argument similar to the proof of part (ii) to analyze those "good" moves. Define

$$\mathcal{G}_1(\gamma) = \left\{ \gamma' \in \mathcal{N}_{\mathrm{s}}(\gamma) \colon B(\gamma, \gamma') \geq \frac{p^{c_1}}{p s_0} \right\}.$$

By Condition (1c), there exists $\mathcal{T}(\gamma) \in \mathcal{G}_1(\gamma)$ such that $B(\gamma, \mathcal{T}(\gamma)) \geq p^{c_1}$. By Lemma S5,

$$-R_1(\gamma, \mathcal{T}(\gamma)) \geq \frac{c_1}{2n\kappa_1}.$$

Similarly, for any $\gamma' \in \mathcal{G}_1(\gamma)$,

$$-R_1(\gamma, \gamma') \geq \frac{(c_1 - 1) - (\log s_0)/(\log p)}{2n\kappa_1} \geq \frac{c_1 - 2}{2n\kappa_1} \geq \frac{c_1}{4n\kappa_1},$$

since $c_1 \geq 4$. By Lemma S4(iii), $Z_s(\gamma') \leq s_0 p^{c_1+1}$. Further, since $c_1 \geq 4$, for any $\gamma' \in \mathcal{G}_1(\gamma)$, $B(\gamma, \gamma') \geq w_s(\gamma' \mid \gamma)$. Applying (17), we obtain that

$$\boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \geq w_s(\gamma' \mid \gamma) \min\left\{\frac{1}{2Z_s(\gamma)}, \frac{ps_0}{2s_0 p^{c_1+1}}\right\} = \frac{w_s(\gamma' \mid \gamma)}{2Z_s(\gamma)}, \quad \forall \gamma' \in \mathcal{G}_1.$$

By calculations similar to (32) and (33), we find that

$$Z_s(\gamma) \leq 2p^{c_1} + \sum_{\gamma' \in \mathcal{G}_1 \setminus \{\mathcal{T}(\gamma)\}} w_s(\gamma' \mid \gamma),$$

which yields

$$-\sum_{\gamma' \in \mathcal{G}_1} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \geq \frac{c_1}{8n\kappa_1}. \tag{34}$$

Let $\mathcal{G}_2(\gamma) = \{\gamma' \in \mathcal{N}_s(\gamma) \colon B(\gamma, \gamma') < 1\}$. We claim that any $\gamma' \in \mathcal{G}_2$ is still underfitted. If $\gamma' \in \mathcal{G}_2$ is overfitted, $\gamma' = (\gamma \cup \{j\}) \setminus \{k\}$ for some $j \in \gamma^* \setminus \gamma$ and $k \in \gamma \setminus \gamma^*$. By Lemma S2(iv), $B(\gamma', \gamma) \leq p^{-(c_0+c_1)}$, which yields the contradiction. Since $\gamma' \in \mathcal{G}_2$ is underfitted, $Z_s(\gamma') \geq p^{c_1}$ by Lemma S4(iii). Consequently,

$$\boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq B(\gamma, \gamma') \boldsymbol{K}_{\mathrm{lit}}(\gamma', \gamma) \leq B(\gamma, \gamma') \frac{w_s(\gamma \mid \gamma')}{2p^{c_1}}.$$

If $B(\gamma, \gamma') < (ps_0)^{-1}$, then $B(\gamma, \gamma') w_s(\gamma \mid \gamma') \leq 1$, and thus, by Lemma 1(i),

$$R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq \frac{e - 1}{2p^{c_1}}.$$

If $B(\gamma, \gamma') \in [(ps_0)^{-1}, 1)$, $w_s(\gamma \mid \gamma') = ps_0$, and by Lemma S5,

$$R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq \frac{2\log(ps_0)}{n\kappa_1 \log p} \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \leq \frac{s_0 \log(ps_0)}{n\kappa_1 p^{c_1-1} \log p} \leq \frac{2s_0}{n\kappa_1 p^{c_1-1}}.$$

Using the assumptions $c_1 \geq 4$, $n = O(p)$ and $\kappa = O(s_0)$, we get

$$R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') = O\left(\frac{s_0}{n\kappa_1 p^3}\right). \tag{35}$$

If $\gamma' \in \mathcal{N}_s(\gamma) \setminus \mathcal{G}_2(\gamma)$, then $R_1(\gamma, \gamma') \leq 0$. Hence, it follows from (34) and (35) that

$$\begin{aligned} \sum_{\gamma' \in \mathcal{N}_s(\gamma)} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') &\leq \sum_{\gamma' \in \mathcal{G}_1} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') + \sum_{\gamma' \in \mathcal{G}_2} R_1(\gamma, \gamma') \boldsymbol{P}_{\mathrm{lit}}(\gamma, \gamma') \\ &\leq -\frac{c_1}{8n\kappa_1} + O\left(\frac{s_0^2}{n\kappa_1 p^2}\right). \end{aligned}$$

Finally, notice that $s_0 \log p = O(n)$ and $n = O(p)$ imply that $s_0 = o(p)$. Hence, the asymptotic term in the last step is $o((n\kappa_1)^{-1})$. □

S4.4. Proof of Theorem 1

Proof. To apply the results provided in Section 4, we need to consider the lazy version of the transition matrix $\boldsymbol{P}_{\text{lit}}$, $\boldsymbol{P}_{\text{lazy}} = (\boldsymbol{P}_{\text{lit}} + \boldsymbol{I})/2$. But this is equivalent to dividing all the proposal probabilities by 2. Hence, by Propositions 1 and 2, for the lazy chain we have

$$\lambda_1 = 1 - \frac{c_1}{16n\kappa_1} + o((n\kappa_1)^{-1}), \qquad \lambda_2 = 1 - \frac{1}{8s_0} + o(s_0^{-1}).$$

For sufficiently large n, we can assume that

$$\lambda_1 \leq 1 - \frac{c_1}{20n\kappa_1}, \qquad \lambda_2 \leq 1 - \frac{1}{10s_0}.$$

Let q be the probability of the chain escaping from the set $\mathcal{O}(\gamma^*, s_0)$. By Condition (1b) and Lemma S2(iv), $q \leq p^{-c_0} \leq p^{-2}$. Since $\kappa_1 \leq \kappa = O(s_0) = o(p)$, we have $q = o((1-\lambda_1) \wedge (1-\lambda_2))$. By Lemma 1(i), $\sup_{\gamma \in \mathcal{O}} V_2(\gamma) \leq e$ and $\sup_{\gamma \in \mathcal{M}(s_0)} V_1(\gamma) \leq e$. Thus, we may assume $\boldsymbol{P}_{\text{lazy}}$ satisfies that assumptions of Corollary 1 with $K = e$, $M = 2e$ and $C = 6/e$ (other values of C will yield slightly different constants in the bound). The asserted upper bound on the mixing time then follows from a routine calculation using Corollary 2. □

S5. Data and code availability

The two GWAS data sets used in Section 6 are the Primary Open-Angle Glaucoma Genes and Environment (GLAUGEN) Study (accession no. phs000308.v1.p1) and the National Eye Institute Glaucoma Human Genetics Collaboration (NEIGHBOR) Consortium Glaucoma Genome-Wide Association Study (accession no. phs000238.v1.p1). Both can be obtained from dbGaP (`https://www.ncbi.nlm.nih.gov/gap/`). The genotype data of both studis were generated using the Illumina Human660W-Quad_v1_A beadchip. The code used for simulation studies described in Section 5 is available at `https://web.stat.tamu.edu/~quan/code/lit_mh.tgz`.

References

David J Aldous. Some inequalities for reversible Markov chains. *Journal of the London Mathematical Society*, 2(3):564–576, 1982.

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

Persi Diaconis and James Allen Fill. Strong stationary times via a new form of duality. *The Annals of Probability*, pages 1483–1522, 1990.

Simon Griffiths, Ross Kang, Roberto Oliveira, and Viresh Patel. Tight inequalities among set hitting times in Markov chains. *Proceedings of the American Mathematical Society*, 142(9): 3285–3298, 2014.

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

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Robert Lund, Ying Zhao, and Peter C Kiessler. A monotonicity in reversible Markov chains. *Journal of Applied Probability*, 43(2):486–499, 2006.

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

JW Pitman. On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35(4):315–322, 1976.

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

Marc Yor. *Intertwinings of Bessel processes*. Department of Statistics, University of California, 1988.

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