COMPLEXITY RESULTS FOR MCMC DERIVED FROM QUANTITATIVE BOUNDS

JUN YANG AND JEFFREY S. ROSENTHAL

Abstract. This paper considers whether MCMC quantitative convergence bounds can be translated into complexity bounds. We prove that a certain realistic Gibbs sampler algorithm converges in constant number of iterations. Our proof uses a new general method of establishing a generalized geometric drift condition defined in a subset of the state space. The subset is chosen to rule out some "bad" states which have poor drift property when the dimension gets large. Using the new general method, the obtained quantitative bounds for the Gibbs sampler algorithm can be translated to tight complexity bounds in high-dimensional setting. It is our hope that the new general approach can be employed in many other specific examples to obtain complexity bounds for high-dimensional Markov chains.

Contents

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are extremely widely used and studied in statistics, e.g. [Bro+11; GRS95], and their running times are an extremely important practical issue. They have been studied from a variety of perspectives, including convergence "diagnostics" via the Markov chain output (e.g. [GR92]), proving weak convergence limits of speed-up versions of the algorithms to diffusion limits [RGG97; RR98], and directly bounding the convergence in total variation distance [MT94; Ros95a; Ros96; RT99; JH01; Ros02; JH04; Bax05; FHJ08].

Department of Statistical Sciences, University of Toronto, Canada
E-mail addresses: jun@utstat.toronto.edu, jeff@math.toronto.edu.

Among the work of directly bounding the total variation distance, most of the quantitative convergence bounds proceed by establishing a *drift condition* and an associated *minorization condition* for the Markov chain in question (see e.g. [MT12] and its first edition in 1993). The most widely employed approach for finding quantitative bounds has been the drift and minorization method set forth by Rosenthal [Ros95a].

Computer scientists take a slightly different perspective, in terms of running time complexity order as the dimension goes to infinity. Complexity results in computer science go back at least to Cobham [Cob65], and took on greater focus with the pioneering NP-complete work of Cook [Coo71]. In the Markov chain context, computer scientists have been bounding convergence times of Markov chain algorithms since at least Sinclair and Jerrum [SJ89], focusing largely on spectral gap bounds for Markov chains on finite state spaces. More recently, attention has turned to bounding spectral gaps of modern Markov chain algorithms on general state spaces, again primarily via spectral gaps, such as [WSH09a; WSH09b]. These bounds often focus on the order of the convergence time in terms of some particular parameter, such as the dimension of the corresponding state space. In recent years, there is much interest in the "large p, large n" or "large p, small n" high-dimensional setting, where p is the number of parameters and n is the sample size. Rajaratnam and Sparks [RS15] use the term convergence complexity to denote the ability of a high-dimensional MCMC scheme to draw samples from the posterior, and how the ability to do so changes as the dimension of the parameter set grows.

Direct total variation bounds for MCMC are sometimes presented in terms of the convergence order, for example, the work by Rosenthal [Ros95b] for a Gibbs sampler for a variance components model. However, current methods for obtaining total variation bounds of such MCMCs typically proceed as if the dimension of the parameter, p, and sample size, n, are fixed. Perhaps because of this, they are often overlooked by the computer science complexity community. Actually, this has caused them to claim (e.g. by Yang, Wainwright, and Jordan [YWJ16]) that little is known about MCMC complexity. It is thus important to bridge the gap between statistics-style convergence bounds, and computer-science-style complexity results.

In one direction, Roberts and Rosenthal [RR16] connect known results about diffusion limits of MCMC to the computer science notion of algorithm complexity. They show that any weak limit of a Markov process implies a corresponding complexity bound in an appropriate metric. For example, under appropriate assumptions, in p dimensions, the Random-Walk Metropolis algorithm takes $\mathcal{O}(p)$ iterations and the

Metropolis-Adjusted Langevin Algorithm takes $\mathcal{O}(p^{1/3})$ iterations to converge to stationarity.

This paper considers whether MCMC quantitative convergence bounds can be translated into complexity bounds. At the first glance, it may seem that an approach to answering the question of convergence complexity may be provided by the method of [Ros95a], since it is the most widely used approach to obtain upper bounds on the total variation distance. However, Rajaratnam and Sparks [RS15] demonstrate that somewhat problematically, such upper bounds tend to 1 as n or p tends to infinity. It is therefore their hope that proposals and developments of new ideas analogous to those of [Ros95a], which are suitable for high-dimensional settings, can be motivated.

In this paper, we attempt to address this concern on how to obtain quantitative bounds that can be translated into tight complexity bounds. We prove that a certain realistic Gibbs sampler algorithm converges in $\mathcal{O}(1)$ iterations. Be more specific, we prove that when the dimension of the model is large, the number of iterations which guarantees small distance of the Gibbs sampler to stationarity is upper bounded by some constant which does not depends on the dimension of the model; see Theorem 3.3. Our proof uses a new general method of establishing a generalized drift condition for a subset of the state space; see Section 2. The generalized geometric drift condition is defined in a "large set" of states, instead of the whole state space. The "large set" is chosen to rule out some "bad" states which have poor drift property when the dimension gets large. As a result, the new quantitative bound is composed of two parts. The first part is an upper bound on the probability the Markov chain will visit the states outside of the "large set"; the second part is an upper bound on the total variation distance of a new Markov chain defined only on the "large set" using the conditional transition kernel. In order to obtain good complexity bounds for high-dimensional setting, the "large set" should be adjusted depending on n and p to balance the complexity order of the two parts. It is our hope that this general method of proof in Section 2 can be employed to other specific examples for obtaining quantitative bounds that can be translated to complexity bounds in high-dimensional setting.

2. Generalized Geometric Drift Conditions and Large Sets

Scaling classical MCMCs to very high dimensions can be problematic. Even the chain is indeed geometrically ergodic for fixed n and p, the convergence of Markov chains may still be quite slow as $p \to \infty$ and

$n \to \infty$. Rajaratnam and Sparks [RS15] demonstrate that the drift and minorization method [Ros95a] has problems when the dimension grows. For more details, recall that for a Markov chain $P(x, \cdot)$ on a state space $\mathcal{X}$, the general method of [Ros95a] proceeds by establishing a *drift condition*

$$\mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x) \leq \lambda f(x) + b, \quad \forall x \in \mathcal{X}, \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}^+$ is the "drift function", some $0 < \lambda < 1$ and $b < \infty$; and an associated *minorization condition*

$$P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R, \tag{2}$$

where $R := \{x \in \mathcal{X} : f(x) \leq d\}$ and $d > 2b/(1 - \lambda)$, for some $\epsilon > 0$ and some probability measure $Q(\cdot)$ on $\mathcal{X}$. By directly translating the existing work by Choi and Hobert [CH13] and Khare and Hobert [KH13], which are both based on the general approach of [Ros95a], Rajaratnam and Sparks [RS15] show that the "small set" R gets large fast as the dimension p increases. This leads the minorization volume ϵ goes to zero exponentially fast as p goes to infinity. As a result, the directly translated upper bound on the mixing time of the Markov chain increases exponentially fast as the dimension p goes to infinity. Also, it could be very hard to find alternative drift functions that lead to good complexity bounds. For a given drift function, when the dimension gets larger, the typical scenario for the drift condition of Eq. (1) seems to be λ going to one, and/or b getting much larger.

We consider the difficulty for finding a good drift function suitable for high-dimensional setting is caused by the existence of some "bad" states when the dimension gets large. A "bad" state is a state that has a large drift function value, but the drift property becomes poor as the dimension gets large. Therefore, once the Markov chain stays in one of these "bad" states, it drifts back toward to the "small set" slowly in high-dimensional cases. Since the definition of drift condition in Eq. (1) requires that all $x \in \mathcal{X}$ need to satisfy, in the existence of "bad" states when dimension goes large, λ has to be close to 1 and/or b has to be large in Eq. (1). This leads to the "small set", R, too big, hence the minorization volume ϵ goes to zero very fast. From another direction, we should consider this is because the definition of drift condition in Eq. (1) is restrictive.

In this section, we propose a new general approach using a generalized drift condition, where the drift function is defined only in a "large set" of states, instead of the whole state space. The goal of choosing the "large set" is to rule out those "bad" states in high-dimensional cases, so that the obtained quantitative bound can be translated into good complexity bound. The new general method is summarized in the following theorem.

Theorem 2.1. *Suppose that a Markov chain $P(x,\cdot)$ on a state space $(\mathcal{X},\mathcal{B})$. Let $R_0 \in \mathcal{B}$ be a subset of $\mathcal{X}$ and π be the stationary distribution. Suppose that*

$$\begin{aligned}\mathbb{E}(f(X^{(1)})\,|\,X^{(0)}=x, X^{(1)}\in R_0) &\le \mathbb{E}(f(X^{(1)})\,|\,X^{(0)}=x)\\ &\le \lambda f(x)+b, \quad \forall x\in R_0,\end{aligned} \tag{3}$$

for some $f:\mathcal{X}\to\mathbb{R}^+$ and some $\lambda<1$ and $b<\infty$. For $R := \{x\in\mathcal{X} : f(x)\le d\}$ where $d>2b/(1-\lambda)$, if the Markov chain also satisfies a minorization condition:

$$P(x,\cdot)\ge \epsilon Q(\cdot), \quad \forall x\in R, \tag{4}$$

for some $\epsilon>0$, some probability measure $Q(\cdot)$ on $\mathcal{X}$. Then for any $0<r<1$, beginning with any initial distribution ν such that $\nu(R_0)=1$, we have

$$\begin{aligned}\|\mathcal{L}(X^{(k)})-\pi\|_{\mathrm{var}} &\le (1-\epsilon)^{rk} + \frac{(\alpha\Lambda)^{rk}\left[1+\mathbb{E}_\nu(f(x))+\frac{b}{1-\lambda}\right]-\alpha^{rk}}{\alpha^k-\alpha^{rk}}\\ &\quad + k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c),\end{aligned} \tag{5}$$

where $\alpha^{-1}=\frac{1+2b+\lambda d}{1+d}$ and $\Lambda = 1+2(\lambda d+b)$.

Proof. See Section 4. $\square$

Remark 2.2. If $R_0=\mathcal{X}$, then Theorem 2.1 reduces to

$$\|\mathcal{L}(X^{(k)})-\pi\|_{\mathrm{var}} \le (1-\epsilon)^{rk} + \frac{(\alpha\Lambda)^{rk}\left[1+\mathbb{E}_\nu(f(x))+\frac{b}{1-\lambda}\right]-\alpha^{rk}}{\alpha^k-\alpha^{rk}}. \tag{6}$$

This bound is almost the same as [Ros95a, Theorem 12], except slightly tighter due to the terms α^{rk}. $\triangleleft$

Remark 2.3. For the new general approach in Theorem 2.1, we have the following comments:

- The new drift condition of Eq. (3) involves checking inequality $\mathbb{E}(f(X^{(1)})\,|\,X^{(0)}=x, X^{(1)}\in R_0)\le \mathbb{E}(f(X^{(1)})\,|\,X^{(0)}=x)$. This implies the "large set" R_0 should be chosen such that the states in R_0 on expectation have smaller values of the drift function.

- The condition $\mathbb{E}(f(X^{(1)})\,|\,X^{(0)} = x, X^{(1)} \in R_0) \le \lambda f(x) + b, \forall x \in R_0$ essentially defines a traditional drift condition in Eq. (1) for a new Markov chain only on the "large set" R_0, using the *conditional* transition kernel $P'(x, B) := P(x, B \cap R_0)/P(x, R_0), \forall x \in R_0, B \in \mathcal{B}$. The first two terms in the upper bound Eq. (5) is indeed an upper bound on the total variation distance of this new Markov chain.
- The last two terms in the upper bound Eq. (5)) is an upper bound of the probability that the Markov chain will visit R_0^c starting from both the initial distribution ν and the stationary distribution π.
- Overall, the "large set" should be chosen to rule out the "bad" states in order to provide a tight upper bound given in the first two terms of the upper bound in Eq. (5). At the same time, the "large set" should be large enough to control the last two terms of the upper bound in Eq. (5). In order to obtain quantitative bounds which can be translated to tight complexity bounds, the "large set" should be chosen to balance the complexity order of the two parts of the new quantitative bound. In general, the "large set" should depend on n and p.

$\triangleleft$

In the next section, we employ this new approach to prove a certain realistic Gibbs sampler algorithm converges in $\mathcal{O}(1)$. We first choose a particular function $f(x)$ as the drift function and identify the "bad" states. In our Gibbs sampler example, there is one hidden parameter A, and the "bad" states correspond to those whose value of A is close to zero. Then we define the "large set" by ruling out the "bad" states. This allow us to obtain a quantitative bound using Theorem 2.1. Finally, under high-dimensional setting, the obtained quantitative bound can be translated to a complexity bound, which shows that the mixing time of the Gibbs sampler is $\mathcal{O}(1)$.

3. Gibbs Sampler Convergence Bound

We concentrate on a particular MCMC model, which is related to the James-Stein estimators [Ros96]:

$$\begin{aligned}
Y_i \mid \theta_i \quad &\sim \mathcal{N}(\theta_i, V), \quad 1 \le i \le n,\\
\theta_i \mid \mu, A &\sim \mathcal{N}(\mu, A), \quad 1 \le i \le n,\\
\mu &\sim \text{ flat prior on } \mathbb{R},\\
A &\sim \mathbf{IG}(a, b),
\end{aligned} \tag{7}$$

where V is assumed to be known, $(Y_1, \dots, Y_n)$ is the observed data, and $x = (A, \mu, \theta_1, \dots, \theta_n)$ are parameters. Note that we have the number of parameters $p = n + 2$ in this example. For simplicity, we will not mention p but only refer to n for this model. The posterior distribution satisfies

$$\begin{aligned}\pi(\cdot) &= \mathcal{L}(A, \mu, \theta_1, \dots, \theta_n \mid Y_1, \dots, Y_n) \\ &\propto \frac{b^a}{\Gamma(a)} A^{-a-1} e^{-b/A} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi A}} e^{-\frac{(\theta_i-\mu)^2}{2A}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i-\theta_i)^2}{2}}.\end{aligned} \tag{8}$$

A Gibbs sampler for the posterior distribution of this model has been originally analyzed in [Ros96] where a quantitative bound has been derived using the approach of [Ros95a]. However, if we translate the quantitative bound in [Ros96] into complexity orders, it requires the width of the "small set" R be $\mathcal{O}(n^2)$, which makes the minorization volume ϵ be exponentially small. This leads to upper bounds on the distance to stationarity which require exponentially large number of iterations to become small. This result also coincides with the observations by Rajaratnam and Sparks [RS15] when translating the work of Khare and Hobert [KH13] and Choi and Hobert [CH13].

In this section, we consider the following order of Gibbs sampling for computing the posterior distribution:

$$\begin{aligned}\mu^{(k+1)} &\sim \mathcal{N}\left(\bar{\theta}^{(k)}, \frac{A^{(k)}}{n}\right), \\ \theta_i^{(k+1)} &\sim \mathcal{N}\left(\frac{\mu^{(k+1)} + Y_i A^{(k)}}{V + A^{(k)}}, \frac{A^{(k)}}{V + A^{(k)}}\right), \\ A^{(k+1)} &\sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(k+1)} - \bar{\theta}^{(k+1)})^2\right).\end{aligned} \tag{9}$$

We use the new general method Theorem 2.1 to prove that the convergence of this Gibbs sampler is actually much faster: the number of iterations required is $\mathcal{O}(1)$. More precisely, we first make the following assumptions on the observed data $\{Y_i\}$: there exists $\delta > 0$ and a positive integer N_0, such that:

$$\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \Theta(1), \quad \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \geq V + \delta, \quad \forall n \geq N_0. \tag{10}$$

Remark 3.1. The assumptions in Eq. (10) are quite natural. For the second assumption, the variance of Y_i is larger than V because of the uncertainty of θ_i. Actually, conditional on A, the variance of the data

$\{Y_i\}$ is $V + A$. Therefore, the second assumption in Eq. (10) is just to assume the observed data is not abnormal when n is large enough. $\triangleleft$

Then we show that, under the assumption Eq. (10), with initial state $\bar{\theta}^{(0)} = \bar{Y}$ and $A^{(0)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} - V$, the mixing time of the Gibbs sampler to guarantee small total variation distance to stationarity is bounded by some constant when n is large enough.

Next, we show the main results. First, we obtain a quantitative bound for large enough n, which is given in the following theorem.

Theorem 3.2. *With initial state $\bar{\theta}^{(0)} = \bar{Y}$ and $A^{(0)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} - V$, there exists a positive integer N, some constants $C_1 > 0, C_2 > 0, C_3 > 0$ and $0 < \gamma < 1$, such that for all $n \geq N$ and for all k, we have*

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq C_1 \gamma^k + C_2 \frac{k(1+k)}{n} + C_3 \frac{k}{\sqrt{n}}. \tag{11}$$

Proof. See Section 5. $\square$

Finally, we translate the quantitative bound in Theorem 3.2 into the convergence complexity in terms of mixing time. The result is given in the following.

Theorem 3.3. *For any $0 < c < 1$, define the mixing time K_c by*

$$K_c(n) := \arg\min_k \|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq c. \tag{12}$$

Then there exists $N_c = \Theta(1)$ and $\bar{K}_c = \Theta(1)$ which do not depend on n, such that

$$K_c(n) \leq \bar{K}_c, \quad \forall n \geq N_c. \tag{13}$$

Proof. See Section 6. $\square$

4. Proof of Theorem 2.1

Consider the joint chains (X, Y) where Y is run on the stationary distribution π. Let ν denote the initial distribution of X. Consider the event $E_k := \{(X_i, Y_i) \in R_0 \times R_0, \forall i \in \{0, \dots, k\}\}$. Then $E_k^c = \{\exists i \in \{0, \dots, k\}, (X_i, Y_i) \notin R_0 \times R_0\}$.

Lemma 4.1. *Let N_k be the number of (X, Y) to return to $R \times R$ in k iterations*

$$\begin{aligned} \|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\| \leq &(1-\epsilon)^j + \mathbb{P}(N_k < j \mid E_k) \\ &+ k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c). \end{aligned} \tag{14}$$

Proof. The results follow from [Ros95a, Theorem 1] using

$$\begin{aligned}\mathbb{P}(N_k < j) &= \mathbb{P}(N_k < j \mid E_k)\mathbb{P}(E_k) + \mathbb{P}(N_k < j \mid E_k^c)\mathbb{P}(E_k^c) \\ &\le \mathbb{P}(N_k < j \mid E_k) + \mathbb{P}(E_k^c)\end{aligned} \tag{15}$$

and $\mathbb{P}(E_k^c) \le k\,\pi(R_0^c) + \sum_{i=1}^k P^i(\nu, R_0^c)$. □

Lemma 4.2. *Let t_i be the hitting times of (X,Y) to $R \times R$ and $r_i = t_i - t_{i-1}$ be the i-th gap of return times. For any $\alpha > 1$ and $k > j$,*

$$\mathbb{P}(N_k < j \mid E_k) \le \frac{1}{\alpha^k - \alpha^j}\left[\mathbb{E}\left(\prod_{i=1}^{j} \alpha^{r_i} \mid E_{t_j}\right) - \alpha^j\right]. \tag{16}$$

Proof. By the Markov property $\mathbb{P}(N_k < j \mid E_k) = \mathbb{P}(N_k < j \mid E_{k'})$ for all $k' \ge k$. Then the result comes from Markov's inequality on

$$\mathbb{P}(N_k < j \mid E_k) = \mathbb{P}(\alpha^{r_1+\cdots+r_j} - \alpha^j > \alpha^k - \alpha^j \mid E_{t_j}). \tag{17}$$

□

Lemma 4.3. *If there exists a function $h \ge 1$ such that*

$$\begin{aligned}&\mathbb{E}(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y, (X^{(1)}, Y^{(1)}) \in R_0 \times R_0) \\ &\quad \le \mathbb{E}(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y) \le \alpha^{-1} h(x,y), \quad \forall (x,y) \in R_0 \times R_0,\end{aligned} \tag{18}$$

then for any $0 < j < k$

$$\begin{aligned}&\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \\ &\le (1-\epsilon)^j + \frac{(\alpha\Lambda)^{j-1}\mathbb{E}_{\nu\times\pi}(h(X^{(0)}, Y^{(0)}) \mid (X^{(0)}, Y^{(0)}) \in R_0 \times R_0) - \alpha^j}{\alpha^k - \alpha^j} \\ &\quad + k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c),\end{aligned} \tag{19}$$

where

$$\Lambda := \sup_{(x,y) \in (R\times R)\cap(R_0\times R_0)} \mathbb{E}(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y). \tag{20}$$

Proof. We bound $\mathbb{E}\left(\prod_{i=1}^j \alpha^{r_i} \mid E_{t_j}\right)$ using

$$\mathbb{E}\left(\alpha^{r_i} \mid E_{t_j}, r_1, \ldots, r_{i-1}\right), i = 2, \ldots, j. \tag{21}$$

Since $t_j \ge t_{i-1}$, we have by the Markov property

$$\mathbb{E}\left(\alpha^{r_i} \mid E_{t_j}, r_1, \ldots, r_{i-1}\right) = \mathbb{E}\left(\alpha^{t_i - t_{i-1}} \mid t_{i-1}, E_{t_i}\right) \tag{22}$$

Under the assumption in Eq. (18), $g_i(k) = \alpha^k h(X^{(k)}, Y^{(k)}) \mathbb{I}_{k \le t_i}$ has non-increasing *conditional* expectation as a function of k for $t_{i-1} \le k \le t_i$. Therefore, we have

$$
\begin{aligned}
&\mathbb{E}(\alpha^{t_i - t_{i-1}} \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, E_{t_i}) \\
&\le \mathbb{E}(\alpha^{-t_{i-1}} g_i(t_i) \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, E_{t_i}) \\
&\le \mathbb{E}(\alpha^{-t_{i-1}} g_i(t_{i-1}+1) \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, E_{t_i}) \\
&= \alpha \mathbb{E}(h(X^{(t_{i-1}+1)}, Y^{(t_{i-1}+1)}) \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, E_{t_{i-1}+1}) \\
&= \alpha \mathbb{E}(h(X^{(1)}, Y^{(1)}) \,|\, (X^{(0)}, Y^{(0)}) \in (R \times R) \cap (R_0 \times R_0), (X^{(1)}, Y^{(1)}) \in R_0 \times R_0) \\
&\le \alpha \sup_{(x,y) \in (R \times R) \cap (R_0 \times R_0)} \mathbb{E}(h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y, (X^{(1)}, Y^{(1)}) \in R_0 \times R_0) \\
&\le \alpha \sup_{(x,y) \in (R \times R) \cap (R_0 \times R_0)} \mathbb{E}(h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y) = \alpha \Lambda.
\end{aligned} \tag{23}
$$

Finally the initial distributions

$$
\mathbb{E}(\alpha^{r_1} \,|\, E_0) \le \mathbb{E}(h(X^{(0)}, Y^{(0)}) \,|\, (X^{(0)}, Y^{(0)}) \in R_0 \times R_0). \tag{24}
$$

Then Lemma 4.3 follows by combing the results in Lemma 4.1 and Lemma 4.2. ☐

Now we prove Theorem 2.1 using Lemma 4.3. We set $h(x, y) = 1 + f(x) + f(y)$ and $R = \{x \in \mathcal{X} \,|\, f(x) \le d\}$. Since

$$
\mathbb{E}(f(X^{(1)})) \,|\, X^{(0)} = x, X^{(1)} \in R_0) \le \lambda f(x) + b, \forall x \in R_0, \tag{25}
$$

we have $\mathbb{E}_\pi(f(X^{(0)}) \,|\, X^{(0)} \in R_0) \le \frac{b}{1-\lambda}$. Also, if $(x, y) \notin R \times R$, we have $h(x, y) \ge 1 + d$. Thus

$$
\begin{aligned}
\mathbb{E}(h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y) \le \left(\frac{1 + 2b + \lambda d}{1 + d} \right) h(x, y), \\
\forall (x, y) \in (R \times R)^c \cap (R_0 \times R_0).
\end{aligned} \tag{26}
$$

Furthermore, we have $\Lambda = 1 + 2 \sup_{x \in R \cap R_0} \mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x) \le 1 + 2(\lambda d + b)$ and

$$
\mathbb{E}_{\nu \times \pi}(h(X^{(0)}, Y^{(0)}) \,|\, (X^{(0)}, Y^{(0)}) \in R_0 \times R_0) \tag{27}
$$

$$
= 1 + \mathbb{E}_\nu(f(X^{(0)}) \,|\, X^{(0)} \in R_0) + \frac{b}{1 - \lambda}. \tag{28}
$$

Combing the results and setting $j = rk+1$, we have

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \le (1-\epsilon)^{rk+1} + \frac{(\alpha\Lambda)^{rk}\left[1 + \mathbb{E}_\nu(f(x)) + \frac{b}{1-\lambda}\right] - \alpha^{rk+1}}{\alpha^k - \alpha^{rk+1}} + k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c). \tag{29}$$

Finally, Theorem 2.1 is proved by slightly relaxing $(1-\epsilon)^{rk+1}$ to $(1-\epsilon)^{rk}$ and α^{rk+1} to α^{rk}.

5. Proof of Theorem 3.2

We first choose the drift function, which is given in the following lemma.

Lemma 5.1. *Let* $\Delta = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ *and*

$$f(x) = n(\bar{\theta} - \bar{Y})^2 + n\left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2, \tag{30}$$

then we have

$$\mathbb{E}[f(x^{(k+1)}) \mid x^{(k)}] \le \left(\frac{V^2 + 2VA^{(k)}}{V^2 + 2VA^{(k)} + (A^{(k)})^2}\right)^2 f(x^{(k)}) + b, \tag{31}$$

where $b = \mathcal{O}(1)$.

Proof. See Section 5.1. □

Note that Eq. (31) is not a valid drift function since $\left(\frac{V^2+2VA^{(k)}}{V^2+2VA^{(k)}+(A^{(k)})^2}\right)^2$ depends on the state $A^{(k)}$. According to Eq. (10), for large enough n, we have $\frac{\Delta}{n-1} > V$. Then, we choose a threshold T such that, for large enough n, we have $0 < T < \frac{\Delta}{n-1} - V$. Defining $\lambda_T := \left(\frac{V^2+2VT}{V^2+2VT+T^2}\right)^2$, we get

$$\mathbb{E}[f(x^{(k+1)}) \mid x^{(k)}] \le \lambda_T\, f(x^{(k)}) + b, \quad \forall x \in R_T. \tag{32}$$

where the "large set", R_T, is defined by

$$R_T := \left\{x \in \mathcal{X} : \left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2 \le \left[\left(\frac{\Delta}{n-1} - V\right) - T\right]^2\right\}. \tag{33}$$

In order to satisfy the new drift condition in Eq. (3), we still need to check

$$\mathbb{E}[f(x^{(k+1)}) \mid x^{(k)}, x^{(k+1)} \in R_T] \le \mathbb{E}[f(x^{(k+1)}) \mid x^{(k)}], \quad \forall x \in R_T. \tag{34}$$

According to the order of Gibbs sampling, if we denote $\eta_i^{(k+1)} := \theta_i^{(k+1)} - \frac{Y_i A^{(k)}}{V+A^{(k)}}$, then $\{\eta_i^{(k+1)}\}$ are i.i.d. samples from a Normal distribution. Therefore, we have $\bar{\eta}^{(k+1)}$ is independent with $\eta_i^{(k+1)} - \bar{\eta}^{(k+1)}$, which implies $\bar{\theta}^{(k+1)}$ is conditional independent with $\theta_i^{(k+1)} - \bar{\theta}^{(k+1)}$ given $A^{(k)}$. Since $A^{(k+1)}$ is sampled from the last step of Gibbs sampling only using $\sum_{i=1}^n (\theta_i^{(k+1)} - \bar{\theta}^{(k+1)})^2$, we have $n(\bar{\theta}^{(k+1)} - \bar{Y})^2$ is conditional independent with $A^{(k+1)}$ given $x^{(k)}$, therefore

$$\mathbb{E}[(\bar{\theta}^{(k+1)} - \bar{Y})^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T] = \mathbb{E}[(\bar{\theta}^{(k+1)} - \bar{Y})^2 \,|\, x^{(k)}]. \tag{35}$$

Furthermore, by the definition of R_T, we have

$$\begin{aligned} &\mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V\right) - A^{(k+1)}\right)^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T\right] \\ &\leq \left(\left(\frac{\Delta}{n-1} - V\right) - T\right)^2 \\ &\leq \mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V\right) - A^{(k+1)}\right)^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T^c\right], \end{aligned} \tag{36}$$

which implies

$$\begin{aligned} &\mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V\right) - A^{(k+1)}\right)^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T\right] \\ &\leq \mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V\right) - A^{(k+1)}\right)^2 \,|\, x^{(k)}\right]. \end{aligned} \tag{37}$$

Therefore, the new drift condition of Eq. (3) is satisfied.

Now we can use Theorem 2.1 to derive a quantitative bound for the Gibbs sampler. We first present some useful lemmas.

Lemma 5.2. *If $T = \Theta(1)$, then $d = \mathcal{O}(1)$ and the minorization volume $\epsilon = \Theta(1)$.*

Proof. See Section 5.2. ∎

Lemma 5.3. *With the initial state $\bar{\theta}^{(0)} = \bar{Y}$ and $A^{(0)} = \frac{\Delta}{n-1} - V$, there exists a positive integer N such that for all $n \geq N$, we have*

$$\begin{aligned} &k\,\pi(R_T^c) + \sum_{i=1}^k P^i(x_0, R_T^c) \\ &\leq \frac{k(1+k)}{2n} \frac{b}{\left[\left(\frac{\Delta}{n-1} - V\right) - T\right]^2} + \frac{k}{\sqrt{n}} \frac{\sqrt{b}(2V/\delta + 1)}{\left|\left(\frac{\Delta}{n-1} - V\right) - T\right|}. \end{aligned} \tag{38}$$

Proof. See Section 5.3. ☐

Now we derive a quantitative bound for the Gibbs sampler for large enough n by combing results together. First, from Lemma 5.1 and Lemma 5.2, we have $b = \mathcal{O}(1)$. By choosing $d = \Theta(1)$, we have $\epsilon = \Theta(1)$. Therefore, $\alpha^{-1} = \frac{1+2b+\lambda d}{1+d} = \Theta(1) < 1$ and $\Lambda = 1 + 2(\lambda d + b) = \Theta(1)$. Next, we choose $r = \log(\alpha)/\log(\alpha\Lambda/(1-\epsilon))$ to balance $(1-\epsilon)^r = \alpha^{-1}(\alpha\Lambda)^r$ and denote it as γ. Then we have $\gamma = \Theta(1)$ and $0 < \gamma < 1$. Since $\frac{b}{1-\lambda} = \Theta(1)$ and $f(x_0) = 0$ with initial state $\bar{\theta}^{(0)} = \bar{Y}$ and $A^{(0)} = \frac{\Delta}{n-1} - V$, we can define $C_1 = 2 + \frac{b}{1-\lambda}$. Furthermore, we have $k\pi(R_T^c) + \sum_{i=1}^k P^i(x_0, R_T^c) \le C_2\frac{k(1+k)}{n} + C_3\frac{k}{\sqrt{n}}$ by Lemma 5.3. Finally, Theorem 3.2 follows from applying Theorem 2.1.

5.1. **Proof of Lemma 5.1.** For simplicity and without loss of generality, we set $V = 1$ and use $\mathcal{O}(1)$ to denote terms that can be upper bounded by some constant that does not depend on the state $x = (A, \mu, \theta_1, \dots, \theta_n)$.

5.1.1. *Expectation over $A^{(1)}$ given $\theta^{(1)}$ and $\mu^{(1)}$.* The first term of $f(x)$, $n(\bar{\theta} - \bar{Y})^2$, is unchanged. Denoting $S := \frac{\sum_i(\theta_i - \bar{\theta})^2}{n-1}$, the second term of $f(x)$ becomes

$$n\mathbb{E}_{A^{(1)}}\left[\left(\frac{\Delta}{n-1} - 1\right) - A\right]^2 = n\left(\frac{\Delta}{n-1} - 1\right)^2 + nS^2 - 2n\left(\frac{\Delta}{n-1} - 1\right)S + \mathcal{O}(1) + \mathcal{O}(1)S + \mathcal{O}(1)S^2, \tag{39}$$

where the order $\mathcal{O}(1)$ terms come from the approximation

$$n\frac{\sum_i(\theta_i - \bar{\theta})^2 + 2b}{n - 1 + 2(a-1)} = nS + \mathcal{O}(1) + \mathcal{O}(1)S \tag{40}$$

and n times the variance of A

$$\frac{n(\sum_i(\theta_i - \bar{\theta})^2/2 + b)^2}{[(n-1)/2 + (a-1)]^2[(n-1)/2 + (a-2)]} = \mathcal{O}(1)S^2 + \mathcal{O}(1/n)S + \mathcal{O}(1/n^2). \tag{41}$$

5.1.2. *Expectation over $\theta^{(1)}$ given $\mu^{(1)}$.* For the first term of $f(x)$, we have

$$n\mathbb{E}_{\theta^{(1)}}(\bar{\theta} - \bar{Y})^2 = \frac{A}{1+A} + \frac{n(\mu^{(1)} - \bar{Y})^2}{(1+A)^2} \le \frac{n(\mu^{(1)} - \bar{Y})^2}{(1+A)^2} + 1. \tag{42}$$

For the second term of $f(x)$, we have

$$n\left(\frac{\Delta}{n-1}-1\right)^2+n\mathbb{E}_{\theta^{(1)}}[S^2]-2n\left(\frac{\Delta}{n-1}-1\right)\mathbb{E}_{\theta^{(1)}}[S] \\ +\mathcal{O}(1)+\mathcal{O}(1)\mathbb{E}_{\theta^{(1)}}[S^2]+\mathcal{O}(1)\mathbb{E}_{\theta^{(1)}}[S]. \tag{43}$$

Next, we apply Lemma 5.4 and combine the two terms of $f(x)$ to get

$$\frac{n(\mu^{(1)}-\bar{Y})^2}{(1+A)^2}+n\left[\frac{A}{1+A}+\frac{\Delta}{n-1}\left(\frac{A}{1+A}\right)^2-\left(\frac{\Delta}{n-1}-1\right)\right]^2+\mathcal{O}(1), \tag{44}$$

where $\mathbb{E}_{\theta^{(1)}}[S]\leq 1+\frac{\Delta}{n-1}$ has been used to simplify the $\mathcal{O}(1)$ terms.

Lemma 5.4.

$$\mathbb{E}_{\theta^{(1)}}[S]=\frac{A}{1+A}+\frac{\Delta}{n-1}\left(\frac{A}{1+A}\right)^2, \quad \mathbb{E}_{\theta^{(1)}}[S^2]=(\mathbb{E}_{\theta^{(1)}}[S])^2+\mathcal{O}(1/n). \tag{45}$$

Proof. We have set $V=1$ in this proof for simplicity. Define $\eta_i:=\theta_i-Y_i\frac{A}{1+A}$ and $\bar{\eta}:=\bar{\theta}-\bar{Y}\frac{A}{1+A}$. Then $\eta_i\sim\mathcal{N}(0,A/(1+A))$ and $\bar{\eta}\sim\mathcal{N}(0,\frac{1}{n}A/(1+A))$. Decomposing $\sum_i(\theta_i-\bar{\theta})^2$ by

$$\sum_i(\theta_i-\bar{\theta})^2=\sum\left((\eta_i-\bar{\eta})^2+\left(\frac{A}{1+A}\right)^2(Y_i-\bar{Y})^2+\frac{2(\eta_i-\bar{\eta})(Y_i-\bar{Y})A}{1+A}\right), \tag{46}$$

and taking expectation over θ yields $(n-1)\frac{A}{1+A}+\left(\frac{A}{1+A}\right)^2\Delta$. Therefore, $\mathbb{E}(S)=\frac{A}{1+A}+\frac{\Delta}{n-1}\left(\frac{A}{1+A}\right)^2$.

Furthermore, for $\mathbb{E}(S^2)$, since

$$\mathbb{E}\left[\frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1}\right]^2=\left[\mathbb{E}\frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1}\right]^2+\mathcal{O}(1/n), \tag{47}$$

we have

$$\mathbb{E}(S^2)=\left(\mathbb{E}\left[\frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1}\right]+\frac{\Delta}{n-1}\left(\frac{A}{1+A}\right)^2\right)^2 \\ +2\left(\frac{A}{1+A}\right)^2\frac{\mathbb{E}\left[\sum_i(\eta_i-\bar{\eta})(Y_i-\bar{Y})\right]^2}{(n-1)^2} \tag{48}$$

where the last term

$$\begin{aligned}
&\frac{\mathbb{E}\left[\sum_i(\eta_i-\bar{\eta})(Y_i-\bar{Y})\right]^2}{(n-1)^2}\\
&=\frac{\mathbb{E}\left[\sum_i(\eta_i-\bar{\eta})^2(Y_i-\bar{Y})^2\right]+\mathbb{E}[\bar{\eta}^2]\sum_{i\neq j}(Y_i-\bar{Y})(Y_j-\bar{Y})}{(n-1)^2}\\
&=\frac{\sum_i(Y_i-\bar{Y})^2}{(n-1)^2}\mathbb{E}(\eta_1-\bar{\eta})^2+\mathcal{O}(1/n)\\
&=\frac{\Delta}{(n-1)^2}\frac{(n-1)\frac{A}{1+A}}{n}+\mathcal{O}(1/n)=\mathcal{O}(1/n).
\end{aligned} \tag{49}$$

Therefore, we have

$$\mathbb{E}(S^2)=\left[\frac{A}{1+A}+\frac{\Delta}{n-1}\left(\frac{A}{1+A}\right)^2\right]+\mathcal{O}(1/n). \tag{50}$$

□

5.1.3. *Expectation over $\mu^{(1)}$.* Only the term $n(\mu^{(1)}-\bar{Y})^2\frac{1}{(1+A)^2}$ involves $\mu^{(1)}$. Since

$$\mathbb{E}_{\mu^{(1)}}\left[n(\mu^{(1)}-\bar{Y})^2\right]=n(\bar{\theta}-\bar{Y})^2+A, \tag{51}$$

we have

$$\frac{n(\bar{\theta}-\bar{Y})^2}{(1+A)^2}+n\left[\frac{A}{1+A}+\frac{\Delta}{n-1}\left(\frac{A}{1+A}\right)^2-\left(\frac{\Delta}{n-1}-1\right)\right]^2+\mathcal{O}(1). \tag{52}$$

Finally, the second term can be simplified as

$$\begin{aligned}
&n\left[\frac{A}{1+A}+\frac{\Delta}{n-1}\left(\frac{A}{1+A}\right)^2-\left(\frac{\Delta}{n-1}-1\right)\right]^2\\
&=n\left[\frac{\Delta}{n-1}\left[\left(\frac{A}{1+A}\right)^2-1\right]+\left(\frac{A}{1+A}+1\right)\right]^2\\
&=n\left(\frac{A}{1+A}+1\right)^2\left[\frac{\Delta}{n-1}\left(\frac{-1}{1+A}\right)+1\right]^2\\
&=\frac{(1+2A)^2}{(1+A)^4}n\left[\frac{\Delta}{n-1}-(A+1)\right]^2.
\end{aligned} \tag{53}$$

Therefore, we have

$$\begin{aligned}
&\frac{1}{(1+A)^2}n(\bar\theta-\bar Y)^2+\frac{(1+2A)^2}{(1+A)^4}n\left[\frac{\Delta}{n-1}-(A+1)\right]^2\\
&=\frac{(1+2A)^2}{(1+A)^4}\left\{\frac{(1+A)^2}{(1+2A)^2}n(\bar\theta-\bar Y)^2+n\left[\frac{\Delta}{n-1}-(A+1)\right]^2\right\}\\
&\le\frac{(1+2A)^2}{(1+A)^4}\left\{n(\bar\theta-\bar Y)^2+n\left[\frac{\Delta}{n-1}-(A+1)\right]^2\right\}.
\end{aligned}\tag{54}$$

5.2. **Proof of Lemma 5.2.** For simplicity, we set $V=1$ in this proof. In order to construct the minorization volume ϵ, instead of considering $(\theta_1,\dots,\theta_n)$, we integrate over $\bar\theta$ and $S=\frac{\sum_i(\theta_i-\bar\theta)^2}{n-1}$. Since we have shown $\bar\theta$ is conditional independent with S given A in the proof of Lemma 5.4, we can integrate them separately. Denoting $\hat A:=\frac{\Delta}{k-1}-1$ and $f_S(A,n;S)$ as the density of S given A, then

$$\begin{aligned}
&\int \mathrm{d}S\mathrm{d}\bar\theta \inf_{|A-\hat A|\le\sqrt{d/n}} f_S(A,n;S)\,\mathcal{N}\left(\frac{\mu}{1+A}+\bar Y\frac{A}{1+A},\frac{A}{n(1+A)};\bar\theta\right)\\
&\ge\int \mathrm{d}S \inf_{|A-\hat A|\le\sqrt{d/n}} f_S(A,n;S)\\
&\quad\cdot\int \mathrm{d}\bar\theta \inf_{|A-\hat A|\le\sqrt{d/n}} \mathcal{N}\left(\frac{\mu}{1+A}+\bar Y\frac{A}{1+A},\frac{A}{n(1+A)};\bar\theta\right)
\end{aligned}\tag{55}$$

5.2.1. *Integration over S.* In the proof of Lemma 5.4, Eqs. (46) and (49) implies that, with $\eta_i=\theta_i-Y_i\frac{A}{1+A}\sim\mathcal{N}(0,\frac{A}{1+A})$, S converges to $\frac{\sum_i(\eta_i-\bar\eta)^2}{n-1}+(\frac{A}{1+A})^2\Delta$ in L^2-norm, conditional on A. Since $\frac{A}{1+A}$ is bounded when $|A-\hat A|\le\sqrt{d/n}$, it suffices to show the minorization volume of $\left\{\frac{\frac{1+A}{A}\sum_i(\eta_i-\bar\eta)^2}{n-1}+(\frac{A}{1+A})\Delta\right\}$ for $|A-\hat A|\le\sqrt{d/n}$ is $\Theta(1)$. Note that there exists some constant C_0 such that

$$\left|\max_{\{A:|A-\hat A|\le\sqrt{d/n}\}}\frac{A}{1+A}\Delta-\min_{\{A:|A-\hat A|\le\sqrt{d/n}\}}\frac{A}{1+A}\Delta\right|\le\frac{C_0}{\sqrt{n-1}}.\tag{56}$$

Defining $f'(z,A;x),\forall z\in\mathbb{R}$ as the density function of a random variable

$$\tilde X_{z,A}:=z+\frac{\frac{1+A}{A}\sum_i(\eta_i-\bar\eta)^2-(n-1)}{\sqrt{2(n-1)}},\tag{57}$$

the minorization volume of $\left\{\frac{\frac{1+A}{A}\sum_i(\eta_i-\bar{\eta})^2}{n-1}+(\frac{A}{1+A})\Delta\right\}$ for $|A-\hat{A}|\leq\sqrt{d/n}$ is asymptotically no more than

$$\begin{aligned}&\min_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\int \mathrm{d}x\min\left\{f'\left(-\frac{C_0}{\sqrt{2}},A;x\right),f'\left(+\frac{C_0}{\sqrt{2}},A;x\right)\right\}\\ &=1-\max_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\int_{-\sqrt{2}C_0}^{\sqrt{2}C_0}\mathrm{d}x f'(0,A;x).\end{aligned}\tag{58}$$

Therefore, it suffices to show

$$1-\max_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\mathbb{P}(-\sqrt{2}C_0\leq\tilde{X}_{0,A}\leq\sqrt{2}C_0)=\Theta(1).\tag{59}$$

Finally, since $\frac{1+A}{A}\sum_i(\eta_i-\bar{\eta})^2\sim\chi^2_{n-1}$ we have $\frac{\frac{1+A}{A}\sum_i(\eta_i-\bar{\eta})^2-(n-1)}{\sqrt{2(n-1)}}\xrightarrow{d}\mathcal{N}(0,1)$. That is $\mathbb{P}(-\sqrt{2}C_0\leq X_{0,A}\leq\sqrt{2}C_0)\to\int_{-\sqrt{2}C_0}^{\sqrt{2}C_0}\mathrm{d}x\mathcal{N}(0,1;x)=\Theta(1)$. Therefore, we have $\max_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\mathbb{P}(-\sqrt{2}C_0\leq\tilde{X}_{0,A}\leq\sqrt{2}C_0)=\Theta(1)$.

5.2.2. *Integration over $\bar{\theta}$.* Next, we consider

$$\int \mathrm{d}\bar{\theta}\inf_{|A-\hat{A}|\leq\sqrt{d/n}}\mathcal{N}\left(\frac{\mu}{1+A}+\bar{Y}\frac{A}{1+A},\frac{A}{n(1+A)};\bar{\theta}\right).$$

Note that there exists some constants C_1 and C_2 such that

$$\max_{|A-\hat{A}|\leq\sqrt{d/n}}\frac{\mu+\bar{Y}A}{1+A}-\min_{|A-\hat{A}|\leq\sqrt{d/n}}\frac{\mu+\bar{Y}A}{1+A}\leq\frac{C_1|\mu|+C_2}{\sqrt{n}}.\tag{60}$$

Then we have

$$\begin{aligned}&\int \mathrm{d}\bar{\theta}\inf_{|A-\hat{A}|\leq\sqrt{d/n}}\mathcal{N}\left(\frac{\mu}{1+A}+\bar{Y}\frac{A}{1+A},\frac{A}{n(1+A)};\bar{\theta}\right)\\ &\geq 2\int_{(C_1|\mu|+C_2)/\sqrt{n}}^{\infty}\mathrm{d}x\,\mathcal{N}(0,C_3/n;x)=2\int_{C_4|\mu|+C_5}^{\infty}\mathrm{d}x\,\mathcal{N}(0,1;x)\end{aligned}\tag{61}$$

for some constant C_3 and $C_4:=C_1/\sqrt{C_3},C_5:=C_2/\sqrt{C_3}$. Note that the last term $g(\mu):=2\int_{C_4|\mu|+C_5}^{\infty}\mathrm{d}x\,\mathcal{N}(0,1;x)=1-\mathrm{erf}(\frac{C_4|\mu|+C_5}{\sqrt{2}})$ does not depends on n.

5.2.3. *Integration over μ.* Finally, we consider

$$\int \mathrm{d}\mu \left\{ g(\mu) \inf_{\left\{|\bar{\theta}-\bar{Y}|\le\sqrt{d/n}, |A-\hat{A}|\le\sqrt{d/n}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\}. \tag{62}$$

We only need to show the above quantity is (asymptotically) bounded away from zero. Note that

$$\begin{aligned} &\inf_{\left\{|\bar{\theta}-\bar{Y}|\le\sqrt{d/n}, |A-\hat{A}|\le\sqrt{d/n}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \\ &\quad = \min\left\{\mathcal{N}(\bar{Y}-\sqrt{d/n}, \frac{C_6}{n}; \mu), \mathcal{N}(\bar{Y}+\sqrt{d/n}, \frac{C_6}{n}; \mu)\right\} \end{aligned} \tag{63}$$

where $C_6 \in [\hat{A}-\sqrt{d/n}, \hat{A}+\sqrt{d/n}]$. Then we have

$$\begin{aligned} &\int \mathrm{d}\mu \left\{ g(\mu) \inf_{\left\{|\bar{\theta}-\bar{Y}|\le\sqrt{d/n}, |A-\hat{A}|\le\sqrt{d/n}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\} \\ &\ge \int_0^{2\bar{Y}} \mathrm{d}\mu \left\{ g(\mu) \inf_{\left\{|\bar{\theta}-\bar{Y}|\le\sqrt{d/n}, |A-\hat{A}|\le\sqrt{d/n}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\} \\ &\ge \left(1-\mathrm{erf}(\frac{C_4|2\bar{Y}|+C_5}{\sqrt{2}})\right) \int_0^{2\bar{Y}} \mathrm{d}\mu \inf_{\left\{|\bar{\theta}-\bar{Y}|\le\sqrt{d/n}, |A-\hat{A}|\le\sqrt{d/n}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \\ &= \left(1-\mathrm{erf}(\frac{C_4|2\bar{Y}|+C_5}{\sqrt{2}})\right) \\ &\quad \cdot \left[\int_0^{\bar{Y}} \mathrm{d}\mu\, \mathcal{N}(\bar{Y}+\sqrt{d/n}, \frac{C_6}{n}; \mu) + \int_{\bar{Y}}^{2\bar{Y}} \mathrm{d}\mu\, \mathcal{N}(\bar{Y}-\sqrt{d/n}, \frac{C_6}{n}; \mu)\right] \\ &= \left(1-\mathrm{erf}(\frac{C_4|2\bar{Y}|+C_5}{\sqrt{2}})\right) \\ &\quad \cdot \left[\int_{-\bar{Y}}^{0} \mathrm{d}\mu\, \mathcal{N}(\sqrt{d/n}, \frac{C_6}{n}; \mu) + \int_0^{\bar{Y}} \mathrm{d}\mu\, \mathcal{N}(-\sqrt{d/n}, \frac{C_6}{n}; \mu)\right] \\ &\to \left(1-\mathrm{erf}(\frac{C_4|2\bar{Y}|+C_5}{\sqrt{2}})\right) \left[1-\int_{-\sqrt{d/n}}^{\sqrt{d/n}} \mathrm{d}x\, \mathcal{N}(0, \frac{\hat{A}}{n}; x)\right] \\ &= \left(1-\mathrm{erf}(\frac{C_4|2\bar{Y}|+C_5}{\sqrt{2}})\right) \left[1-\int_{-\sqrt{d}}^{\sqrt{d}} \mathrm{d}x\, \mathcal{N}(0, \hat{A}; x)\right] = \Theta(1). \end{aligned} \tag{64}$$

5.3. **Proof of Lemma 5.3.** We first consider a Markov chain starting from initial state x_0. Since $f(x_0) = 0$, we have $\mathbb{E}(f(x^{(1)})) \le b$ from Lemma 5.1. Furthermore, we can continue to get upper bounds $\mathbb{E}(f(x^{(i)})) \le ib$ for all $i = 1, \ldots, k$. This implies

$$\mathbb{E}\left[\left(\frac{\Delta}{n-1} - V\right) - A^{(i)}\right]^2 \le i\frac{b}{n}, \quad i = 1, \ldots, k. \tag{65}$$

By Markov's inequality, we have

$$\mathbb{P}\left(\left|A^{(i)} - \left(\frac{\Delta}{n-1} - V\right)\right| \ge \left|T - \left(\frac{\Delta}{n-1} - V\right)\right|\right) \le \frac{i}{n}\frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}, \tag{66}$$

for $i = 1, \ldots, k$. Therefore, we have

$$\sum_{i=1}^{k} P^i(x_0, R_T^c) \le \frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}\sum_{i=1}^{k}\frac{i}{n} = \frac{k(1+k)}{2n}\frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}. \tag{67}$$

Next, we consider a Markov chain starting from π. According to Lemma 5.1, we have

$$\begin{aligned} &\mathbb{E}_\pi\left[\left(1 - \left(\frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right)^2\right) f(x)\right] \\ &= \mathbb{E}_\pi\left[\left(1 + \frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right)\left(1 - \frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right) f(x)\right] \\ &= \mathbb{E}_\pi\left[\left(1 + \frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right)\left(\frac{A}{V+A}\right)^2 f(x)\right] \le b. \end{aligned} \tag{68}$$

Note that by reverse Hölder's inequality

$$\begin{aligned} &\mathbb{E}_\pi\left[\left(1 + \frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right)\left(\frac{A}{V+A}\right)^2 f(x)\right] \\ &\ge \mathbb{E}_\pi\left[\left(\frac{A}{V+A}\right)^2 f(x)\right] \\ &\ge [\mathbb{E}_\pi(f(x)^{\frac{1}{2}})]^2\left[\mathbb{E}_\pi\left(\frac{A}{V+A}\right)^{-1}\right]^{-2} \\ &= [\mathbb{E}_\pi(f(x)^{\frac{1}{2}})]^2[\mathbb{E}_\pi(1 + V/A)]^{-2}. \end{aligned} \tag{69}$$

Therefore, we have

$$\mathbb{E}_\pi(f(x)^{\frac{1}{2}}) \le \sqrt{b}[1 + V\mathbb{E}_\pi(1/A)] \tag{70}$$

By Lemma 5.5, we have $1 + V\mathbb{E}_\pi(1/A) \le 1 + 2V/\delta$ for large enough n. Therefore, we get

$$\mathbb{E}_\pi\left(\left|\left(\frac{\Delta}{n-1} - V\right) - A\right|\right) \le \sqrt{\frac{b}{n}}(2V/\delta + 1).$$

Thus, by Markov's inequality

$$\begin{aligned}\pi(R_T^c) &= \mathbb{P}_\pi\left(\left|\left(\frac{\Delta}{n-1} - V\right) - A\right| \ge \left|\left(\frac{\Delta}{n-1} - V\right) - T\right|\right)\\ &\le \frac{\sqrt{\frac{b}{n}}(2V/\delta+1)}{\left|\left(\frac{\Delta}{n-1} - V\right) - T\right|}.\end{aligned} \tag{71}$$

Finally, we have

$$\begin{aligned}&k\,\pi(R_T^c) + \sum_{i=1}^{k} P^i(x_0, R_T^c)\\ &\le \frac{k(1+k)}{2n}\frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2} + \frac{k}{\sqrt{n}}\frac{\sqrt{b}(2V/\delta+1)}{\left|\left(\frac{\Delta}{n-1} - V\right) - T\right|}.\end{aligned} \tag{72}$$

Lemma 5.5. *There exists a positive integer N such that for all $n \ge N$, we have*

$$\mathbb{E}_\pi(1/A) \le 2/\delta. \tag{73}$$

Proof. For simplicity, we set $V = 1$ in this proof. Note that

$$\pi(x \mid Y_1, \dots, Y_n) = \frac{f_a(x, Y_1, \dots, Y_n)}{\int f_a(x, Y_1, \dots, Y_n)\mathrm{d}x}, \tag{74}$$

where we use $f_a(x, Y_1, \dots, Y_n)$ to denote the joint distribution of x and $\{Y_i\}$ when $\mathbf{IG}(a, b)$ is used as the prior for A. That is,

$$\begin{aligned}&f_a(x, Y_1, \dots, Y_n)\\ &= \frac{b^a}{\Gamma(a)}A^{-a-1}e^{-b/A}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi A}}e^{-\frac{(\theta_i-\mu)^2}{2A}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(Y_i-\theta_i)^2}{2}}\\ &= \frac{1}{(2\pi)^n}\frac{b^a}{\Gamma(a)}A^{-a-1-\frac{n}{2}}e^{-b/A}\exp\left[-\sum_{i=1}^{n}\left(\frac{(\theta_i-\mu)^2}{2A} + \frac{(Y_i-\theta_i)^2}{2}\right)\right].\end{aligned} \tag{75}$$

Now using $\frac{1}{A} f_a(x, Y_1, \ldots, Y_n) = \frac{a}{b} f_{a+1}(x, Y_1, \ldots, Y_n)$, we have

$$\mathbb{E}_\pi(1/A) = \frac{a}{b} \frac{\int f_{a+1}(x, Y_1, \ldots, Y_n) \mathrm{d}x}{\int f_a(x, Y_1, \ldots, Y_n) \mathrm{d}x}. \tag{76}$$

Therefore, it suffices to show the ratio of $\int f_{a+1}(x, Y_1, \ldots, Y_n)\mathrm{d}x$ and $\int f_a(x, Y_1, \ldots, Y_n)\mathrm{d}x$ is (asymptotically) bounded. Now using the facts that

$$\int \exp\left[-\left(\frac{(\theta_i - \mu)^2 + A(Y_i - \theta_i)^2}{2A}\right)\right] \mathrm{d}\theta_i \propto \sqrt{2\pi \frac{A}{1+A}} \exp\left[-\frac{(Y_i - \mu)^2}{2(1+A)}\right] \tag{77}$$

and

$$\int \exp\left[-\frac{\sum_{i=1}^n (Y_i - \mu)^2}{2(1+A)}\right] \mathrm{d}\mu \propto \exp\left[-\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{2(1+A)}\right] \sqrt{2\pi \frac{1+A}{n}}, \tag{78}$$

we have $\mathbb{E}_\pi(1/A) = g_n(\Delta)$, where

$$g_n(\Delta) := \frac{\int A^{-a-2} e^{-b/A} (1+A)^{\frac{-n+1}{2}} \exp\left[-\frac{\Delta}{2(1+A)}\right] \mathrm{d}A}{\int A^{-a-1} e^{-b/A} (1+A)^{\frac{-n+1}{2}} \exp\left[-\frac{\Delta}{2(1+A)}\right] \mathrm{d}A}. \tag{79}$$

Next, we show $g_n((n-1)(c+1))$ is (asymptotically) bounded for any fixed $c > 0$. Note that

$$\int A^{-a-1} e^{-b/A} (1+A)^{\frac{-n+1}{2}} \exp\left[-\frac{\Delta}{2(1+A)}\right] \mathrm{d}A \tag{80}$$

$$= \int A^{-a-1} e^{-b/A} \left\{ \frac{1}{\sqrt{1+A}} \exp\left[-\frac{\frac{\Delta}{n-1}}{2(1+A)}\right] \right\}^{n-1} \mathrm{d}A. \tag{81}$$

We change variable $y = \frac{1}{\sqrt{1+A}}$ and apply the Laplace approximation. Note that for any $c > 0$, let $y_0 = \arg\max_y \left[y \exp\left(-\frac{c+1}{2} y^2\right)\right]$, then $y_0 = \frac{1}{\sqrt{c+1}}$. Therefore, by the Laplace approximation [ZC04, Thm. 1, Chp. 19.2.4], we have

$$\begin{aligned} g_n((n-1)(c+1)) &= \frac{c^{-a-2} e^{-b/c} \left[y_0 \exp\left(-\frac{c+1}{2} y_0^2\right)\right]^{n-1} (1 + \mathcal{O}(n^{-\frac{1}{2}}))}{c^{-a-1} e^{-b/c} \left[y_0 \exp\left(-\frac{c+1}{2} y_0^2\right)\right]^{n-1} (1 + \mathcal{O}(n^{-\frac{1}{2}}))} \\ &= \frac{1}{c}(1 + \mathcal{O}(n^{-1/2})). \end{aligned} \tag{82}$$

Finally, since for all $n \geq N_0$ we have $\Delta \geq (n-1)(1+\delta)$, this implies $g_n(\Delta) \leq \frac{1}{\delta}(1 + \mathcal{O}(n^{-1/2})), \forall n \geq N_0$. Therefore, there exists large enough positive integer N such that for all $n \geq N$, we have $\mathbb{E}_\pi(1/A) = g_n(\Delta) \leq \frac{1}{\delta}(1 + \mathcal{O}(n^{-1/2})) \leq \frac{2}{\delta}$. $\square$

6. Proof of Theorem 3.3

Using Theorem 3.2, one sufficient condition for

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq c \tag{83}$$

is that $n \geq N$ and

$$C_1\gamma^k \leq \frac{c}{3}, \quad C_2\frac{(1+k)^2}{n} \leq \frac{c}{3}, \quad C_3\frac{k}{\sqrt{n}} \leq \frac{c}{3}. \tag{84}$$

This requires the number of iterations, k, satisfies

$$\frac{\log(C_1) - \log(c/3)}{\log(1/\gamma)} \leq k \leq \max\left\{\sqrt{\frac{c/3}{C_3}\sqrt{n}} - 1, \frac{c/3}{C_3}\sqrt{n}\right\}. \tag{85}$$

Note that any k (if exists) satisfying the above equation provides an upper bound for the mixing time $K_c(n)$.

That is, for any $n \geq N$ such that

$$\frac{\log(C_1) - \log(c/3)}{\log(1/\gamma)} \leq \max\left\{\sqrt{\frac{c/3}{C_3}\sqrt{n}} - 1, \frac{c/3}{C_3}\sqrt{n}\right\}, \tag{86}$$

which is equivalent to

$$n \geq \max\left\{N, \left[\bar{K}_c\frac{3C_3}{c}\right]^2, \left[\left(\bar{K}_c + 1\right)\sqrt{\frac{3C_3}{c}}\right]^2\right\} =: N_c, \tag{87}$$

we have $\bar{K}_c := \frac{\log(C_1) - \log(c) + \log(3)}{\log(1/\gamma)}$ is an upper bound of the mixing time.

Finally, it can be seen that both $\bar{K}_c = \Theta(1)$ and $N_c = \Theta(1)$.

Acknowledgments

The authors thank Jim Hobert and Gareth Roberts for very helpful discussions. This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[Bax05] P. H. Baxendale. “Renewal theory and computable convergence rates for geometrically ergodic Markov chains”. *The Annals of Applied Probability* 15.1B (2005), pp. 700–738.

[Bro+11] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[CH13] H. M. Choi and J. P. Hobert. "The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic". *Electronic Journal of Statistics* 7 (2013), pp. 2054–2064.

[Cob65] A. Cobham. "The Intrinsic Computational Difficulty of Functions". In: *Logic, Methodology and Philosophy of Science: Proceedings of the 1964 International Congress (Studies in Logic and the Foundations of Mathematics)*. Ed. by Y. Bar-Hillel. North-Holland Publishing, 1965, pp. 24–30.

[Coo71] S. A. Cook. "The complexity of theorem-proving procedures". In: *Proceedings of the third annual ACM symposium on Theory of computing*. ACM. 1971, pp. 151–158.

[FHJ08] J. M. Flegal, M. Haran, and G. L. Jones. "Markov chain Monte Carlo: Can we trust the third significant figure?" *Statistical Science* (2008), pp. 250–260.

[GR92] A. Gelman and D. B. Rubin. "Inference from iterative simulation using multiple sequences". *Statistical Science* (1992), pp. 457–472.

[GRS95] W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

[JH01] G. L. Jones and J. P. Hobert. "Honest exploration of intractable probability distributions via Markov chain Monte Carlo". *Statistical Science* (2001), pp. 312–334.

[JH04] G. L. Jones and J. P. Hobert. "Sufficient burn-in for Gibbs samplers for a hierarchical random effects model". *The Annals of Statistics* 32.2 (2004), pp. 784–817.

[KH13] K. Khare and J. P. Hobert. "Geometric ergodicity of the Bayesian lasso". *Electronic Journal of Statistics* 7 (2013), pp. 2150–2163.

[MT12] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[MT94] S. P. Meyn and R. L. Tweedie. "Computable bounds for geometric convergence rates of Markov chains". *The Annals of Applied Probability* (1994), pp. 981–1011.

[RGG97] G. O. Roberts, A. Gelman, and W. R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". *The Annals of Applied Probability* 7.1 (1997), pp. 110–120.

[Ros02] J. S. Rosenthal. "Quantitative convergence rates of Markov chains: A simple account". *Electronic Communications in Probability* 7 (2002), pp. 123–128.

[Ros95a] J. S. Rosenthal. "Minorization conditions and convergence rates for Markov chain Monte Carlo". *Journal of the American Statistical Association* 90.430 (1995), pp. 558–566.

[Ros95b] J. S. Rosenthal. "Rates of convergence for Gibbs sampling for variance component models". *The Annals of Statistics* (1995), pp. 740–761.

[Ros96] J. S. Rosenthal. "Analysis of the Gibbs sampler for a model related to James-Stein estimators". *Statistics and Computing* 6.3 (1996), pp. 269–275.

[RR16] G. O. Roberts and J. S. Rosenthal. "Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits". *Journal of Applied Probability* 53.2 (2016), pp. 410–420.

[RR98] G. O. Roberts and J. S. Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1 (1998), pp. 255–268.

[RS15] B. Rajaratnam and D. Sparks. "MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains". *arXiv preprint arXiv:1508.00947* (2015).

[RT99] G. O. Roberts and R. L. Tweedie. "Bounds on regeneration times and convergence rates for Markov chains". *Stochastic Processes and their applications* 80.2 (1999), pp. 211–229.

[SJ89] A. Sinclair and M. Jerrum. "Approximate counting, uniform generation and rapidly mixing Markov chains". *Information and Computation* 82.1 (1989), pp. 93–133.

[WSH09a] D. Woodard, S. Schmidler, and M. Huber. "Sufficient conditions for torpid mixing of parallel and simulated tempering". *Electronic Journal of Probability* 14 (2009), pp. 780–804.

[WSH09b] D. B. Woodard, S. C. Schmidler, and M. Huber. "Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions". *The Annals of Applied Probability* (2009), pp. 617–640.

[YWJ16] Y. Yang, M. J. Wainwright, and M. I. Jordan. "On the computational complexity of high-dimensional Bayesian variable selection". *The Annals of Statistics* 44.6 (2016), pp. 2497–2532.

[ZC04] V. A. Zorich and R. Cooke. *Mathematical analysis II*. Springer Science & Business Media, 2004.