COMPLEXITY RESULTS FOR MCMC DERIVED FROM QUANTITATIVE BOUNDS

JUN YANG AND JEFFREY S. ROSENTHAL

Abstract. This paper considers how to obtain MCMC quantitative convergence bounds which can be translated into tight complexity bounds in high-dimensional setting. We propose a modified drift-and-minorization approach, which establishes a generalized drift condition defined in a subset of the state space. The subset is called the "large set", and is chosen to rule out some "bad" states which have poor drift property when the dimension gets large. Using the "large set" together with a "centered" drift function, a quantitative bound can be obtained which can be translated into a tight complexity bound. As a demonstration, we analyze a certain realistic Gibbs sampler algorithm and obtain a complexity upper bound for the mixing time, which shows that the number of iterations required for the Gibbs sampler to converge is constant. It is our hope that this modified drift-and-minorization approach can be employed in many other specific examples to obtain complexity bounds for high-dimensional Markov chains.

Contents

1. Introduction 1
2. Generalized Geometric Drift Conditions and Large Sets 5
3. Gibbs Sampler Convergence Bound 11
4. Proof of Lemma 3.3 19
5. Proof of Theorem 3.6 26
Acknowledgments 27
References 27
A. Proof of Theorem 2.1 31
B. Proof of Lemma 3.4 35
C. Proof of Lemma 3.5 42

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are extremely widely used and studied in statistics, e.g. [Bro+11; GRS95], and their running

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

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]. 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 "size" of the problem 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 [LV03; Vem05; LV06; WSH09a; WSH09b] and the references therein. 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 how to obtain MCMC quantitative convergence bounds that can be translated into tight complexity bounds in high-dimensional setting. At the first glance, it may seem that an approach to answering the question of convergence complexity may be provided by the drift-and-minorization 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, a few specific upper bounds in the literature obtained by the drift-and-minorization method tend to 1 as n or p tends to infinity. For example, 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" gets large fast as the dimension p increases. And this seems to happen generally when the drift-and-minorization approach is applied to statistical problems. Rajaratnam and Sparks [RS15] also discuss special cases when the method of [Ros95a] can still be used to obtain tight bounds on the convergence rate. However, the conditions proposed in [RS15] are very restrictive. First, it requires the MCMC algorithm to be analyzed is a Gibbs sampler. Second, the Gibbs sampler must have only one high-dimensional parameter which must be drawn in the last step of the Gibbs sampling cycle. Unfortunately, other than some tailored examples [RS15], most realistic MCMC algorithms do not satisfy these conditions. It is unclear whether some particular drift functions lead to bad complexity bounds or the drift-and-minorization approach itself has some limitations. It is therefore the hope by Rajaratnam and Sparks [RS15] 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 the concern on how to obtain quantitative bounds that can be translated into tight complexity bounds. We note that although Rajaratnam and Sparks [RS15] provide evidence

for the claim that many published bounds have poor dependence on n and p, the statistics literature has not focused on controlling the complexity order on n and p. We give some intuitions why most directly translated complexity bounds are quite loose and provide advices on how to obtain tight complexity bounds for high-dimensional Markov chains. The key ideas are (1) the drift function should "capture" the posterior modes as n and/or p goes to infinity and (2) "bad" states which have poor drift property when n and/or p gets large should be ruled out when establishing the drift condition. In order to get tight complexity bounds, we propose a modified drift-and-minorization approach by establishing a generalized drift condition for a subset of the state space, which is called the "large set", instead of the whole state space; see Section 2. The "large set" is chosen to rule out some "bad" states which have poor drift property when the dimension gets large. By establishing the generalized drift condition, a new quantitative bound is obtained, which 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 restricted 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 drift function should be chosen to "capture" the posterior modes (this is called a "centered" drift function in [QH17]), and the "large set" should be adjusted depending on n and p to balance the complexity order of the two parts.

As a demonstration, we prove that a certain realistic Gibbs sampler algorithm converges in $\mathcal{O}(1)$ iterations. To 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 depend on the dimension of the model; see Theorem 3.6. As far as we know, this is the first successful example for analyzing the convergence complexity of a *non-trivial* realistic MCMC algorithm using the (modified) drift-and-minorization approach. Several months after we uploaded this manuscript to arXiv, Qin and Hobert [QH17] successfully analyzed another realistic MCMC algorithm using the drift-and-minorization approach. Although the analysis by Qin and Hobert [QH17] does not make use of the "large set" technique proposed in this paper, they do make use of a "centered" drift function. We explain in this paper that when there exists some "bad" states, using a "centered" drift function might not enough to establish a tight complexity bound. Our modified drift-and-minorization method combining the "large set" technique with

"centered" drift function provides a flexible tool for analyzing convergence complexity. It is our hope that this modified drift-and-minorization 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.

Notations: We use $\xrightarrow{d}$ for weak convergence and $\pi(\cdot)$ to denote the stationary distribution of the Markov chain. The total variance distance is denoted by $\|\cdot\|_{\text{var}}$ and the law of a random variable X denoted by $\mathcal{L}(X)$. We adopt the Big-O, Little-O, Theta, and Omega notations. Formally, $T(n) = \mathcal{O}(f(n))$ if and only if for some constants c and n_0, $T(n) \leq cf(n)$ for all $n \geq n_0$; $T(n) = \Omega(f(n))$ if and only if $T(n) \geq cf(n)$ for all $n \geq n_0$; $T(n)$ is $\Theta(f(n))$ if and only if both $T(n) = \mathcal{O}(f(n))$ and $T(n) = \Omega(f(n))$; $T(n) = o(f(n))$ if and only if $T(n) = \mathcal{O}(f(n))$ and $T(n) \neq \Omega(f(n))$.

2. Generalized Geometric Drift Conditions and Large Sets

Scaling classical MCMCs to very high dimensions can be problematic. Even though 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 $n \to \infty$. For a Markov chain $\{X^{(i)}, i = 0, 1, \dots\}$ on a state space $(\mathcal{X}, \mathcal{B})$ with transition kernel $P(x, \cdot)$, defined by

$$P(x, B) = \mathbb{P}(X^{(i+1)} \in B \,|\, X^{(i)} = x), \quad \forall x \in \mathcal{X}, B \in \mathcal{B} \tag{1}$$

the general method of [Ros95a] proceeds by establishing a *drift condition*

$$\mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x) \leq \lambda f(x) + b, \quad \forall x \in \mathcal{X}, \tag{2}$$

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

$$P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R, \tag{3}$$

where $R := \{x \in \mathcal{X} : f(x) \leq d\}$ is called the "small set", and $d > 2b/(1 - \lambda)$, for some $\epsilon > 0$ and some probability measure $Q(\cdot)$ on $\mathcal{X}$. However, it is observed, for example, in [RS15; QH17], that for many specific bounds obtained by drift-and-minorization method, when the dimension gets larger, the typical scenario for the drift condition of Eq. (2) seems to be λ going to one, and/or b getting much larger. This makes the "size" of the small set R grows too fast, which leads to the minorization volume ϵ goes to 0 exponentially fast. In the following, we give an intuitive explanation what makes a "good" drift condition in the high-dimensional setting.

2.1. **Intuition.** It is useful to think of the drift function $f(x)$ as an energy function [JH01]. Then the drift condition in Eq. (2) implies the chain tends to "drift" toward states which have "lower energy" in expectation. It is well-known that a "good" drift condition is established when both λ and b are small. Intuitively, λ being small implies that when the chain is in a "high-energy" state, then it tends to "drift" back to "low-energy" states fast; and b being small implies when the chain is in a "low-energy" state, then it tends to remain in a "low-energy" state in the next iteration too. In a high-dimensional setting as the dimension grows to infinity, for a collection of drift conditions to be "good", we would like it to satisfy the following two properties:

P1. λ is small, in the sense that it converges to 1 slowly or is bounded away from 1;

P2. b is small, in the sense that it grows at a slower rate than do typical values of the drift function.

One way to understand this intuition is to think of it as controlling the complexity order of the size of the "small set", $R = \{x \in \mathcal{X} : f(x) \leq d\}$. Since $d > 2b/(1-\lambda)$, if λ converges to 1 slowly or is bounded away from 1, and if b is growing at a slower rate than typical values of $f(x)$, then the size of the small set parameter d can be chosen to have a small complexity order. This in turn makes to the minorization volume ϵ converge to 0 sufficiently slowly (or even remain bounded away from 0).

Next, we provide some advices on how to establish such a "good" drift condition in high-dimensional setting.

For clarity, we first assume that λ is bounded away from 1, and focus on conditions required for b to grow at a slower rate than typical values of $f(x)$. Assume for definiteness that p is fixed and $n \to \infty$, and the drift function is scaled in such a way that $f(x) = \mathcal{O}(1)$ and there is a fixed typical state $\tilde{x}$ with $f(\tilde{x}) = \Theta(1)$ regardless of dimension. Then, to satisfy property P2 above, we require that $b = o(1)$. On the other hand, taking expectation over $x \sim \pi(\cdot)$ on both sides of Eq. (2) yields $b \geq \mathbb{E}_\pi[f(x)]/(1-\lambda)$, so $b = \Omega(\mathbb{E}_\pi[f(x)])$. To make $b = o(1)$ implies that the drift function should be chosen such that

$$\mathbb{E}_\pi[f(x)] \to 0.$$

Therefore, to get a small b in a high-dimensional setting, we require a (properly scaled) drift function $f(\cdot)$ whose values $f(x)$, where $x \sim \pi(\cdot)$, concentrate around 0. In particular, if the stationary distribution $\pi(\cdot)$ concentrates near multiple modes as $n \to \infty$, then to make $\mathbb{E}_\pi[f(x)] \to 0$, we require a drift function which "captures" the modes in the sense of nearly vanishing near them. In this paper, we use the name "centered"

drift functions [QH17] to denote drift functions that "capture" the modes of the stationary distribution $\pi(\cdot)$ in this sense.

Note that in the literature, the drift functions used to establish the drift condition are usually not "centered". This is because in the traditional setting where n and p are fixed, a "good" drift condition is established whenever λ and b are small enough for specific fixed values of n and p. The complexity orders of λ and b as functions of n and/or p are not essential, so the property of "capturing" the posterior modes is not necessary for establishing a good drift condition. As a result, many existing quantitative bounds cannot be directly translated into tight complexity bounds, since the size of the small set does not have a small complexity order. At the very least, one has to re-analyze such MCMC algorithms using "centered" drift functions. For example, several months after we uploaded this manuscript to arXiv, Qin and Hobert [QH17] successfully analyzed the Albert and Chib's chain [AC93] using "centered" drift functions. By contrast, the original bounds for this model by Roy and Hobert [RH07] cannot be directly translated into tight complexity orders.

Next, we focus on establishing λ that is either bounded away from 1 or converges to 1 slowly, assuming the drift function is already chosen to be "centered". Intuitively, λ describes the behavior of the Markov chain when its current state has a "high energy". If λ goes to 1 very fast when n and/or p goes to infinity, this may suggest the existence of some "bad" states, i.e. states which have "high energy", but the drift property becomes poor as n and/or p gets large. Therefore, in high dimensions, once the Markov chain visits in one of these "bad" states, it only slowly drifts back toward to the corresponding small set. Since the drift condition in Eq. (2) must hold for all $x \in \mathcal{X}$, the existence of "bad" states forces λ to go to to 1 very fast. And since the small set is defined as $R = \{x \in \mathcal{X} : f(x) \leq d\}$ where $d > 2b/(1-\lambda)$, the scenario $\lambda \to 1$ very fast forces R to become very large, and hence the minorization volume ϵ goes to zero very fast. One perspective on this problem is that the definition of drift condition in Eq. (2) is too restrictive, since it must hold for all states x, even the bad ones.

In summary, we are able to establish a small b as in P2 above by simply using a "centered" drift function. However, we the main difficulty in establishing a small λ as in P1 above is the existence of some "bad" states when n and/or p gets large. Since the traditional drift condition defined in Eq. (2) is restrictive, the traditional drift-and-minorization method is not flexible enough to deal with these "bad" states. In this following, we instead propose a modified drift-and-minorization approach using a generalized drift condition, where the drift function

is defined only in a "large set". This allows us to rule out those "bad" states in high-dimensional cases.

2.2. **New Quantitative Bound.** We propose a new quantitative bound, which is based on a generalized drift condition using a "large set". The main result is summarized in the following theorem.

Theorem 2.1. *Let $\{X^{(k)}\}$ be a Markov chain on a state space $(\mathcal{X}, \mathcal{B})$ with a transition kernel $P(x, \cdot), \forall x \in \mathcal{X}$. Let $P^k(x, \cdot)$ be the k-step transition kernel and π be the stationary distribution of the Markov chain. Suppose for a subset of $\mathcal{X}$, $R_0 \in \mathcal{B}$, there exists a drift function $f : \mathcal{X} \to \mathbb{R}^+$ such that for some $\lambda < 1$ and $b < \infty$, the generalized drift condition holds:*

$$\begin{aligned} \mathbb{E}(f(X^{(1)})) \,|\, X^{(0)} = x, X^{(1)} \in R_0) &\leq \mathbb{E}(f(X^{(1)})) \,|\, X^{(0)} = x) \\ &\leq \lambda f(x) + b, \quad \forall x \in R_0. \end{aligned} \tag{4}$$

Furthermore, suppose that for a "small set" $R := \{x \in \mathcal{X} : f(x) \leq d\}$ where $d > 2b/(1-\lambda)$, the Markov chain also satisfies a minorization condition:

$$P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R, \tag{5}$$

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

(6)

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

where $\alpha^{-1} = \frac{1+2b+\lambda d}{1+d}$, $\Lambda = 1 + 2(\lambda d + b)$ and $\mathbb{E}_\nu[f(x)]$ denotes the expectation of $f(x)$ over $x \sim \nu(\cdot)$.

Proof. See Appendix A. □

Note that the new bound in Theorem 2.1 assumes the Markov chain begins with an initial distribution ν such that $\nu(R_0) = 1$. This assumption is not very restrictive since the "large set" ideally should include all "good" states. In high-dimensional setting, the Markov chain is not expected to converge fast beginning with any state (see Section 3.2.4 for discussions on initial states).

In order to verify the new drift condition of Eq. (4), one has to check a new inequality $\mathbb{E}(f(X^{(1)})) \,|\, X^{(0)} = x, X^{(1)} \in R_0) \leq \mathbb{E}(f(X^{(1)})) \,|\, X^{(0)} =$

x). This implies the "large set" R_0 should be chosen such that the states in R_0 have "lower energy" on expectation. This is intuitive since we assume the "bad" states all have "high energy" and poor drift property when n and/or p gets large. We believe this condition is not very difficult to establish in practice. One trick is to choose R_0 by ruling out more states with "high energy" even if the states are not "bad". In Section 3, we demonstrate the use of this trick to select the "large set" R_0 so that $\mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x, X^{(1)} \in R_0) \leq \mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x)$ can be easily verified.

The condition $\mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x, X^{(1)} \in R_0) \leq \lambda f(x) + b, \forall x \in R_0$ essentially defines a traditional drift condition in Eq. (2) for a new Markov kernel only on the "large set" R_0, which equals 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. (6) are indeed an upper bound on the total variation distance of this "restricted" Markov chain. Note that the general idea of studying the restriction of a Markov chain to some "good" subset of the state space has appeared in the literature, such as [MR00; DF03; Jer+04; Eft+16; MS17] and the references therein, in which different ways of restrictions have been considered for different reasons. Our condition is to assume that the chain has never left the "large set". The goal of considering this "restricted" chain is to obtain better control on the dependence on n and p for the upper bound.

The last two terms in the upper bound Eq. (6) give an upper bound of the probability that the Markov chain will visit R_0^c starting from either the initial distribution ν or the stationary distribution π. Therefore, the proposed method in Theorem 2.1 is a generalized version of the classic drift-and-minorization method [Ros95a] by allowing the drift condition is established in a chosen "large set". Indeed, if we choose $R_0 = \mathcal{X}$, then Eq. (6) reduces to

(7)
$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq (1-\epsilon)^{rk} + \frac{(\alpha\Lambda)^{rk}\left[1 + \mathbb{E}_\nu(f(x)) + \frac{b}{1-\lambda}\right] - \alpha^{rk}}{\alpha^k - \alpha^{rk}}.$$

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

One more note about Eq. (6) is that the new bound does not decrease exponentially with k. For example, the term $k\,\pi(R_0^c)$ is linear increasing with k for fixed n and p. We emphasize that we do not aim to prove a Markov chain is geometrically ergodic here. An upper bound which decreases exponentially with k for fixed n and p does not guarantee to have a tight complexity order, which has been discussed in [RS15].

Instead, our new bound in Eq. (6) is designed for controlling complexity orders of n and/or p for high-dimensional Markov chains.

Furthermore, the Markov chain to be analyzed in Theorem 2.1 does not have to be geometrically ergodic. The proof of Eq. (6) only implies that, after ruling out "bad" states, the "restricted" Markov chain defined in the "large set", using the *conditional* transition kernel $P'(x, B) := P(x, B \cap R_0)/P(x, R_0), \forall x \in R_0, B \in \mathcal{B}$ is geometrically ergodic. Therefore, technically speaking, the new bound in Eq. (6) can be used to analyze non-geometrically ergodic high-dimensional Markov chains.

2.3. **Complexity Bound.** The proposed new bound in Theorem 2.1 can be used to obtain complexity bounds in high-dimensional setting. The key is to balance the complexity orders of k required for both the first two terms and the last two terms of the upper bound in Eq. (6) to be small. The complexity order of k for the first two terms to be small can be controlled by adjusting the "large set". The "large set" should be kept as large as possible provided that "bad" states have been ruled out. The complexity order of k required for the last two terms to be small is determined by

$$k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c) \to 0. \tag{8}$$

This may involve (carefully) bounding the tail probability of the transition kernel, depending on the definition of the "large set" and the complexity order aimed to establish.

In the next section, we employ the modified drift-and-minorization method to prove a certain realistic Gibbs sampler algorithm converges in $\mathcal{O}(1)$. We first choose a particular "centered" drift function $f(x)$ and identify the "bad" states. In our Gibbs sampler example, one coordinate of the state x corresponds to one particular parameter of the MCMC model, and the "bad" states correspond to those whose value of this particular parameter 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 into 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]:

$$\begin{aligned} Y_i \mid \theta_i \quad &\sim \mathcal{N}(\theta_i, V), \quad 1 \le i \le n, \\ \theta_i \mid \mu, A &\sim \mathcal{N}(\mu, A), \quad 1 \le i \le n, \\ \mu &\sim \text{ flat prior on } \mathbb{R}, \\ A &\sim \mathbf{IG}(a, b), \end{aligned} \tag{9}$$

where V is assumed to be known, $(Y_1, \ldots, Y_n)$ is the observed data, and $x = (A, \mu, \theta_1, \ldots, \theta_n)$ 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

$$\begin{aligned} \pi(\cdot) &= \mathcal{L}(A, \mu, \theta_1, \ldots, \theta_n \mid Y_1, \ldots, Y_n) \\ &\propto \frac{b^a}{\Gamma(a)} A^{-a-1} e^{-b/A} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi A}} e^{-\frac{(\theta_i - \mu)^2}{2A}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i - \theta_i)^2}{2V}}. \end{aligned} \tag{10}$$

A Gibbs sampler for the posterior distribution of this model has been originally analyzed in [Ros96]. A quantitative bound has been derived by Rosenthal [Ros96] using the drift-and-minorization method with a drift function $f(x) = \sum_{i=1}^{n} (\theta_i - \bar{Y})^2$. We first observe that this drift function is not "centered". For example, select a "typical" state $\tilde{x} = (\tilde{A}, \tilde{\mu}, \tilde{\theta}_1, \ldots, \tilde{\theta}_n)$ such that $\tilde{\theta}_i = Y_i$, we get $f(\tilde{x}) = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Under reasonable assumptions on the observed data $\{Y_i\}$, we can get the properly scaled drift function $f(\tilde{x})/n = \Theta(1)$. Then if the drift function is "centered", we hope the established b satisfies $b/n = o(1)$. However, $b/n = \sum_{i=1}^{n} (Y_i - \bar{Y})^2/n + \frac{n+1/4}{n} V = \Theta(1)$ in [Ros96]. Furthermore, the established λ in [Ros96] converges to 1 very fast, satisfying $1/(1-\lambda) = \Omega(n)$. Therefore, if we translate the quantitative bound in [Ros96] into complexity orders, it requires the size of the "small set" R be $\Omega(n^2)$, 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].

We demonstrate the use of the modified drift-and-minorization approach by analyzing a Gibbs sampler for this MCMC model. Defining $x^{(k)} = (A^{(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_n^{(k)})$ to be the state of the Markov chain at the k-th iteration, we consider the following order of Gibbs sampling

for computing the posterior distribution:

$$
\begin{aligned}
\mu^{(k+1)} &\sim \mathcal{N}\left(\bar{\theta}^{(k)}, \frac{A^{(k)}}{n}\right), \\
\theta_i^{(k+1)} &\sim \mathcal{N}\left(\frac{\mu^{(k+1)}V + Y_i A^{(k)}}{V + A^{(k)}}, \frac{A^{(k)}V}{V + A^{(k)}}\right), \quad i = 1, \ldots, n, \\
A^{(k+1)} &\sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(k+1)} - \bar{\theta}^{(k+1)})^2\right).
\end{aligned} \tag{11}
$$

We prove that convergence of this Gibbs sampler is actually very fast: 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:

$$
\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \Theta(1), \quad \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \geq V + \delta, \quad \forall n \geq N_0. \tag{12}
$$

Remark 3.1. The assumptions in Eq. (12) are quite natural. For the second assumption, note that our MCMC model implies that the variance of Y_i is larger than V because of the uncertainty of θ_i. Actually, under the MCMC model, conditional on the parameter A, the variance of the data $\{Y_i\}$ equals $V + A$. Therefore, the second assumption in Eq. (12) is just to assume the observed data is not abnormal under the MCMC model when n is large enough. Note that only the existence of δ is required for establishing our main results. More precisely, the existence of δ is needed to obtain an upper bound for $\pi(R_0^c)$. If such δ does not exist, the MCMC model is seriously misspecified so the posterior distribution of the parameter A, which corresponds to the variance of a Normal distribution, may concentrate on 0. In that case, our upper bound on $\pi(R_0^c)$ does not hold. ◁

Then we show that, under the assumption Eq. (12), with initial state

$$
\bar{\theta}^{(0)} = \bar{Y}, \quad A^{(0)} = \begin{cases} \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} - V, & \text{if } \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} > V, \\ \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}, & \text{otherise,} \end{cases} \tag{13}
$$

and $\mu^{(0)}$ arbitrary (since $\mu^{(0)}$ will be updated in the first step of the Gibbs sampler), 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.

3.1. **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 Eq. (13), there exists a positive integer N which does not depend on k, 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*

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq C_1\gamma^k + C_2\frac{k(1+k)}{n} + C_3\frac{k}{\sqrt{n}}. \tag{14}$$

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

Lemma 3.3. *Let $\Delta = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and $x = (A, \mu, \theta_1, \ldots, \theta_n)$. Define the drift function $f(x)$ by*

$$f(x) := n(\bar{\theta} - \bar{Y})^2 + n\left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2. \tag{15}$$

Let $x^{(k)} = (A^{(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_n^{(k)})$ be the state of the Markov chain at the k-th iteration, then we have

$$\mathbb{E}[f(x^{(k+1)}) \,|\, x^{(k)}] \leq \left(\frac{V^2 + 2VA^{(k)}}{V^2 + 2VA^{(k)} + (A^{(k)})^2}\right)^2 f(x^{(k)}) + b, \quad \forall x^{(k)} \in \mathcal{X} \tag{16}$$

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

Proof. See Section 4. □

Note that in Eq. (16), the term $\left(\frac{V^2+2VA^{(k)}}{V^2+2VA^{(k)}+(A^{(k)})^2}\right)^2$ depends on the coordinate $A^{(k)}$ of the state $x^{(k)}$ and is not bounded away from 1, since $A^{(k)}$ can be arbitrarily close to 0. Therefore, $\left(\frac{V^2+2VA^{(k)}}{V^2+2VA^{(k)}+(A^{(k)})^2}\right)^2$ cannot be bounded by some λ such that $0 < \lambda < 1$ and we cannot directly establish the traditional drift condition Eq. (2) by Eq. (16). In the following, we establish the generalized drift condition Eq. (4) using a "large set".

According to Eq. (12), 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 $\lambda_T := \left(\frac{V^2+2VT}{V^2+2VT+T^2}\right)^2 < 1$, we get

$$\mathbb{E}[f(x^{(k+1)}) \,|\, x^{(k)}] \leq \lambda_T\, f(x^{(k)}) + b, \quad \forall x \in R_T. \tag{17}$$

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

$$R_T := \left\{ x \in \mathcal{X} : \left[\left(\frac{\Delta}{n-1} - V \right) - A \right]^2 \leq \left[\left(\frac{\Delta}{n-1} - V \right) - T \right]^2 \right\}. \tag{18}$$

In order to satisfy the new drift condition in Eq. (4), we still need to check

$$\mathbb{E}[f(x^{(k+1)}) \,|\, x^{(k)}, x^{(k+1)} \in R_T] \leq \mathbb{E}[f(x^{(k+1)}) \,|\, x^{(k)}], \quad \forall x \in R_T. \tag{19}$$

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 conditional on $A^{(k)}$, $\{\eta_i^{(k+1)}\}$ are i.i.d. samples from a Normal distribution. Therefore, we have $\bar{\eta}^{(k+1)}$ is conditional independent with $\eta_i^{(k+1)} - \bar{\eta}^{(k+1)}$, which implies $\bar{\theta}^{(k+1)}$ is independent with $\theta_i^{(k+1)} - \bar{\theta}^{(k+1)}$ conditional on $A^{(k)}$. 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 independent with $A^{(k+1)}$ conditional on $x^{(k)}$, therefore

$$\mathbb{E}[(\bar{\theta}^{(k+1)} - \bar{Y})^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T] = \mathbb{E}[(\bar{\theta}^{(k+1)} - \bar{Y})^2 \,|\, x^{(k)}]. \tag{20}$$

Furthermore, by the definition of R_T, we have

$$\begin{aligned} &\left[\left(\frac{\Delta}{n-1} - V \right) - A^{(k+1)} \right]^2 \leq \left[\left(\frac{\Delta}{n-1} - V \right) - T \right]^2, \quad \forall x^{(k+1)} \in R_T \\ &\left[\left(\frac{\Delta}{n-1} - V \right) - A^{(k+1)} \right]^2 > \left[\left(\frac{\Delta}{n-1} - V \right) - T \right]^2, \quad \forall x^{(k+1)} \notin R_T. \end{aligned} \tag{21}$$

Then we have

$$\begin{aligned} &\mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V \right) - A^{(k+1)} \right)^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T \right] \\ &\leq \mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V \right) - T \right)^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T \right] \\ &= \left(\left(\frac{\Delta}{n-1} - V \right) - T \right)^2 \qquad (22) \\ &= \mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V \right) - T \right)^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T^c \right] \\ &\leq \mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V \right) - A^{(k+1)} \right)^2 \,|\, x^{(k)}, x^{(k+1)} \in R_T^c \right], \end{aligned}$$

which implies

$$\begin{aligned}&\mathbb{E}\left[\left(\left(\frac{\Delta}{n-1}-V\right)-A^{(k+1)}\right)^2 \mid x^{(k)}, x^{(k+1)} \in R_T\right] \\ &\leq \mathbb{E}\left[\left(\left(\frac{\Delta}{n-1}-V\right)-A^{(k+1)}\right)^2 \mid x^{(k)}\right].\end{aligned} \tag{23}$$

Therefore, the new drift condition of Eq. (4) 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 3.4. *If $T = \Theta(1)$, by choosing the size of the "small set" $R = \{x \in \mathcal{X} : f(x) \leq d\}$ to satisfy $d = \mathcal{O}(1)$ and $d > \frac{b}{1-\lambda_T}$, the Markov chain satisfies a minorization condition in Eq. (5) with the minorization condition $\epsilon = \Theta(1)$.*

Proof. See Appendix B. □

Lemma 3.5. *With the initial state given by Eq. (13), there exists a positive integer N, which does not depend on k, such that for all $n \geq N$, we have*

$$\begin{aligned}&k\,\pi(R_T^c) + \sum_{i=1}^{k} P^i(x^{(0)}, R_T^c) \\ &\leq \frac{k(1+k)}{2n}\frac{b}{\left[\left(\frac{\Delta}{n-1}-V\right)-T\right]^2} + \frac{k}{\sqrt{n}}\frac{\sqrt{b}(2V/\delta+1)}{\left|\left(\frac{\Delta}{n-1}-V\right)-T\right|}.\end{aligned} \tag{24}$$

Proof. See Appendix C. □

Now we derive a quantitative bound for the Gibbs sampler for large enough n by combing results together. First, from Lemma 3.3, we have $b = \mathcal{O}(1)$. Recall that $\lambda_T = \left(\frac{V^2+2VT}{V^2+2VT+T^2}\right)^2$. We obtain $\frac{b}{1-\lambda_T} = \mathcal{O}(1)$ by choosing $T = \Theta(1)$. Since $d > \frac{b}{1-\lambda_T}$, we can choose the size of small set to be $d = \mathcal{O}(1)$. Then by Lemma 3.4, we obtain the minorization volume $\epsilon = \Theta(1)$. Furthermore, by definition $\alpha^{-1} = \frac{1+2b+\lambda_T d}{1+d} < 1$, it can be verified that α^{-1} is bounded away from 0 when $T = \Theta(1)$ and $d = \mathcal{O}(1)$. Next, since $\Lambda = 1 + 2(\lambda_T d + b) = \Theta(1)$, we choose $r = \log(\alpha)/\log(\alpha\Lambda/(1-\epsilon))$ to balance the order of $(1-\epsilon)^r$ and $\alpha^{-1}(\alpha\Lambda)^r$ and define $\gamma := (1-\epsilon)^r = \alpha^{-1}(\alpha\Lambda)^r$. Then we have $\gamma = \Theta(1)$ and $0 < \gamma < 1$. Furthermore, since $f(x^{(0)}) = 0$ for large enough n and $\frac{b}{1-\lambda_T} = \mathcal{O}(1)$, we can pick a constant C_1 such that $C_1 \geq 2 + \frac{b}{1-\lambda_T}$ for large enough

n. Finally, 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 3.5, then Theorem 3.2 follows from Theorem 2.1. □

Next, we translate the quantitative bound in Theorem 3.2 into the convergence complexity in terms of mixing time. We show the convergence complexity is $\mathcal{O}(1)$. Intuitively, to make the term $C_1\gamma^k$ in Eq. (14) arbitrarily small, k needs to have a complexity order of $\mathcal{O}(1)$ since γ does not depend on n. The residual terms $C_2 \frac{k(1+k)}{n} + C_3 \frac{k}{\sqrt{n}} \to 0$ when $k = o(\sqrt{n})$. Therefore, the complexity bound on the mixing time of the Gibbs sampler equals the smaller complexity order between $\mathcal{O}(1)$ and $o(\sqrt{n})$, which is $\mathcal{O}(1)$. The formal result is given in the following.

Theorem 3.6. *For any $0 < c < 1$, define the mixing time K_c by*

$$K_c(n) := \arg\min_k \left\{ \|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \le c \right\}. \tag{25}$$

Then with the initial state $x^{(0)}$ given by Eq. (13), there exists $N_c = \Theta(1)$ and $\bar{K}_c = \Theta(1)$ such that

$$K_c(n) \le \bar{K}_c, \quad \forall n \ge N_c. \tag{26}$$

Proof. See Section 5. □

3.2. **Discussions.** We give further comments and discussions on the analysis of the Gibbs sampler.

3.2.1. *Drift function.* In the proof of Theorem 3.2, we have used a "centered" drift function shown in Eq. (15). To check this, we select a "typical" state $\tilde{x} = (\tilde{A}, \tilde{\mu}, \tilde{\theta}_1, \ldots, \tilde{\theta}_n)$ such that $\tilde{\theta}_i = Y_i$ and $\tilde{A} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$ then the scaled drift function $f(\tilde{x})/n = nV^2/n = \Theta(1)$. We then hope to establish b such that $b/n = o(1)$, or equivalently, $b = o(n)$. Indeed, the established generalized drift condition has $b = \mathcal{O}(1) = o(n)$, which implies the drift function is "centered".

3.2.2. *"Large set".* The result in Eq. (16) implies that those states whose value of A are close to zero are "bad" states. Therefore, the goal of choosing the "larget set" in Eq. (18) is to rule out those states. Note that we have applied the trick that ruling more states with "high energy" could make the condition Eq. (4) easier to establish. In the "large set" R_T defined by Eq. (18), we have also ruled out the states x whose value of A are larger than $\left|\left(\frac{\Delta}{n-1} - V\right) - T\right| + \left(\frac{\Delta}{n-1} - V\right)$. Note that these states are not "bad" states. However, by ruling out them, the condition Eq. (4) can be verified easily.

3.2.3. *The upper bound in Eq.* (24). Although the upper bound of $k\,\pi(R_T^c) + \sum_{i=1}^{k} P^i(x^{(0)}, R_T^c)$ shown in Eq. (24) is loose, it is already enough for showing the mixing time of the Gibbs sampler is $\mathcal{O}(1)$. The proof of Lemma 3.5 only makes use of the form of drift function and the definition of "large set", and does not depend on the particular form of the transition kernel of the Gibbs sampler. We expect that, in general, tighter upper bound on $k\,\pi(R_T^c) + \sum_{i=1}^{k} P^i(x^{(0)}, R_T^c)$ could be obtained, depending on the choice of "large set" and the MCMC algorithm to be analyzed. This may involve carefully bounding the tail probability of the transition kernel.

3.2.4. *Initial state.* The main results in Theorem 3.2 and Theorem 3.6 hold for a particular initial state given in Eq. (13). We discuss other initial states than the one given in Eq. (13). Note that the new bound in Lemma 3.3 holds for any initial state that is in the "large set". Therefore, we can extend the results in Theorem 3.2 to get bounds when the Markov chain starts from some other initial states in the "large set". Recall the assumption on the observed data $\{Y_i\}$ in Eq. (12), we have assumed there exists $\delta > 0$ such that $\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1} \geq V + \delta$ for large enough n. Note that the existence of such δ is sufficient to obtain the results in Theorem 3.2 and Theorem 3.6. In order to get bounds when the MCMC algorithm starts from other initial states, we assume δ is known and establish upper bounds using δ explicitly. We define the "large set" Eq. (18) using $T = \delta$ and the extension of Theorem 3.2 is given in the following.

Theorem 3.7. *Let* $\Delta = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. *If the Markov chain starts with any initial state* $x^{(0)} \in R_\delta$ *where*

$$R_\delta := \left\{ x \in \mathcal{X} : \left[\frac{\Delta}{n-1} - V - A\right]^2 \leq \left[\frac{\Delta}{n-1} - V - \delta\right]^2 \right\}, \tag{27}$$

there exists a positive integer N, *which does not depend on* k, *some constants* $C_1 > 0, C_2 > 0, C_3 > 0, C_4 > 0$ *and* $0 < \gamma < 1$, *such that for all* $n \geq N$ *and for all* k, *we have*

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq [C_1 + f(x^{(0)})]\gamma^k + C_2\frac{k(1+k)}{n} + C_3\frac{k}{\sqrt{n}} + C_4 f(x^{(0)})\frac{k}{n}, \tag{28}$$

where $f(\cdot)$ *is the drift function defined in Eq.* (15).

Proof. Following the same proof of Theorem 3.2 by keeping the term $f(x^{(0)})$, the first two terms of the upper bound given in Eq. (6) can be replaced by $[C_1 + f(x^{(0)})]\gamma^k$ and the last term of the upper bound in

Eq. (6) can be replaced by $\sum_{i=1}^{k} P^i(x^{(0)}, R_\delta^c) \leq C_2 \frac{k(1+k)}{n} + C_4 f(x^{(0)}) \frac{k}{n}$. □

From Theorem 3.7, we can immediately obtain a complexity bound when the Markov chain starts within a subset of the "large set", which is given in the following. This result suggests that if the Markov chain starts from an initial state which is not "too far" from the state given in Eq. (13), the Markov chain still mixes fast. The mixing time becomes $\mathcal{O}(\log n)$ instead of $\mathcal{O}(1)$.

Corollary 3.8. *For an initial state $x^{(0)} \in \{x \in R_\delta : f(x) = o(n/\log n)\}$, the mixing time of the Gibbs sampler is $\mathcal{O}(\log n)$.*

Note that $\{x \in R_\delta : f(x) = o(n/\log n)\}$ defines a subset of the "large set" R_δ, and the above result shows that the mixing time is $\mathcal{O}(\log n)$ if the initial state is in this subset. We conjecture the same complexity order of $\mathcal{O}(\log n)$ on the mixing time may hold even if the initial state is in a larger subset, for example $\left\{x^{(0)} \in R_\delta : f(x^{(0)}) = \Theta(n)\right\}$. However, in order to prove this, we need to derive tighter upper bound of $\sum_{i=1}^{k} P^i(x^{(0)}, R_\delta^c)$ which is a non-trivial task. We therefore leave it as an open problem.

Finally, we do not have upper bounds for the Markov chain when the initial state is outside of the "large set" since the new bound in Theorem 2.1 requires the Markov chain starts within the "large set". For this particular Gibbs sampler example, numerical experiments suggest that, if the Markov chain starts from a "bad" state, the number of iterations required for the Markov chain to mix can be much larger than $\mathcal{O}(\log n)$. In high-dimensional setting, when the dimension of the state space goes to infinity, the Markov chain may not mix fast starting from any state. This observation is loosely consistent with various observations made by Hairer, Mattingly, and Scheutzow [HMS11].

3.2.5. *Relation to spectral gaps.* Many approaches in MCMC literature bound the spectral gap of the corresponding Markov operator[LV03; Vem05; LV06; WSH09a; WSH09b]. However, on general state spaces, the spectral gap is zero for Markov chains which are not geometrically ergodic, even if they do converge to stationarity. Our results do not require the Markov chain to be geometrically ergodic. Instead, we only require the constructed "restricted" chain on the "large set" in our proof is geometrically ergodic. Therefore, we cannot connect our results to bounds on spectral gaps. Furthermore, we do not require the Markov chain to be reversible. So our results apply even in the non-reversible cases, which makes spectral gaps harder to study or interpret. For these reasons, we do not present the main results in terms of spectral gaps.

4. Proof of Lemma 3.3

Recall that the order of Gibbs sampling for computing the first scan is:

$$\begin{aligned}
\mu^{(1)} &\sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right), \\
\theta_i^{(1)} &\sim \mathcal{N}\left(\frac{\mu^{(1)}V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{V + A^{(0)}}\right), \qquad (29) \\
A^{(1)} &\sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right).
\end{aligned}$$

It suffices to show that for $\Delta = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and

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

we have

$$\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] \le \left(\frac{V^2 + 2VA^{(0)}}{V^2 + 2VA^{(0)} + (A^{(0)})^2}\right)^2 f(x^{(0)}) + b, \qquad (31)$$

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

Note that we can compute the expectation in $\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}]$ by three steps, according to the reverse order of the Gibbs sampling. To simplify the notation, we define σ-algebras that we condition on:

$$\begin{aligned}
\mathcal{G}_A &:= \sigma(A^{(0)}, \{\theta_i^{(1)}\}, \mu^{(1)}), \\
\mathcal{G}_\theta &:= \sigma(A^{(0)}, \{\theta_i^{(0)}\}, \mu^{(1)}), \qquad (32) \\
\mathcal{G}_\mu &:= \sigma(A^{(0)}, \{\theta_i^{(0)}\}, \mu^{(0)}).
\end{aligned}$$

Then we have

$$\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] = \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_\mu] = \mathbb{E}[\mathbb{E}[\mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A] \,|\, \mathcal{G}_\theta] \,|\, \mathcal{G}_\mu]. \qquad (33)$$

The three steps are as follows:

(1) Compute the expectation over $A^{(1)}$ given $\{\theta_i^{(1)}\}$ and $\mu^{(1)}$. This is to compute the conditional expectation

$$f'(x^{(1)}) := \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A], \qquad (34)$$

where we write $\mathbb{E}[\cdot \,|\, \mathcal{G}_A]$ to denote the the expectation is over (recall that a and b are constants from the prior $\mathbf{IG}(a, b)$)

$$A^{(1)} \sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right) \qquad (35)$$

for given $\theta^{(1)}$ and $\mu^{(1)}$.

(2) Compute the expectation over $\{\theta_i^{(1)}\}$ given $\mu^{(1)}$. This is to compute the conditional expectation

$$f''(x^{(1)}) := \mathbb{E}[f'(x^{(1)}) \,|\, \mathcal{G}_\theta], \tag{36}$$

where we use $\mathbb{E}[\cdot \,|\, \mathcal{G}_\theta]$ to denote the expectation is over

$$\theta_i^{(1)} \sim \mathcal{N}\left(\frac{\mu^{(1)}V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{V + A^{(0)}}\right) \tag{37}$$

for given $\mu^{(1)}$ and $A^{(0)}$.

(3) Compute the expectation over $\mu^{(1)}$. This is to compute the conditional expectation

$$\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] = \mathbb{E}[f''(x^{(1)}) \,|\, \mathcal{G}_\mu], \tag{38}$$

where we have used $\mathbb{E}[\cdot \,|\, \mathcal{G}_\mu]$ to denote the expectation is over

$$\mu^{(1)} \sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right) \tag{39}$$

for given $\{\theta_i^{(0)}\}$ and $A^{(0)}$.

In the following, we compute the three steps, respectively. We use $\mathcal{O}(1)$ to denote terms that can be upper bounded by some constant that does not depend on the state.

4.1. **Compute $f'(x^{(1)}) = \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A]$.** The first term of $f(x^{(1)})$ is $n(\bar{\theta}^{(1)} - \bar{Y})^2$, which does not involve $A^{(1)}$. Thus, we only need to compute the conditional expectation of the second term of $f(x^{(1)})$. That is,

$$\begin{aligned} f'(x^{(1)}) &= \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A] \\ &= n(\bar{\theta}^{(1)} - \bar{Y})^2 + n\mathbb{E}\left\{\left[\left(\frac{\Delta}{n-1} - V\right) - A^{(1)}\right]^2 \,|\, \mathcal{G}_A\right\}. \end{aligned} \tag{40}$$

Note that

$$\begin{aligned} &n\mathbb{E}\left\{\left[\left(\frac{\Delta}{n-1} - V\right) - A^{(1)}\right]^2 \,|\, \mathcal{G}_A\right\} \\ &= n\left(\frac{\Delta}{n-1} - V\right)^2 + n\mathbb{E}[(A^{(1)})^2 \,|\, \mathcal{G}_A] - 2n\left(\frac{\Delta}{n-1} - V\right)\mathbb{E}[A^{(1)} \,|\, \mathcal{G}_A]. \end{aligned} \tag{41}$$

Recall that $\mathbb{E}[\cdot \,|\, \mathcal{G}_A]$ denotes that the expectation is over

$$A^{(1)} \sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right), \tag{42}$$

where a and b are constants from the prior $\mathbf{IG}(a, b)$. The mean and variance of $A^{(1)}$ can be written in closed forms since $A^{(1)}$ follows from an inverse Gamma distribution. Denoting $S := \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1}$, we can write the mean of $A^{(1)}$ using S as follows:

$$\begin{aligned}
\mathbb{E}[A^{(1)} \,|\, \mathcal{G}_A] &= \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2 + 2b}{n - 1 + 2(a-1)} \\
&= \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1} + \frac{2b}{n-1+2(a-1)} \\
&\quad - \left(\frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1} \right) \left(\frac{2(a-1)}{n-1+2(a-1)} \right) \\
&= S + \mathcal{O}(1/n) + \mathcal{O}(1/n) S.
\end{aligned} \tag{43}$$

Similarly, the variance of $A^{(1)}$ can be written in terms of S as well:

$$\begin{aligned}
\mathrm{var}[A^{(1)} \,|\, \mathcal{G}_A] &= \frac{(\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2/2 + b)^2}{[(n-1)/2 + (a-1)]^2 [(n-1)/2 + (a-2)]} \\
&= \frac{1}{(n-1)/2 + (a-2)} \left(\mathbb{E}[A^{(1)} \,|\, \mathcal{G}_A] \right)^2 \\
&= \mathcal{O}(1/n) \left(S + \mathcal{O}(1/n) + \mathcal{O}(1/n) S \right)^2 \\
&= \mathcal{O}(1/n) S^2 + \mathcal{O}(1/n^2) S + \mathcal{O}(1/n^3).
\end{aligned} \tag{44}$$

Substituting the mean and variance of $A^{(1)}$ in terms of S, we get

$$\begin{aligned}
& n \mathbb{E} \left\{ \left[\left(\frac{\Delta}{n-1} - V \right) - A^{(1)} \right]^2 \,|\, \mathcal{G}_A \right\} \\
&= n \left(\frac{\Delta}{n-1} - V \right)^2 + n S^2 - 2n \left(\frac{\Delta}{n-1} - V \right) S \\
&\quad + \mathcal{O}(1) + \mathcal{O}(1) S + \mathcal{O}(1) S^2.
\end{aligned} \tag{45}$$

Therefore, we have

$$\begin{aligned}
f'(x^{(1)}) &= \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A] \\
&= n(\bar{\theta}^{(1)} - \bar{Y})^2 + n \left(\frac{\Delta}{n-1} - V \right)^2 + n S^2 - 2n \left(\frac{\Delta}{n-1} - V \right) S \\
&\quad + \mathcal{O}(1) + \mathcal{O}(1) S + \mathcal{O}(1) S^2.
\end{aligned} \tag{46}$$

4.2. **Compute $f''(x^{(1)}) = \mathbb{E}[f'(x^{(1)}) \,|\, \mathcal{G}_\theta]$.** Note that the terms in $f'(x^{(1)})$ involving $\{\theta_i^{(1)}\}$ are $(\bar{\theta}^{(1)} - \bar{Y})^2$ and $S = \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1}$. Then

$$
\begin{aligned}
f''(x^{(1)}) &= \mathbb{E}[f'(x^{(1)}) \,|\, \mathcal{G}_\theta] \\
&= n\mathbb{E}\left[(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_\theta\right] + n\left(\frac{\Delta}{n-1} - V\right)^2 \\
&\quad + n\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta] - 2n\left(\frac{\Delta}{n-1} - V\right)\mathbb{E}[S \,|\, \mathcal{G}_\theta] \\
&\quad + \mathcal{O}(1) + \mathcal{O}(1)\mathbb{E}[S \,|\, \mathcal{G}_\theta] + \mathcal{O}(1)\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta].
\end{aligned} \tag{47}
$$

Therefore, it suffices to compute the following terms

$$
\mathbb{E}\left[(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_\theta\right], \quad \mathbb{E}[S \,|\, \mathcal{G}_\theta], \quad \mathbb{E}[S^2 \,|\, \mathcal{G}_\theta], \tag{48}
$$

where $\mathbb{E}[\cdot \,|\, \mathcal{G}_\theta]$ denotes the expectation is over

$$
\theta_i^{(1)} \sim \mathcal{N}\left(\frac{\mu^{(1)}V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{V + A^{(0)}}\right). \tag{49}
$$

Note that $\{\theta_i^{(1)}\}$ are independent (but not identically distributed) conditional on $\mathcal{G}_\theta$. For the first term $\mathbb{E}\left[(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_\theta\right]$, we have

$$
\begin{aligned}
\mathbb{E}\left[(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_\theta\right] &= \mathbb{E}\left[\left(\bar{\theta}^{(1)} - \frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}} + \frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}} - \bar{Y}\right)^2 \,|\, \mathcal{G}_\theta\right] \\
&= \mathbb{E}\left[\left(\bar{\theta}^{(1)} - \frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}\right)^2 \,|\, \mathcal{G}_\theta\right] + \left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}} - \bar{Y}\right)^2 \\
&\quad + 2\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}} - \bar{Y}\right)\mathbb{E}\left[\left(\bar{\theta}^{(1)} - \frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}\right) \,|\, \mathcal{G}_\theta\right] \\
&= \mathrm{var}[\bar{\theta}^{(1)} \,|\, \mathcal{G}_\theta] + \left(\frac{V}{V + A^{(0)}}\right)^2 \left(\mu^{(1)} - \bar{Y}\right)^2 \\
&= \frac{1}{n}\frac{A^{(0)}V}{V + A^{(0)}} + \left(\frac{V}{V + A^{(0)}}\right)^2 \left(\mu^{(1)} - \bar{Y}\right)^2
\end{aligned} \tag{50}
$$

For the other two terms involving S, we have the following lemma.

Lemma 4.1. *For $S = \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1}$, we have*

$$
\mathbb{E}[S \,|\, \mathcal{G}_\theta] = \frac{A^{(0)}V}{V + A^{(0)}} + \left(\frac{A^{(0)}}{V + A^{(0)}}\right)^2 \frac{\Delta}{n-1}, \quad \mathrm{var}[S \,|\, \mathcal{G}_\theta] = \mathcal{O}(1/n). \tag{51}
$$

Proof. Define $\eta_i := \theta_i^{(1)} - \frac{Y_i A^{(0)}}{V+A^{(0)}}$ then $\bar{\eta} = \bar{\theta}^{(1)} - \frac{\bar{Y} A^{(0)}}{V+A^{(0)}}$. Note that $\{\eta_i\}$ are i.i.d. conditional on $\mathcal{G}_\theta$ with

(52)
$$\eta_i \sim \mathcal{N}\left(\frac{\mu^{(1)}V}{V+A^{(0)}}, \frac{A^{(0)}V}{V+A^{(0)}}\right), \quad \bar{\eta} \sim \mathcal{N}\left(\frac{\mu^{(1)}V}{V+A^{(0)}}, \frac{1}{n}\frac{A^{(0)}V}{V+A^{(0)}}\right).$$

Next, we decompose $\sum_{i=1}^n (\theta_i^{(1)} - \bar{\theta}^{(1)})^2$ by

(53)
$$\begin{aligned}\sum_{i=1}^n (\theta_i^{(1)} - \bar{\theta}^{(1)})^2 &= \sum_{i=1}^n \left(\eta_i + \frac{Y_i A^{(0)}}{V+A^{(0)}} - \bar{\eta} - \frac{\bar{Y} A^{(0)}}{V+A^{(0)}}\right)^2 \\ &= \sum_{i=1}^n \left((\eta_i - \bar{\eta})^2 + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2 (Y_i - \bar{Y})^2 + \frac{2(\eta_i - \bar{\eta})(Y_i - \bar{Y})A^{(0)}}{V+A^{(0)}}\right).\end{aligned}$$

Then we can obtain $\mathbb{E}[S \,|\, \mathcal{G}_\theta]$ by

(54)
$$\begin{aligned}\mathbb{E}[S \,|\, \mathcal{G}_\theta] &= \mathbb{E}\left\{\left[\frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1}\right] \,|\, \mathcal{G}_\theta\right\} \\ &= \mathbb{E}\left\{\left[\frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1}\right] \,|\, \mathcal{G}_\theta\right\} + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2 \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} \\ &= \frac{A^{(0)}V}{V+A^{(0)}} + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2 \frac{\Delta}{n-1}.\end{aligned}$$

For $\mathrm{var}[S \,|\, \mathcal{G}_\theta]$, using Cauchy—Schwarz inequality

(55)
$$\begin{aligned}&\mathrm{var}[S \,|\, \mathcal{G}_\theta] = \mathbb{E}\left[(S - \mathbb{E}[S \,|\, \mathcal{G}_\theta])^2 \,|\, \mathcal{G}_\theta\right] \\ &= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n (\eta_i - \bar{\eta})^2}{n-1} - \mathbb{E}_{\{\eta_i\}}\left[\frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1}\right] + 2\frac{A^{(0)}}{V+A^{(0)}}\frac{\sum_{i=1}^n (\eta_i - \bar{\eta})(Y_i - \bar{Y})}{n-1}\right)^2 \,|\, \mathcal{G}_\theta\right] \\ &\leq 2\,\mathrm{var}\left[\frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1} \,|\, \mathcal{G}_\theta\right] + 8\left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2 \frac{\mathbb{E}\left\{\left[\sum_i (\eta_i - \bar{\eta})(Y_i - \bar{Y})\right]^2 \,|\, \mathcal{G}_\theta\right\}}{(n-1)^2}.\end{aligned}$$

Note that $\{\eta_i\}$ are i.i.d conditional on $\mathcal{G}_\theta$, we know

(56)
$$\mathbb{E}\left\{\left[\frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1}\right]^2 \,|\, \mathcal{G}_\theta\right\} = \left\{\mathbb{E}\left[\frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1} \,|\, \mathcal{G}_\theta\right]\right\}^2 + \mathcal{O}(1/n).$$

That is, $\mathrm{var}\left[\frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1} \,|\, \mathcal{G}_\theta\right] = \mathcal{O}(1/n)$. Finally, the term

$$
\begin{aligned}
&\frac{\mathbb{E}\left\{\left[\sum_i(\eta_i-\bar{\eta})(Y_i-\bar{Y})\right]^2 \,|\, \mathcal{G}_\theta\right\}}{(n-1)^2} \\
&= \frac{\mathbb{E}\left\{\left[\sum_i(\eta_i-\bar{\eta})^2(Y_i-\bar{Y})^2\right] \,|\, \mathcal{G}_\theta\right\} + \mathbb{E}[\bar{\eta}^2 \,|\, \mathcal{G}_\theta]\sum_{i\neq j}(Y_i-\bar{Y})(Y_j-\bar{Y})}{(n-1)^2} \\
&= \frac{\sum_i(Y_i-\bar{Y})^2}{(n-1)^2}\mathbb{E}\left[(\eta_1-\bar{\eta})^2 \,|\, \mathcal{G}_\theta\right] + \mathcal{O}(1/n) \\
&= \frac{\Delta}{(n-1)^2}\frac{(n-1)\frac{A^{(0)}V}{V+A^{(0)}}}{n} + \mathcal{O}(1/n) = \mathcal{O}(1/n).
\end{aligned} \tag{57}
$$

Therefore, we have $\mathrm{var}[S \,|\, \mathcal{G}_\theta] = \mathcal{O}(1/n)$. □

Next, using the following results

$$
\begin{aligned}
\mathbb{E}[S \,|\, \mathcal{G}_\theta] &= \frac{A^{(0)}V}{V+A^{(0)}} + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2\frac{\Delta}{n-1} \\
&\leq V + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2\frac{\Delta}{n-1} = \mathcal{O}(1), \\
\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta] &= \left(\mathbb{E}[S \,|\, \mathcal{G}_\theta]\right)^2 + \mathcal{O}(1/n) = \mathcal{O}(1),
\end{aligned} \tag{58}
$$

we can first write $f''(x^{(1)})$ by

$$
\begin{aligned}
f''(x^{(1)}) =& n\mathbb{E}\left[(\bar{\theta}^{(1)}-\bar{Y})^2 \,|\, \mathcal{G}_\theta\right] + n\left(\frac{\Delta}{n-1} - V\right)^2 \\
&+ n\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta] - 2n\left(\frac{\Delta}{n-1} - V\right)\mathbb{E}[S \,|\, \mathcal{G}_\theta] + \mathcal{O}(1).
\end{aligned} \tag{59}
$$

Then, using

$$
\begin{aligned}
n\mathbb{E}\left[(\bar{\theta}^{(1)}-\bar{Y})^2 \,|\, \mathcal{G}_\theta\right] &= \frac{A^{(0)}V}{V+A^{(0)}} + n\left(\frac{V}{V+A^{(0)}}\right)^2\left(\mu^{(1)}-\bar{Y}\right)^2 \\
&\leq V + \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}
\end{aligned} \tag{60}
$$

we further bound the terms

(61)
$$
\begin{aligned}
&n\mathbb{E}\left[(\bar{\theta}^{(1)}-\bar{Y})^2\,|\,\mathcal{G}_\theta\right]+n\left(\frac{\Delta}{n-1}-V\right)^2\\
&\quad+n\mathbb{E}[S^2\,|\,\mathcal{G}_\theta]-2n\left(\frac{\Delta}{n-1}-V\right)\mathbb{E}[S\,|\,\mathcal{G}_\theta]\\
&\leq\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}+n\left[\left(\frac{\Delta}{n-1}-V\right)-\mathbb{E}[S\,|\,\mathcal{G}_\theta]\right]^2\\
&=\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}+n\left[\frac{A^{(0)}V}{V+A^{(0)}}+\left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2\frac{\Delta}{n-1}-\left(\frac{\Delta}{n-1}-V\right)\right]^2\\
&=\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}+n\left[\frac{\Delta}{n-1}\left[\left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2-1\right]+\left(\frac{A^{(0)}V}{V+A^{(0)}}+V\right)\right]^2\\
&=\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}+n\left(\frac{A^{(0)}}{V+A^{(0)}}+1\right)^2\left[\frac{\Delta}{n-1}\left(\frac{-V}{V+A^{(0)}}\right)+V\right]^2\\
&=\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}+\frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1}-(A^{(0)}+V)\right]^2.
\end{aligned}
$$

Finally, combing all the results yields

(62)
$$
f''(x^{(1)})=\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}+\frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1}-(A^{(0)}+V)\right]^2+\mathcal{O}(1).
$$

4.3. **Compute $\mathbb{E}[f(x^{(1)})\,|\,x^{(0)}]=\mathbb{E}[f''(x^{(1)})\,|\,\mathcal{G}_\mu]$.** Recall that the expectation $\mathbb{E}[\cdot\,|\,\mathcal{G}_\mu]$ is over

(63)
$$
\mu^{(1)}\sim\mathcal{N}\left(\bar{\theta}^{(0)},\frac{A^{(0)}}{n}\right).
$$

In the obtained expression of $f''(x^{(1)})$ from previous step, the only term involves $\mu^{(1)}$ is $\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}$. Since

(64)
$$
\mathbb{E}\left[(\mu^{(1)}-\bar{Y})^2\,|\,\mathcal{G}_\mu\right]=(\bar{\theta}^{(0)}-\bar{Y})^2+A^{(0)}/n,
$$

we have

(65)
$$\begin{aligned}\mathbb{E}[f(x^{(1)})\,|\,x^{(0)}] &= \mathbb{E}[f''(x^{(1)})\,|\,\mathcal{G}_\mu] \\ &\le \frac{nV^2}{(V+A^{(0)})^2}\left((\bar{\theta}^{(0)}-\bar{Y})^2+\frac{A^{(0)}}{n}\right) \\ &\quad + \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1}-(A^{(0)}+V)\right]^2+\mathcal{O}(1) \\ &= \frac{nV^2(\bar{\theta}^{(0)}-\bar{Y})^2}{(V+A^{(0)})^2} \\ &\quad + \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1}-(A^{(0)}+V)\right]^2+\mathcal{O}(1).\end{aligned}$$

Finally, we complete the proof by

(66)
$$\begin{aligned}&\frac{nV^2(\bar{\theta}^{(0)}-\bar{Y})^2}{(V+A^{(0)})^2}+\frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1}-(A^{(0)}+V)\right]^2+\mathcal{O}(1) \\ &= \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left\{\frac{(V+A^{(0)})^2}{(V+2A^{(0)})^2}(\bar{\theta}^{(0)}-\bar{Y})^2+\left[\frac{\Delta}{n-1}-(A^{(0)}+V)\right]^2\right\}+\mathcal{O}(1) \\ &\le \frac{V^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left\{n(\bar{\theta}^{(0)}-\bar{Y})^2+n\left[\frac{\Delta}{n-1}-(A^{(0)}+V)\right]^2\right\}+\mathcal{O}(1) \\ &= \left(\frac{V^2+2VA^{(0)}}{V^2+2VA^{(0)}+(A^{(0)})^2}\right)^2 f(x^{(0)})+\mathcal{O}(1).\end{aligned}$$

5. Proof of Theorem 3.6

Using Theorem 3.2, one sufficient condition for

(67)
$$\|\mathcal{L}(X^{(k)})-\pi\|_{\text{var}} \le c$$

is that $n \ge N$ and

(68)
$$C_1\gamma^k \le \frac{c}{3}, \quad C_2\frac{(1+k)^2}{n} \le \frac{c}{3}, \quad C_3\frac{k}{\sqrt{n}} \le \frac{c}{3}.$$

This requires the number of iterations, k, satisfies

(69)
$$\frac{\log(C_1)-\log(c/3)}{\log(1/\gamma)} \le k \le \min\left\{\sqrt{\frac{c/3}{C_3}}\sqrt{n}-1, \frac{c/3}{C_3}\sqrt{n}\right\}.$$

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

$$\frac{\log(C_1) - \log(c/3)}{\log(1/\gamma)} \leq \min\left\{\sqrt{\frac{c/3}{C_3}}\sqrt{n} - 1, \frac{c/3}{C_3}\sqrt{n}\right\}, \tag{70}$$

which is equivalent to

$$n \geq \max\left\{N, \left[\bar{K}_c \frac{3C_3}{c}\right]^2, \left[\left(\bar{K}_c + 1\right)\sqrt{\frac{3C_3}{c}}\right]^2\right\} =: N_c, \tag{71}$$

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. The authors also thank the associate editor and two anonymous referees for very valuable comments and suggestions. This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[AC93] J. H. Albert and S. Chib. "Bayesian analysis of binary and polychotomous response data". *Journal of the American statistical Association* 88.422 (1993), pp. 669–679.

[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.

[DF03] M. Dyer and A. Frieze. "Randomly coloring graphs with lower bounds on girth and maximum degree". *Random Structures & Algorithms* 23.2 (2003), pp. 167–179.

[Eft+16] C. Efthymiou, T. P. Hayes, D. Štefankovic, E. Vigoda, and Y. Yin. "Convergence of MCMC and loopy BP in the tree uniqueness region for the hard-core model". In: *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE. 2016, pp. 704–713.

[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.

[HMS11] M. Hairer, J. C. Mattingly, and M. Scheutzow. "Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations". *Probability Theory and Related Fields* 149.1-2 (2011), pp. 223–259.

[Jer+04] M. Jerrum, J.-B. Son, P. Tetali, and E. Vigoda. "Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains". *Annals of Applied Probability* (2004), pp. 1741–1765.

[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.

[LV03] L. Lovász and S. Vempala. "Hit-and-run is fast and fun". *preprint, Microsoft Research* (2003).

[LV06] L. Lovász and S. Vempala. "Hit-and-run from a corner". *SIAM Journal on Computing* 35.4 (2006), pp. 985–1005.

[MR00] R. A. Martin and D. Randall. "Sampling adsorbing staircase walks using a new Markov chain decomposition method". In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE. 2000, pp. 492–502.

[MS17] O. Mangoubi and A. Smith. "Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions". *arXiv preprint arXiv:1708.07114* (2017).

[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.

[QH17] Q. Qin and J. P. Hobert. "Asymptotically Stable Drift and Minorization for Markov Chains with Application to Albert and Chib's Algorithm". *arXiv preprint arXiv:1712.08867* (2017).

[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.

[RH07] V. Roy and J. P. Hobert. "Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4 (2007), pp. 607–623.

[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.

[Vem05] S Vempala. "Geometric random walk: a survey". *Combinatorial and computational geometry* 52 (2005), pp. 577–616.

[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.

A. Proof of Theorem 2.1

This proof is a generalization of the original proof of drift-and-minorization method using coupling in [Ros95a]. The main difference is that our proof is by conditioning on the fact that the Markov chain always stays in the "large set".

Suppose that $X^{(m)}$ and $Y^{(m)}$ are two realizations of the Markov chain, where $X^{(m)}$ starts with the initial distribution $\nu(\cdot)$ and $Y^{(m)}$ starts with an initial distribution $\pi(\cdot)$. We define the natural filtrations

$$\mathcal{F}_i := \sigma\left((X^{(m)}, Y^{(m)}), m = 0, 1, \ldots, i\right). \tag{72}$$

Recall that R denotes the "small set" and R_0 denotes the "large set". We define a stopping time

$$\tau := \inf\{m \geq 0 : (X^{(m)}, Y^{(m)}) \notin R_0 \times R_0\} \tag{73}$$

Then τ denotes the first time that either $X^{(m)}$ or $Y^{(m)}$ leaves the large set R_0. Next, we define the hitting times of $(X^{(m)}, Y^{(m)})$ to $R \times R$ as follows.

$$\begin{aligned} t_1 :&= \inf\{m \geq 0 : (X^{(m)}, Y^{(m)}) \in R \times R\}, \\ t_i :&= \inf\{m \geq t_{i-1} + 1 : (X^{(m)}, Y^{(m)}) \in R \times R\}, \quad \forall i > 1. \end{aligned} \tag{74}$$

Let $N_k := \max\{i : t_i < k\}$. Then N_k denotes the number of $(X^{(m)}, Y^{(m)})$ to hit $R \times R$ in the first k iterations. The following result gives an upper bound for $\|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\|_{\text{var}}$.

Lemma A.1. *When the Markov chain satisfies the minorization condition in Eq. (5), for any $j > 0$, we have*

$$\begin{aligned} \|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\|_{\text{var}} \leq &(1-\epsilon)^j + \mathbb{P}(N_k < j \,|\, \tau > k) \\ &+ k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c). \end{aligned} \tag{75}$$

Proof. The result follows from [Ros95a, Theorem 1] together with

$$\begin{aligned} \mathbb{P}(N_k < j) &= \mathbb{P}(N_k < j \,|\, \tau > k)\mathbb{P}(\tau > k) + \mathbb{P}(N_k < j \,|\, \tau \leq k)\mathbb{P}(\tau \leq k) \\ &\leq \mathbb{P}(N_k < j \,|\, \tau > k) + \mathbb{P}(\tau \leq k) \end{aligned} \tag{76}$$

and

$$\begin{aligned}\mathbb{P}(\tau \le k) &= \mathbb{P}\left(\exists\, m \le k : (X^{(m)}, Y^{(m)}) \notin R_0 \times R_0\right) \\ &\le \sum_{m=1}^{k} \mathbb{P}\left(Y^{(m)} \notin R_0\right) + \sum_{m=1}^{k} \mathbb{P}\left(X^{(m)} \notin R_0\right) \\ &\le k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c).\end{aligned} \tag{77}$$

□

Next, we further upper bound the term $\mathbb{P}(N_k < j \,|\, \tau > k)$. Define the i-th gap of return times by $r_i := t_i - t_{i-1}, \forall i > 1$, then

Lemma A.2. *For any $\alpha > 1$, $j > 0$, and $k > j$,*

$$\mathbb{P}(N_k < j \,|\, \tau > k) \le \frac{1}{\alpha^k - \alpha^j}\left[\mathbb{E}\left(\prod_{i=1}^{j} \alpha^{r_i} \,|\, \tau > t_j\right) - \alpha^j\right]. \tag{78}$$

Proof. Note that $\{N_k < j\} = \{t_j \ge k\} = \{t_j \le k-1\}^c \in \mathcal{F}_{k-1}$ and $\{\tau > k\} = \{(X^{(m)}, Y^{(m)}) \in R_0 \times R_0, \forall m \in \{0, \dots, k\}\} \in \mathcal{F}_k$. Therefore, by the Markov property, for all $k' > k$ the events $E_{k,k'} := \{(X^{(m)}, Y^{(m)}) \in R_0 \times R_0, \forall m \in \{k+1, \dots, k'\})\}$ are independent with $\{N_k < j\}$ conditional on $\{\tau > k\}$. Then for any $k' > k$ we get

$$\begin{aligned}\mathbb{P}(N_k < j \,|\, \tau > k') &= \mathbb{P}(N_k < j \,|\, \tau > k, E_{k,k'}) \\ &= \frac{\mathbb{P}(N_k < j, E_{k,k'} \,|\, \tau > k)}{\mathbb{P}(E_{k,k'} \,|\, \tau > k)} \\ &= \frac{\mathbb{P}(N_k < j \,|\, \tau > k)\mathbb{P}(E_{k,k'} \,|\, \tau > k)}{\mathbb{P}(E_{k,k'} \,|\, \tau > k)} \\ &= \mathbb{P}(N_k < j \,|\, \tau > k).\end{aligned} \tag{79}$$

Finally, note that $\{N_k < j\} = \{t_j \ge k\} = \{r_1 + \cdots + r_j \ge k\}$ and $r_1 + \cdots + r_j \ge j$ by definition. Then the result comes from the Markov's inequality

$$\begin{aligned}\mathbb{P}(N_k < j \,|\, \tau > k) &= \mathbb{P}(r_1 + \cdots + r_j \ge k \,|\, \tau > t_j) \\ &= \mathbb{P}(\alpha^{r_1 + \cdots + r_j} - \alpha^j \ge \alpha^k - \alpha^j \,|\, \tau > t_j) \\ &\le \frac{1}{\alpha^k - \alpha^j}\left[\mathbb{E}\left(\prod_{i=1}^{j} \alpha^{r_i} \,|\, \tau > t_j\right) - \alpha^j\right].\end{aligned} \tag{80}$$

□

In the following, we bound $\mathbb{E}\left(\prod_{i=1}^{j}\alpha^{r_i}\,|\,\tau>t_j\right)$ and combine the results together.

Lemma A.3. *If there exists a function $h\geq 1$ such that*

(81)
$$\begin{aligned}&\mathbb{E}[h(X^{(1)},Y^{(1)})\,|\,X^{(0)}=x,Y^{(0)}=y,(X^{(1)},Y^{(1)})\in R_0\times R_0]\\&\quad\leq\mathbb{E}[h(X^{(1)},Y^{(1)})\,|\,X^{(0)}=x,Y^{(0)}=y]\leq\alpha^{-1}h(x,y),\quad\forall(x,y)\in R_0\times R_0,\end{aligned}$$

then for any $0<j<k$, we have

(82)
$$\begin{aligned}&\|\mathcal{L}(X^{(k)})-\pi\|_{\mathrm{var}}\\&\leq(1-\epsilon)^j+\frac{(\alpha\Lambda)^{j-1}\mathbb{E}_{\nu\times\pi}[h(X^{(0)},Y^{(0)})\,|\,(X^{(0)},Y^{(0)})\in R_0\times R_0]-\alpha^j}{\alpha^k-\alpha^j}\\&\quad+k\,\pi(R_0^c)+\sum_{i=1}^{k}P^i(\nu,R_0^c),\end{aligned}$$

where $\mathbb{E}_{\nu\times\pi}[\cdot]$ denotes expectation over $X^{(0)}\sim\mu$ and $Y^{(0)}\sim\pi$, and

$$\Lambda:=\sup_{(x,y)\in(R\times R)\cap(R_0\times R_0)}\mathbb{E}[h(X^{(1)},Y^{(1)})\,|\,X^{(0)}=x,Y^{(0)}=y].\tag{83}$$

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

$$\mathbb{E}\left(\alpha^{r_i}\,|\,\tau>t_j,r_1,\ldots,r_{i-1}\right),i=2,\ldots,j.\tag{84}$$

Note that we have $\sigma(\alpha^{r_i})\subseteq\mathcal{F}_{t_i}$ and $\{\tau>t_j\}\in\mathcal{F}_{t_j}$. Note that $t_j\geq t_{i-1}$ for all $i=2,\ldots,j$. Similar to the argument in Eq. (79), by Markov property, for any $i\leq j$ we have

$$\begin{aligned}\mathbb{E}\left(\alpha^{r_i}\,|\,\tau>t_j,r_1,\ldots,r_{i-1}\right)&=\mathbb{E}\left(\alpha^{r_i}\,|\,\tau>t_i,r_1,\ldots,r_{i-1}\right)\\&=\mathbb{E}\left(\alpha^{t_i-t_{i-1}}\,|\,\tau>t_i,t_1,\ldots,t_{i-1}\right)\\&=\mathbb{E}\left(\alpha^{t_i-t_{i-1}}\,|\,\tau>t_i,t_{i-1}\right)\end{aligned}\tag{85}$$

The residual of the proof is a modification of the original proof of [Ros95a, Lemma 4 and Theorem 5]. Under the assumption in Eq. (81), $g_i(k):=\alpha^k h(X^{(k)},Y^{(k)})1_{\{k\leq t_i\}}$ has non-increasing *conditional* expectation as a function of k for $t_{i-1}\leq k\leq t_i$. Using the Markov property several

times, we have

(86)
$$\begin{aligned}
&\mathbb{E}(\alpha^{t_i - t_{i-1}} \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, \tau > t_i) \\
&\leq \mathbb{E}(\alpha^{-t_{i-1}} g_i(t_i) \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, \tau > t_i) \\
&\leq \mathbb{E}(\alpha^{-t_{i-1}} g_i(t_{i-1}+1) \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, \tau > t_i) \\
&= \alpha \mathbb{E}[h(X^{(t_{i-1}+1)}, Y^{(t_{i-1}+1)}) \,|\, X^{(t_{i-1})}, Y^{(t_{i-1})}, \tau > t_{i-1}+1] \\
&= \alpha \mathbb{E}[h(X^{(1)}, Y^{(1)}) \,|\, (X^{(0)}, Y^{(0)}) \in (R \times R) \cap (R_0 \times R_0), (X^{(1)}, Y^{(1)}) \in R_0 \times R_0] \\
&\leq \alpha \sup_{(x,y) \in (R \times R) \cap (R_0 \times R_0)} \mathbb{E}[h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y, (X^{(1)}, Y^{(1)}) \in R_0 \times R_0] \\
&\leq \alpha \sup_{(x,y) \in (R \times R) \cap (R_0 \times R_0)} \mathbb{E}[h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y] = \alpha \Lambda.
\end{aligned}$$

Finally, the first hitting time $t_1 = r_1$, and

$$\mathbb{E}(\alpha^{r_1} \,|\, \tau > 0) \leq \mathbb{E}[h(X^{(0)}, Y^{(0)}) \,|\, (X^{(0)}, Y^{(0)}) \in R_0 \times R_0]. \tag{87}$$

Then Lemma A.3 follows by combing the results in Lemma A.1 and Lemma A.2. □

Now we prove Theorem 2.1 using Lemma A.3. We set $h(x, y) = 1 + f(x) + f(y)$ and $R = \{x \in \mathcal{X} \,|\, f(x) \leq d\}$. We have from Lemma A.3 that

$$\mathbb{E}[f(X^{(1)}) \,|\, X^{(0)} = x, X^{(1)} \in R_0] \leq \lambda f(x) + b, \forall x \in R_0, \tag{88}$$

Note that for any $x \in R_0$ we have

$$\begin{aligned}
&\mathbb{E}[f(X^{(1)}) \,|\, X^{(0)} = x, X^{(1)} \in R_0] \\
&= \mathbb{E}[f(X^{(1)}) \,|\, X^{(0)} = x, X^{(1)} \in R_0, X^{(0)} \in R_0].
\end{aligned} \tag{89}$$

Consider a "restricted" chain on R_0 with the transition kernel $P'(x, B) = P(x, B \cap R_0)/P(x, R_0), \forall x \in R_0, B \in \mathcal{B}$. Denote $\pi'(\cdot)$ as the stationary distribution of this "restricted" chain. Then we have

(90)
$$\mathbb{E}[f(X^{(1)}) \,|\, X^{(0)} = x, X^{(1)} \in R_0, X^{(0)} \in R_0] = \mathbb{E}_{\pi'}[f(X^{(1)}) \,|\, X^{(0)} = x],$$

where $\mathbb{E}_{\pi'}[\cdot \,|\, X^{(0)} = x]$ denotes the expectation is over $X^{(1)}$ which follows from the conditional transition kernel $P'(x, \cdot)$.

Taking expectations over $x \sim \pi'(\cdot)$ on both side of

$$\mathbb{E}_{\pi'}[f(X^{(1)}) \,|\, X^{(0)} = x] \leq \lambda f(x) + b, \forall x \in R_0, \tag{91}$$

we get

$$\mathbb{E}_{\pi'}[f(x)] \leq \lambda \mathbb{E}_{\pi'}[f(x)] + b, \tag{92}$$

which implies $\mathbb{E}[f(X^{(0)}) \,|\, X^{(0)} \in R_0] = \mathbb{E}_\pi[f(x) \,|\, x \in R_0] = \mathbb{E}_{\pi'}[f(x)] \leq \frac{b}{1-\lambda}$. Also, if $(x,y) \notin R \times R$, we have $h(x,y) \geq 1+d$. Thus

$$\mathbb{E}[h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y] \leq \left(\frac{1+2b+\lambda d}{1+d}\right) h(x,y), \quad \forall (x,y) \in (R \times R)^c \cap (R_0 \times R_0). \tag{93}$$

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

$$\begin{aligned} &\mathbb{E}_{\nu \times \pi}[h(X^{(0)}, Y^{(0)}) \,|\, (X^{(0)}, Y^{(0)}) \in R_0 \times R_0] \\ &= 1 + \mathbb{E}_\nu[f(X^{(0)}) \,|\, X^{(0)} \in R_0] + \frac{b}{1-\lambda}. \end{aligned} \tag{94}$$

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

$$\begin{aligned} \|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq (1-\epsilon)^{rk+1} + \frac{(\alpha\Lambda)^{rk}\left[1 + \mathbb{E}_\nu(f(x)) + \frac{b}{1-\lambda}\right] - \alpha^{rk+1}}{\alpha^k - \alpha^{rk+1}} \\ + k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c). \end{aligned} \tag{95}$$

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

B. Proof of Lemma 3.4

Recall that the small set is defined by $R = \{x \in \mathcal{X} : f(x) \leq d\}$ where $d > 2b/(1-\lambda_T)$ and $x = (\mu, A, \theta_1, \ldots, \theta_n)$. When $b = \mathcal{O}(1)$ and $\lambda_T = \Theta(1)$, we can choose $d = \mathcal{O}(1)$. Our goal is to show the minorization volume ϵ satisfying

$$P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R, \tag{96}$$

is asymptotically bounded away from 0. Denoting $\hat{A} := \frac{\Delta}{n-1} - V$, we have

$$\begin{aligned} R &= \left\{ x \in \mathcal{X} : n(\bar{\theta} - \bar{Y})^2 + n\left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2 \leq d \right\} \\ &\subseteq \left\{ x \in \mathcal{X} : |\bar{\theta} - \bar{Y}| \leq \sqrt{\frac{d}{n}} \right\} \bigcap \left\{ x \in \mathcal{X} : |A - \hat{A}| \leq \sqrt{\frac{d}{n}} \right\} \end{aligned} \tag{97}$$

Denoting

$$R' := \left\{ x \in \mathcal{X} : |\bar{\theta} - \bar{Y}| \leq \sqrt{\frac{d}{n}}, |A - \hat{A}| \leq \sqrt{\frac{d}{n}} \right\} \tag{98}$$

since $R \subseteq R'$, it suffices to show the minorization volume ϵ satisfying

$$P(x^{(0)}, \cdot) \geq \epsilon Q(\cdot), \quad \forall x^{(0)} \in R', \tag{99}$$

is asymptotically bounded away from 0. One common technique to obtain ϵ is by integrating the infimum of densities of $P(x^{(0)}, \cdot)$ where in our case the infimum is over all $\bar{\theta}^{(0)}$ and $A^{(0)}$ such that $|\bar{\theta}^{(0)} - \bar{Y}| \leq \sqrt{\frac{d}{n}}$ and $|A^{(0)} - \hat{A}| \leq \sqrt{\frac{d}{n}}$.

Note that the intuition behind the proof is: since R' is determined by $|\bar{\theta}^{(0)} - \bar{Y}| \leq \sqrt{\frac{d}{n}}$ and $|A^{(0)} - \hat{A}| \leq \sqrt{\frac{d}{n}}$. The size of uncertainties of the initial $\bar{\theta}^{(0)}$ and $A^{(0)}$ is of order $\mathcal{O}(1/\sqrt{n})$. Therefore, for any fixed initial state $x^{(0)} \in R'$, if the transition kernel $P(x^{(0)}, \cdot)$ concentrates at a rate of $\Omega(1/\sqrt{n})$ then ϵ is bounded away from 0.

For the density function of the Markov transition kernel $P(x^{(0)}, \cdot)$, recall the order of Gibbs sampler

$$\begin{aligned}
\mu^{(1)} &\sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right), \\
\theta_i^{(1)} &\sim \mathcal{N}\left(\frac{\mu^{(1)} V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)} V}{V + A^{(0)}}\right), \quad i = 1, \ldots, n \\
A^{(1)} &\sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n} (\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right).
\end{aligned} \tag{100}$$

Then ϵ can be computed using the three steps of integration according to the reverse order of the Gibbs sampler:

(1) For given $\mu^{(1)}$ and $\{\theta_i^{(1)}\}$, integrating the infimum of the density of $A^{(1)}$. Note that the infimum is over a subset of $\bar{\theta}^{(0)}$ and $A^{(0)}$. However,

$$A^{(1)} \sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n} (\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right) \tag{101}$$

does not depend on $\bar{\theta}^{(0)}$ and $A^{(0)}$. Therefore, the integration of the infimum of the density in this step always equals one;

(2) For given $\mu^{(1)}$, integrating the infimum of the densities of $\{\theta_i\}$. We first note that $\{\theta_i\}$ appear in the densities only in the forms of $\bar{\theta}$ and $S = \frac{\sum_i (\theta_i - \bar{\theta})^2}{n-1}$. Therefore, instead of integrating over

$(\theta_1, \ldots, \theta_n)$ we can integrate over $\bar{\theta}$ and S. Furthermore, we have shown $\bar{\theta}$ is conditional independent with S given A in the proof of Lemma 4.1, we can integrate them separately. Finally, we note that the infimum is over $\left\{A^{(0)} : |A^{(0)} - \hat{A}| \leq \sqrt{\frac{d}{n}}\right\}$. Overall, we need to lower bound $\tilde{g}_n(\mu^{(1)})$ which is defined by

(102)

$$\begin{aligned}\tilde{g}_n(\mu^{(1)}) &:= \int \mathrm{d}S\mathrm{d}\bar{\theta} \inf_{x^{(0)} \in R'} \left\{ f_S(A^{(0)}, n; S)\, \mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}; \bar{\theta}\right)\right\} \\ &\geq \left[\int \mathrm{d}S \inf_{x^{(0)} \in R'} f_S(A^{(0)}, n; S)\right] \\ &\quad \cdot \left[\int \mathrm{d}\bar{\theta} \inf_{x^{(0)} \in R'} \mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}; \bar{\theta}\right)\right],\end{aligned}$$

where $f_S(A^{(0)}, n; S)$ denotes the density function of $S = \frac{\sum_i (\theta_i - \bar{\theta})^2}{n-1}$ for given $A^{(0)}$, with

$$\theta_i \sim \mathcal{N}\left(\frac{\mu^{(1)}V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{V + A^{(0)}}\right), \quad i = 1, \ldots, n, \tag{103}$$

and $\mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}; \bar{\theta}\right)$ denotes the density function of

$$\bar{\theta} \sim \mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}\right). \tag{104}$$

(3) Finally, we integrate the infimum of the densities of $\mu^{(1)}$ to get ϵ. That is,

$$\epsilon = \int \mathrm{d}\mu \left\{ \tilde{g}_n(\mu) \inf_{x^{(0)} \in R'} \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}; \mu\right)\right\}. \tag{105}$$

In the following, we show ϵ is lower bounded away from 0 in three steps.

First, it is easy to see that the density of S does not depend on $\mu^{(1)}$. We show

$$\int \mathrm{d}S \inf_{x^{(0)} \in R'} f_S(A^{(0)}, n; S) = \Theta(1). \tag{106}$$

Second, we show

(107)

$$\int \mathrm{d}\bar{\theta} \inf_{x^{(0)} \in R'} \mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}; \bar{\theta}\right) \geq 1 - \mathrm{erf}\left(\frac{C|\mu| + C'}{\sqrt{2}}\right)$$

where $\mathrm{erf}(z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t$ and C and C' are some constants.

Finally, we complete the proof by showing

$$\int \mathrm{d}\mu \left\{ \left(1 - \mathrm{erf}(\frac{C|\mu| + C'}{\sqrt{2}})\right) \inf_{x^{(0)} \in R'} \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}; \mu\right) \right\} = \Theta(1). \tag{108}$$

B.1. **Proof of Eq. (106).** We omit the superscripts for simplicity. That is, we show

$$\int \mathrm{d}S \inf_{\left\{A:|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} f_S(A, n; S) = \Theta(1). \tag{109}$$

Following the proof of Lemma 4.1 from Eq. (53) to Eq. (57), defining

$$\eta_i := \theta_i - \frac{Y_i A}{V + A} \sim \mathcal{N}\left(\frac{\mu V}{V + A}, \frac{AV}{V + A}\right), \tag{110}$$

we know

$$\mathbb{E}\left[\left|S - \frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1} - \left(\frac{A}{V+A}\right)^2 \frac{\Delta}{n-1}\right|^2\right] = \mathcal{O}(1/n). \tag{111}$$

Therefore, defining

$$S' := \frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1} + \left(\frac{A}{V+A}\right)^2 \frac{\Delta}{n-1} \tag{112}$$

and denoting $f'_{S'}(A, n; S')$ as the density of S', it suffices to show

$$\int \mathrm{d}S' \inf_{\left\{A:|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} f'_{S'}(A, n; S') = \Theta(1). \tag{113}$$

Furthermore, note that under $|A - \hat{A}| \le \sqrt{\frac{d}{n}}$, we have $\frac{V+A}{AV} = \frac{V+\hat{A}}{\hat{A}V} + \mathcal{O}(1/\sqrt{n}) = \Theta(1)$. Then it suffices to show

$$\int \mathrm{d}S'' \inf_{\left\{A:|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} f''_{S''}(A, n; S'') = \Theta(1), \tag{114}$$

where

$$S'' := \frac{V+A}{AV} S' = \frac{V+A}{AV} \frac{\sum_i (\eta_i - \bar{\eta})^2}{n-1} + \frac{1}{V}\left(\frac{A}{V+A}\right)\frac{\Delta}{n-1} \tag{115}$$

and $f''_{S''}(A, n; S'')$ is the density function of S''.

Next, note that $\frac{V+A}{AV}\sum_i (\eta_i - \bar{\eta})^2 \sim \chi^2_{n-1}$, we have

$$\frac{\frac{V+A}{AV}\sum_i (\eta_i - \bar{\eta})^2 - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{d} \mathcal{N}(0, 1), \tag{116}$$

which does not depend on n. We define $\tilde{f}(z, A; x), \forall z \in \mathbb{R}$ as the density function of a random variable

$$\tilde{X}_{z,A} := z + \frac{\frac{V+A}{AV}\sum_i(\eta_i - \bar{\eta})^2 - (n-1)}{\sqrt{2(n-1)}}, \tag{117}$$

then we know $\tilde{X}_{z,A} \xrightarrow{d} \mathcal{N}(z, 1)$.

The rest of the proof is first to lower bound $\int \mathrm{d}S'' \inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} f''_{S''}(A, n; S'')$ using the density function $\tilde{f}(z, A; x)$ and then show it is asymptotically lower bounded away from 0.

Notice that $\frac{1}{V}\left(\frac{A}{V+A}\right)\frac{\Delta}{n-1}$ is not random, and there exists a constant C_0 such that

$$\left(\max_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}} \frac{A}{V+A} - \min_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}} \frac{A}{V+A}\right)\frac{\Delta/V}{n-1} \leq \frac{C_0}{\sqrt{n-1}}. \tag{118}$$

Finally we have

$$\begin{aligned}
&\int \mathrm{d}S'' \inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} f''_{S''}(A, n; S'') \\
&\geq \inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} \int \mathrm{d}x \min\left\{\tilde{f}\left(-\frac{C_0}{\sqrt{2}}, A; x\right), \tilde{f}\left(+\frac{C_0}{\sqrt{2}}, A; x\right)\right\} \\
&= 1 - \sup_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}} \int_{-\sqrt{2}C_0}^{\sqrt{2}C_0} \mathrm{d}x \tilde{f}(0, A; x) \\
&= 1 - \sup_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}} \mathbb{P}(-\sqrt{2}C_0 \leq \tilde{X}_{0,A} \leq \sqrt{2}C_0) \\
&\to 1 - \int_{-\sqrt{2}C_0}^{\sqrt{2}C_0} \mathrm{d}x \mathcal{N}(0, 1; x) = \Theta(1).
\end{aligned} \tag{119}$$

B.2. **Proof of Eq. (107).** We again omit the subscripts for simplicity. The goal is to lower bound

$$\int \mathrm{d}\bar{\theta} \inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\frac{\mu V + \bar{Y}A}{V+A}, \frac{AV}{n(V+A)}; \bar{\theta}\right) \tag{120}$$

Note that there exists some constants C_1 and C_2 such that

$$\max_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\frac{\mu V+\bar{Y}A}{V+A}-\min_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\frac{\mu V+\bar{Y}A}{V+A}\leq\frac{C_1|\mu|+C_2}{\sqrt{n}}, \tag{121}$$

and another constant C_3 such that

$$\min_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\frac{AV}{n(V+A)}\geq\frac{C_3}{n}. \tag{122}$$

Therefore, we have

$$\begin{aligned}
&\int \mathrm{d}\bar{\theta}\inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\mathcal{N}\left(\frac{\mu V+\bar{Y}A}{V+A},\frac{AV}{n(V+A)};\bar{\theta}\right)\\
&\geq 2\int_{(C_1|\mu|+C_2)/\sqrt{n}}^{\infty}\mathrm{d}x\,\mathcal{N}(0,C_3/n;x)\\
&=2\int_{C_4|\mu|+C_5}^{\infty}\mathrm{d}x\,\mathcal{N}(0,1;x)\\
&=1-\mathrm{erf}\left(\frac{C_4|\mu|+C_5}{\sqrt{2}}\right),
\end{aligned} \tag{123}$$

where $C_4:=\frac{C_1}{\sqrt{C_3}}$ and $C_5:=\frac{C_2}{\sqrt{C_3}}$.

B.3. **Proof of Eq. (108).** We omit the subscripts for simplicity. We show the following is asymptotically bounded away from 0:

$$\int \mathrm{d}\mu\left\{\left(1-\mathrm{erf}\left(\frac{C_4|\mu|+C_5}{\sqrt{2}}\right)\right)\inf_{x\in R'}\mathcal{N}\left(\bar{\theta},\frac{A}{n};\mu\right)\right\} \tag{124}$$

Note that there exists $A'_n\in[\hat{A}-\sqrt{d/n},\hat{A}+\sqrt{d/n}]$ such that

$$\begin{aligned}
&\inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\leq\sqrt{\frac{d}{n}},|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\mathcal{N}\left(\bar{\theta},\frac{A}{n};\mu\right)\\
&=\min\left\{\mathcal{N}\left(\bar{Y}-\sqrt{\frac{d}{n}},\frac{A'_n}{n};\mu\right),\mathcal{N}\left(\bar{Y}+\sqrt{\frac{d}{n}},\frac{A'_n}{n};\mu\right)\right\}
\end{aligned} \tag{125}$$

Therefore, we have

(126)

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathrm{d}\mu \left\{ \left(1 - \operatorname{erf}\left(\frac{C_4|\mu| + C_5}{\sqrt{2}}\right)\right) \inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\le\sqrt{\frac{d}{n}},|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\} \\
&\ge \int_{0}^{2\bar{Y}} \mathrm{d}\mu \left\{ \left(1 - \operatorname{erf}\left(\frac{C_4|\mu| + C_5}{\sqrt{2}}\right)\right) \inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\le\sqrt{\frac{d}{n}},|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\} \\
&\ge \left(1 - \operatorname{erf}\left(\frac{C_4|2\bar{Y}| + C_5}{\sqrt{2}}\right)\right) \int_{0}^{2\bar{Y}} \mathrm{d}\mu \inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\le\sqrt{\frac{d}{n}},|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \\
&= \left(1 - \operatorname{erf}\left(\frac{C_4|2\bar{Y}| + C_5}{\sqrt{2}}\right)\right) \\
&\quad \cdot \left[\int_{0}^{\bar{Y}} \mathrm{d}\mu\, \mathcal{N}\left(\bar{Y} + \sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_{\bar{Y}}^{2\bar{Y}} \mathrm{d}\mu\, \mathcal{N}\left(\bar{Y} - \sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \right] \\
&= \left(1 - \operatorname{erf}\left(\frac{C_4|2\bar{Y}| + C_5}{\sqrt{2}}\right)\right) \\
&\quad \cdot \left[\int_{-\bar{Y}}^{0} \mathrm{d}\mu\, \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_{0}^{\bar{Y}} \mathrm{d}\mu\, \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \right]
\end{aligned}$$

Finally, we show

$$\int_{-\bar{Y}}^{0} \mathrm{d}\mu\, \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_{0}^{\bar{Y}} \mathrm{d}\mu\, \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \tag{127}$$

is asymptotically bounded away from 0. Note that when $n \to \infty$, we have $A'_n \to \hat{A}$. So the density functions $\mathcal{N}\left(\pm\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right)$ concentrate

on 0. Therefore

$$
\begin{aligned}
&\int_{-\bar{Y}}^{0} \mathrm{d}\mu\, \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_{0}^{\bar{Y}} \mathrm{d}\mu\, \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \\
&\to \int_{-\infty}^{0} \mathrm{d}\mu\, \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{\hat{A}}{n}; \mu\right) + \int_{0}^{\infty} \mathrm{d}\mu\, \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{\hat{A}}{n}; \mu\right) \\
&= 1 - \int_{-\sqrt{d/n}}^{\sqrt{d/n}} \mathrm{d}x\, \mathcal{N}\left(0, \frac{\hat{A}}{n}; x\right) \\
&= 1 - \int_{-\sqrt{d}}^{\sqrt{d}} \mathrm{d}x\, \mathcal{N}(0, \hat{A}; x) = \Theta(1).
\end{aligned} \tag{128}
$$

C. Proof of Lemma 3.5

We first consider a Markov chain starting from initial state $x^{(0)}$ defined by Eq. (13). By Eq. (12), we have $A^{(0)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} - V$ for large enough n, which implies $f(x^{(0)}) = 0$. Therefore, for large enough n, we have $\mathbb{E}(f(x^{(1)})) \leq b$ from Lemma 3.3. Furthermore, we can continue to get upper bounds $\mathbb{E}(f(x^{(i)})) \leq ib$ for all $i = 1, \ldots, k$. This implies

$$
\mathbb{E}\left[\left(\frac{\Delta}{n-1} - V\right) - A^{(i)}\right]^2 \leq i\frac{b}{n}, \quad i = 1, \ldots, k. \tag{129}
$$

By the Markov's inequality, we have

$$
\mathbb{P}\left(\left|A^{(i)} - \left(\frac{\Delta}{n-1} - V\right)\right| \geq \left|T - \left(\frac{\Delta}{n-1} - V\right)\right|\right) \leq \frac{i}{n}\frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}, \tag{130}
$$

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

$$
\sum_{i=1}^{k} P^i(x^{(0)}, R_T^c) \leq \frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2} \sum_{i=1}^{k} \frac{i}{n} = \frac{k(1+k)}{2n} \frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}. \tag{131}
$$

Next, we consider a Markov chain starting from π. According to Lemma 3.3, we have

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(1-\left(\frac{V^2+2VA}{V^2+2VA+A^2}\right)^2\right)f(x)\right] \\
&=\mathbb{E}_\pi\left[\left(1+\frac{V^2+2VA}{V^2+2VA+A^2}\right)\left(1-\frac{V^2+2VA}{V^2+2VA+A^2}\right)f(x)\right] \\
&=\mathbb{E}_\pi\left[\left(1+\frac{V^2+2VA}{V^2+2VA+A^2}\right)\left(\frac{A}{V+A}\right)^2 f(x)\right]\le b,
\end{aligned} \tag{132}$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expectation is over $x\sim\pi(\cdot)$. Note that by reverse Hölder's inequality

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(1+\frac{V^2+2VA}{V^2+2VA+A^2}\right)\left(\frac{A}{V+A}\right)^2 f(x)\right] \\
&\ge \mathbb{E}_\pi\left[\left(\frac{A}{V+A}\right)^2 f(x)\right] \\
&\ge [\mathbb{E}_\pi(f(x)^{\frac{1}{2}})]^2\left[\mathbb{E}_\pi\left(\frac{A}{V+A}\right)^{-1}\right]^{-2} \\
&= [\mathbb{E}_\pi(f(x)^{\frac{1}{2}})]^2[\mathbb{E}_\pi(1+V/A)]^{-2}.
\end{aligned} \tag{133}$$

Therefore, we have

$$\mathbb{E}_\pi(f(x)^{\frac{1}{2}})\le\sqrt{b}[1+V\mathbb{E}_\pi(1/A)] \tag{134}$$

Next, we show $\mathbb{E}_\pi(1/A)\le 2/\delta$ for large enough n.

Lemma C.1. *There exists a positive integer N, which only depends on a, b, V, and δ, such that for all $n\ge N$, we have*

$$\mathbb{E}_\pi(1/A)\le 2/\delta. \tag{135}$$

Proof. The posterior distribution can be written as

$$\pi(x\,|\,Y_1,\ldots,Y_n)=\frac{f_a(x,Y_1,\ldots,Y_n)}{\int f_a(x,Y_1,\ldots,Y_n)\mathrm{d}x}, \tag{136}$$

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

(137)
$$\begin{aligned} & f_a(x, Y_1, \ldots, Y_n) \\ &= \frac{b^a}{\Gamma(a)} A^{-a-1} e^{-b/A} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi A}} e^{-\frac{(\theta_i-\mu)^2}{2A}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i-\theta_i)^2}{2V}} \\ &= \frac{1}{(2\pi)^n} \frac{b^a}{\Gamma(a)} A^{-a-1-\frac{n}{2}} e^{-b/A} \exp\left[-\sum_{i=1}^{n} \left(\frac{(\theta_i-\mu)^2}{2A} + \frac{(Y_i-\theta_i)^2}{2V} \right) \right]. \end{aligned}$$

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

$$\mathbb{E}_\pi(1/A) = \frac{a}{b} \frac{\int f_{a+1}(x, Y_1, \ldots, Y_n) \mathrm{d}x}{\int f_a(x, Y_1, \ldots, Y_n) \mathrm{d}x}. \tag{138}$$

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.

Using the fact that

$$\begin{aligned} & \int \exp\left[-\left(\frac{V(\theta_i-\mu)^2 + A(Y_i-\theta_i)^2}{2AV} \right) \right] \mathrm{d}\theta_i \\ &= \left(\int \exp\left[-\frac{\left(\theta - \frac{V\mu + YA}{A+V}\right)^2}{\frac{2AV}{A+V}} \right] \mathrm{d}\theta_i \right) \left(\exp\left[-\frac{(Y_i-\mu)^2}{2(V+A)} \right] \right) \\ &= \sqrt{2\pi \frac{2AV}{V+A}} \exp\left[-\frac{(Y_i-\mu)^2}{2(V+A)} \right], \end{aligned} \tag{139}$$

and

$$\begin{aligned} & \int \exp\left[-\frac{\sum_{i=1}^{n} (Y_i-\mu)^2}{2(V+A)} \right] \mathrm{d}\mu \\ &= \left(\int \exp\left[-\frac{(\mu-\bar{Y})^2}{2(V+A)/n} \right] \mathrm{d}\mu \right) \left(\exp\left[-\frac{\sum_i Y_i^2 - n\bar{Y}^2}{2(V+A)} \right] \right) \\ &= \exp\left[-\frac{\sum_{i=1}^{n} (Y_i-\bar{Y})^2}{2(V+A)} \right] \sqrt{2\pi \frac{2(V+A)}{n}}, \end{aligned} \tag{140}$$

we can write $\mathbb{E}_\pi(1/A)$ as a function of $\Delta = \sum_i (Y_i - \bar{Y})^2$. Denote $h_n(\Delta) := \mathbb{E}_\pi(1/A)$, then we have

$$h_n(\Delta) := \frac{\int A^{-a-2} e^{-b/A} (V+A)^{\frac{-n+1}{2}} \exp\left[-\frac{\Delta}{2(V+A)} \right] \mathrm{d}A}{\int A^{-a-1} e^{-b/A} (V+A)^{\frac{-n+1}{2}} \exp\left[-\frac{\Delta}{2(V+A)} \right] \mathrm{d}A}. \tag{141}$$

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

$$\begin{aligned}&\int A^{-a-1}e^{-b/A}(V+A)^{\frac{-n+1}{2}}\exp\left[-\frac{\Delta}{2(V+A)}\right]\mathrm{d}A\\&=\int A^{-a-1}e^{-b/A}\left\{\frac{1}{\sqrt{V+A}}\exp\left[-\frac{\frac{\Delta}{n-1}}{2(V+A)}\right]\right\}^{n-1}\mathrm{d}A.\end{aligned}\tag{142}$$

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

$$\begin{aligned}h_n((n-1)(c+V)) &= \frac{c^{-a-2}e^{-b/c}\left[y_0\exp\left(-\frac{c+V}{2}y_0^2\right)\right]^{n-1}(1+\mathcal{O}(n^{-\frac{1}{2}}))}{c^{-a-1}e^{-b/c}\left[y_0\exp\left(-\frac{c+V}{2}y_0^2\right)\right]^{n-1}(1+\mathcal{O}(n^{-\frac{1}{2}}))}\\&=\frac{1}{c}(1+\mathcal{O}(n^{-1/2})),\end{aligned}\tag{143}$$

where the term $\mathcal{O}((n^{-1/2})$ only depends on constants a, b, and V. Finally, since for all $n \geq N_0$ we have $\Delta \geq (n-1)(V+\delta)$, this implies $h_n(\Delta) \leq \frac{1}{\delta}(1+\mathcal{O}(n^{-1/2})), \forall n \geq N_0$. Therefore, there exists large enough positive integer N, which only depends on a, b, V, and δ, such that for all $n \geq N$, we have $\mathbb{E}_\pi(1/A) = h_n(\Delta) \leq \frac{1}{\delta}(1+\mathcal{O}(n^{-1/2})) \leq \frac{2}{\delta}$. □

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

$$\mathbb{E}_\pi\left(\left|\left(\frac{\Delta}{n-1}-V\right)-A\right|\right) \leq \sqrt{\frac{b}{n}}(2V/\delta+1).\tag{144}$$

Thus, by the Markov's inequality

$$\begin{aligned}\pi(R_T^c) &= \mathbb{P}_\pi\left(\left|\left(\frac{\Delta}{n-1}-V\right)-A\right| \geq \left|\left(\frac{\Delta}{n-1}-V\right)-T\right|\right)\\&\leq \frac{\sqrt{\frac{b}{n}}(2V/\delta+1)}{\left|\left(\frac{\Delta}{n-1}-V\right)-T\right|}.\end{aligned}\tag{145}$$

Finally, we have

$$
\begin{aligned}
& k\,\pi(R_T^c) + \sum_{i=1}^{k} P^i(x^{(0)}, R_T^c) \\
& \leq \frac{k(1+k)}{2n} \frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2} + \frac{k}{\sqrt{n}} \frac{\sqrt{b}(2V/\delta + 1)}{\left|\left(\frac{\Delta}{n-1} - V\right) - T\right|}.
\end{aligned} \tag{146}
$$