OPTIMAL SCALING OF RANDOM-WALK METROPOLIS ALGORITHMS ON GENERAL TARGET DISTRIBUTIONS

JUN YANG[1], GARETH O. ROBERTS[2], AND JEFFREY S. ROSENTHAL[1]

ABSTRACT. One main limitation of the existing optimal scaling results for Metropolis–Hastings algorithms is that the assumptions on the target distribution are unrealistic. In this paper, we consider optimal scaling of random-walk Metropolis algorithms on general target distributions in high dimensions arising from practical MCMC models from Bayesian statistics. For optimal scaling by maximizing expected squared jumping distance (ESJD), we show the asymptotically optimal acceptance rate 0.234 can be obtained under general realistic sufficient conditions on the target distribution. The new sufficient conditions are easy to be verified and may hold for some general classes of MCMC models arising from Bayesian statistics applications, which substantially generalize the product i.i.d. condition required in most existing literature of optimal scaling. Furthermore, we show one-dimensional diffusion limits can be obtained under slightly stronger conditions, which still allow dependent coordinates of the target distribution. We also connect the new diffusion limit results to complexity bounds of Metropolis algorithms in high dimensions.

CONTENTS

[1]DEPARTMENT OF STATISTICAL SCIENCES, UNIVERSITY OF TORONTO, CANADA

[2]DEPARTMENT OF STATISTICS, UNIVERSITY OF WARWICK, UK

E-mail addresses: jun@utstat.toronto.edu, gareth.o.roberts@warwick.ac.uk, jeff@math.toronto.edu.

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms [BGJM11; GRS95; Liu08; MT12; RC04] are now routinely used in many fields to obtain approximations of integrals that could not be tackled by common numerical methods, because of the simplicity and the scalability to high-dimensional settings. The running times of MCMC algorithms are an extremely important issue of practice. They have been studied from a variety of perspectives, including convergence "diagnostics" via the Markov chain output [GR92], proving weak convergence limits of sped-up versions of the algorithms to diffusion limits [RGG97; RR98], directly bounding the convergence in total variation distance [MT94; Ros95; Ros96; RT99; JH01; Ros02; JH04; Bax05; FHJ08], and non-asymptotic guarantees when the target distribution has a smooth and log-concave density, e.g. [BREZ18; Dal17; DCWY18; DK19] and the references therein.

The optimal scaling framework [RGG97; RR98; RR01] is one of the most successful and practically useful ways of performing asymptotic analysis of MCMC methods in high-dimensions. Optimal scaling results (e.g. [CRR05; NR06; Béd08; BR08; NR08; NR11; NRY12; JLM15; JLM14; RR14; ZBK17]) facilitate optimization of MCMC performance by providing clear and mathematically-based guidance on how to tune the parameters defining the proposal distribution for Metropolis–Hastings algorithms [MRRT+53; Has70]. For instance, classical results include tuning the acceptance probabilities to 0.234 for random-walk Metropolis algorithm (RWM) [RGG97] and 0.574 for Metropolis-adjusted Langevin algorithm (MALA) [RR98]. Moreover, optimal scaling results have been used to analyze and compare a wide variety of MCMC algorithms, such as Hamiltonian Monte Carlo (HMC) [BPRSS+13], Pseudo-Marginal MCMC [STRR15], multiple-try MCMC [BDM12]. This yields guidance which is widely used by practitioners, especially via self-tuning or Adaptive MCMC methodologies [AT08; Ros11].

In the original paper, Roberts, Gelman, and Gilks [RGG97] dealt with the RWM algorithm starting in stationarity for target distributions which have i.i.d. product forms. The i.i.d. condition for the target and the assumption for the chain to start in stationarity are two main limitations of the optimal scaling framework. Particularly, the product i.i.d. condition is very restrictive. From a practitioner's perspective, target distributions of the i.i.d. forms are too limited a class of probability distributions to be useful, since they can be tackled by sampling a single one-dimensional target due to the product structure. To this day, optimal scaling results have mainly been proved for target distributions with a product structure, which severely limits their applicability. On the other hand, practitioners use these tuning criteria far outside the class of target distributions of product i.i.d. forms. For example, extensive simulations [RR01; SFR10] show that these optimality results also hold for more complex target distributions.

There exists only a few extensions for correlated targets and most of them are derived for very specific models. For example, Breyer and Roberts [BR00] studied target densities which are Gibbs measures and Roberts and Rosenthal [RR01] studied inhomogeneous target densities. Breyer, Piccioni, and Scarlatti [BPS04] studied target distributions arising in nonlinear regression and have a mean field structure. Neal and Roberts [NR06] considered

the case where updates of high-dimensional Metropolis algorithms are lower dimensional than the target density itself. Later, Bédard and Rosenthal [BR08] studied independent targets with different scales (see also [Béd07; Béd08]) and Bédard [Béd19] studied a special family of hierarchical target distributions. Neal and Roberts [NR08] studied spherically constrained target distributions and non-Gaussian proposals [NR11]. Sherlock and Roberts [SR09] considered elliptically symmetric unimodal targets. Neal, Roberts, and Yuen [NRY12] studied densities with bounded support. Durmus, Le Corff, Moulines, and Roberts [DLCMR17] considered target distributions which are differentiable in L^p mean. Recently, Mattingly, Pillai, and Stuart [MPS12] studied diffusion limits for a class of high-dimensional measures found from the approximation of measures on a Hilbert space which are absolutely continuous with respect to a Gaussian reference measure (See also [PST12; BRS09; BRSV08; CRSW13]). Important examples of this scenario required by [MPS12] in uncertainty quantification problems are given in [HSV11; Stu10; CDPS18]. However, in this paper we shall concentrate on the situation where absolute continuity with respect to a Gaussian is not a reasonable assumption, as is the case in many Bayesian statistics applications.

Furthermore, we do not consider the transient phase of the Metropolis–Hasting algorithms in this paper. The transient phase of high-dimensional Metropolis–Hasting algorithms are studied for example in [CRR05; JLM14; JLM15; KOS18; KOS19]. Kuntz, Ottobre, and Stuart [KOS19] studied the RWM algorithm starting out of stationarity in the settings of [MPS12; JLM15] when non-product target distributions are defined in a Hilbert space being absolute continuous with respect to some Gaussian measures. Such target distributions in [KOS19] can arise for example in Bayesian nonparametric settings, but not in many other Bayesian statistics applications which we focus on in this paper.

In this paper, we consider optimal scaling of RWM algorithms on general target distributions in high dimensions arising from practical MCMC models in Bayesian statistics. First, for optimal scaling by maximizing expected squared jumping distance (ESJD), we show the asymptotically optimal acceptance rate 0.234 can be obtained under general sufficient conditions on the target distribution. Very briefly speaking, 0.234 is asymptotically optimal if (i) each coordinate of the Markov chain is only strongly dependent with a subset of other coordinates (see assumptions A1 and A3); (ii) the target distribution satisfies some smoothness conditions (see assumptions A2 and A4); (iii) as the dimension goes to infinity, a key quantity of "roughness" of the target concentrates to a nonzero value (see assumption A5). The new sufficient conditions are easy to check in practice and may hold for some general classes of practical MCMC models. Our results substantially generalize the commonly used product i.i.d. condition. Furthermore, we show one-dimensional diffusion limits can also be obtained under relaxed conditions which still allow dependent coordinates of the target distribution. Finally, we also connect the new results of diffusion limits to complexity bounds of RMW algorithms in high dimensions. Note that although the whole paper is focused on RWM algorithm, we believe the technical proofs in this paper can be used to relax restrictive conditions on the target distribution for more general Metropolis algorithms.

The paper is organized as follows. In Section 2, we give a brief background review of optimal scaling for Metropolis–Hastings algorithms and complexity bounds via diffusion limits. In Section 3, we present our main results, which include three parts: optimal

scaling by maximizing ESJD, optimal scaling via diffusion limits, and complexity bounds via diffusion limits. In Section 4, we demonstrate the new optimal scaling result holds for some useful MCMC models. In Appendix A, we prove Theorem 3.10, which is one of our main results. The proofs of lemmas used for proving Theorem 3.10 and other main results, such as Theorems 3.19 and 3.21, are delayed to Appendices B to D.

2. Background on Optimal Scaling

Practical implementations of Metropolis–Hastings algorithms suffer from slow mixing for at least two reasons: the Markov chain moves very slowly to the target distribution when the proposed jumps are too short; the Markov chain stays at a state for most of the time when the proposed jumps are long but the chain ends up in low probability areas of the target distribution. The optimal scaling problem [RGG97] considers the choice of proposed distribution to optimize mixing of the Metropolis–Hastings algorithm. We focus on one of the most popular MCMC algorithms, the RWM algorithm. This algorithm proceeds by running a Markov chain $\{X^d(t), t = 0, \ldots, \infty\}$ as follows. Given a target distribution π^d on the state space $\mathbb{R}^d$ and the current state $X^d(t) = x^d$, a new state is proposed by $Y^d \sim \mathcal{N}(x^d, \sigma_d^2 I)$, which is sampled from a multivariate Gaussian distribution centered at x^d, then the proposal is accepted with probability $\min\{1, \pi^d(Y^d)/\pi^d(x^d)\}$ so that $X^d(t+1) = Y^d$. Otherwise the proposal is rejected and $X^d(t + 1) = x^d$. This is precisely to ensure the Markov chain is reversible with respect to the target distribution π^d. It can be shown that the normal proposals automatically make the RWM algorithm π^d-irreducible, aperiodic, and hence ergodic [RS94; MT96]. Therefore, it will converge asymptotically to π^d in law. Note that the only computational cost involved in calculating the acceptance probabilities is the relative ratio of densities. Within the class of all Metropolis–Hastings algorithms, the RWM algorithm is still widely used in many applications because of its simplicity and robustness.

2.1. **Optimal Scaling via Diffusion Limits.** The most common technique to prove optimal scaling results is to show a weak convergence to diffusion limits as the dimension of a sequence of target densities converges to infinity [RGG97; RR98]. More specifically, even though different coordinates of the Markov chain are *not* independent *nor* even individually Markovian, when the proposal is appropriately scaled according to the dimension, the sequence of sped-up stochastic processes formed by one fixed coordinate of each Markov chain converges to an appropriate Markovian Langevin diffusion process. The limiting diffusion limit admits a straightforward efficiency maximization problem which leads to asymptotically optimal acceptance rate of the proposed moves for the Metropolis–Hastings algorithm. In [RGG97], the target distribution π^d is assumed to be an d-dimensional product density with respect to Lebesgue measure, that is

$$\pi^d(x^d) = \prod_{i=1}^{d} f(x_i), \tag{1}$$

where $x^d = (x_1, x_2, \ldots, x_d)$. It is shown that with the choice of scaling $\sigma_d^2 = \ell^2/(d-1)$ for some fixed $\ell > 0$, individual components of the resulting Markov chain converge to the solution of a stochastic differential equation (SDE). More specifically, denoting $X^d =$

$(X_1^d, X_2^d, \ldots, X_d^d)$, the first coordinate of the RWM algorithm, X_1^d, sped up by a factor of d, i.e. $\{X_1^d(\lfloor dt \rfloor), t = 0, 1, \ldots\}$, converges weakly in the usual Skorokhod topology to a limiting ergodic Langevin diffusion.

Proposition 2.1. *[RGG97, Theorem 1.1] Suppose density f satisfies that f'/f is Lipschitz continuous and*

$$\int \left[\frac{f'(x)}{f(x)} \right]^8 f(x) \mathrm{d}x < \infty, \quad \int \left[\frac{f''(x)}{f(x)} \right]^4 f(x) \mathrm{d}x < \infty. \tag{2}$$

Then for $U^d(t) := X_1^d(\lfloor dt \rfloor)$, as $d \to \infty$, we have $U^d \Rightarrow U$, where $\Rightarrow$ denotes weak convergence in Skorokhod topology, and U satisfies the following Langevin SDE

$$\mathrm{d}U(t) = (h(\ell))^{1/2} \mathrm{d}B(t) + h(\ell) \frac{f'(U(t))}{2f(U(t))} \mathrm{d}t, \tag{3}$$

with $h(\ell) := 2\ell^2 \Phi(-\ell\sqrt{\tilde{I}}/2)$ is the speed measure for the diffusion process, $\tilde{I} := \int \left[\frac{f'(x)}{f(x)} \right]^2 f(x) \mathrm{d}x$, and Φ being the standard Gaussian cumulative density function.

This weak convergence result leads to the interpretation that, started in stationarity and applied to target measures of the i.i.d. form, the RWM algorithm will take on the order of d steps to explore the invariant measure. Furthermore, it may be shown that the value of ℓ which maximizes the speed measure $h(\ell)$ and, therefore, maximizes the speed of convergence of the limiting diffusion, leads to a universal acceptance probability, for the RWM algorithm applied to targets of i.i.d. forms, of approximately 0.234. Proposition 2.1 is proved in [RGG97] using the generator approach [EK86]. The same method of proof has also been applied to derive optimal scaling results for other types of MCMC algorithms: for example, the convergence of MALA to diffusion limits when $\sigma_d^2 = \ell^2/d^{1/3}$ (see e.g. [RR98; RR01; BPS04; CRR05; NR06]) with asymptotically optimal acceptance rate 0.574.

2.2. **Optimal Scaling by maximizing ESJD.** Another popular technique to prove optimal scaling is by maximizing expected squared jumping distance (ESJD) [PG10; ARR11; RR14], which is defined as follows.

Definition 2.2. (Expected Squared Jumping Distance)

$$\mathrm{ESJD}(d) := \mathbb{E}_{X^d \sim \pi^d} \mathbb{E}_{Y^d} \left[\|Y^d - X^d\|^2 \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right] \tag{4}$$

where the expectation over Y^d is taken for $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1}I)$ for given $X^d = x^d$, and $\| \cdot \|$ denotes the Euclidean distance, i.e. $\|Y^d - X^d\|^2 = \sum_{i=1}^d (Y_i - X_i)^2$.

Choosing a proposal variance to maximize ESJD is equivalent to minimizing the first-order auto-correlation of the Markov chain, and thus maximizing the efficiency if the higher order auto-correlations are monotonically increasing with respect to the first-order auto-correlation [PG10]. Furthermore, if weak convergence to a diffusion limit is established, then the ESJD converges to the quadratic variance of the diffusion limit. This suggests that maximizing the ESJD is a reasonable problem. For example, Atchadé, Roberts, and Rosenthal [ARR11]

considered to maximize the ESJD to choose optimal temperature spacings for Metropolis-coupled Markov chain Monte Carlo and simulated tempering algorithms. Later, Roberts and Rosenthal [RR14] proved a diffusion limit for the simulated tempering algorithms. Using a new comparison of asymptotic variance of diffusions, Roberts and Rosenthal [RR14] showed the results in the choice of temperatures in [ARR11] does indeed minimize the asymptotic variance of all functionals. Another example is the optimal scaling result for HMC, with asymptotically optimal acceptance rate 0.651 when $\sigma_d^2 = \ell^2/d^{1/4}$ [BPRSS+13], is proven by maximizing the ESJD.

Although establishing weak convergence of diffusion limits gives stronger guarantee than maximizing ESJD, the price to pay is to require stronger conditions on the target distribution. Maximizing ESJD instead can lead to (much) weaker conditions on the target distribution. Later in this paper, we will show that we are able to relax the restrictive product i.i.d. condition on the target distribution for both cases. In particular, the new sufficient conditions on the target distribution for maximizing ESJD are weak enough to allow target distributions arising from realistic MCMC models.

2.3. **Background on Complexity Bounds.** Because of the big data world, in recent years, there is much interest in the "large d, large n" or "large d, small n" high-dimensional regime, where d 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. This requires the study of computer-science-style complexity bounds [Cob65; Coo71] in terms of running time complexity order as the "size" of the problem goes to infinity. In the Markov chain context, computer scientists have been bounding convergence times of Markov chain algorithms focusing largely on spectral gap bounds for Markov chains [SJ89; LV03; Vem05; LV06; WSH09a; WSH09b]. In contrast, statisticians usually study total variation distance or other metric for MCMC algorithms. In order to bridge the gap between statistics-style convergence bounds, and computer-science-style complexity results, in one direction, Yang and Rosenthal [YR17] recently show that complexity bounds for MCMC can be obtained by quantitative bounds using a modified drift-and-minorization approach. In another direction, Roberts and Rosenthal [RR16] connect existing results on diffusion limits of MCMC algorithm to the computer science notion of algorithm complexity. The main result in [RR16] states that any weak limit of a Markov process implies a corresponding complexity bound in an appropriate metric. More specifically, Roberts and Rosenthal [RR16] connect the diffusion limits to complexity bound using the Wasserstein metric. Let $(\mathcal{X}, \mathcal{F}, \rho)$ be a general measurable metric space, the distance of a stochastic process $\{X(t)\}$ on $(\mathcal{X}, \mathcal{F})$ to its stationary distribution π is defined by the KR distance

$$\|\mathcal{L}_x(X(t)) - \pi\|_{\mathrm{KR}} := \sup_{g \in \mathrm{Lip}_1^1} |\mathbb{E}[g(X(t))] - \pi(g)| \tag{5}$$

where $\mathcal{L}_x(X(t))$ denotes the law of $X(t)$ conditional on starting at $X(0) = x$, $\pi(g) := \int g(x)\pi(\mathrm{d}x)$ is the expected value of g with respect to π, 'KR' stands for 'Kantorovich–Rubinstein', and Lip_1^1 is the set of all functions g from $\mathcal{X}$ to $\mathbb{R}$ with Lipschitz constant no

larger than 1 and with $|g(x)| \leq 1$ for all $x \in \mathcal{X}$, i.e.

$$\mathrm{Lip}_1^1 := \{g : \mathcal{X} \to \mathbb{R}, |g(x) - g(y)| \leq \rho(x, y), \forall x, y \in \mathcal{X}, |g| \leq 1\}. \tag{6}$$

Note that the KR distance defined in Eq. (5) is exactly the 1-st Wasserstein metric. Then it can be shown that the π-average of the KR distance to stationarity from all initial states $X(0)$ in $\mathcal{X}$ is non-increasing, which leads to the following complexity linking proposition.

Proposition 2.3. *[RR16, Theorem 1] Let $X^d = \{X^d(t), t \geq 0\}$ be a stochastic process on $(\mathcal{X}, \mathcal{F}, \rho)$, for each $d \in \mathbb{N}$. Suppose X^d converges weakly in the Skorokhod topology as $d \to \infty$ to a càdlàg process X^∞. Assume these processes all have the same stationary distribution π and that X^∞ converges weakly to π. Then for any $\epsilon > 0$, there are $D < \infty$ and $T < \infty$ such that*

$$\mathbb{E}_{X^d(0) \sim \pi} \|\mathcal{L}_{X^d(0)}(X^d(t)) - \pi\|_{KR} < \epsilon, \quad \forall t \geq T, d \geq D. \tag{7}$$

Proposition 2.3 allows us to bound the convergence of the sequence of processes uniformly over all sufficiently large d, if the sequence of Markov processes converges weakly to a limiting ergodic process. Combining Proposition 2.3 with previously-known MCMC diffusion limit results, Roberts and Rosenthal [RR16] prove that the RWM algorithm in d dimensions takes $\mathcal{O}(d)$ iterations to converge to stationarity. However, in [RR16], the target distribution needs to be product i.i.d. with density satisfies all the assumptions of Proposition 2.1. Furthermore, the condition Eq. (2) is replaced by a stronger condition

$$\int \left[\frac{f'(x)}{f(x)}\right]^{12} f(x)\mathrm{d}x < \infty, \quad \int \left[\frac{f''(x)}{f(x)}\right]^{6} f(x)\mathrm{d}x < \infty. \tag{8}$$

3. Main Results

In this section, we show our main results on optimal scaling of RWM algorithms on general target distributions. We first consider optimal scaling by maximizing ESJD in Section 3.1. We show asymptotic form of the ESJD in Theorem 3.10 under very mild conditions on the target distribution. Then we show in Theorem 3.13 that if we directly maximize the asymptotic ESJD, we can obtain 0.234 as an upper bound of the asymptotically optimal acceptance rate. Next, we show the acceptance rate 0.234 is asymptotically optimal under one more weak law of large number (WLLN) condition on the target distribution in Theorem 3.14. In order to give the reader a brief idea that to what extend the class of target distributions can be enlarged. We first present an example of a non-product non-i.i.d. class of distributions, which is a straightforward corollary of our main result in Theorem 3.14. Note that our main result includes much more general class of distributions that this simple example. Recall that a (probabilistic) graphical model is a family of probability distributions defined in terms of a directed or undirected graph [Jor04]. Suppose that the statistical model can be represented as a graphical model, then we have the following corollary.

Corollary 3.1. *(A Simple Corollary of Theorem 3.14) If the following three conditions hold, 0.234 is indeed the asymptotic acceptance rate: (i) in the graph representation, each node of the graph has at most $o(d^{1/4})$ links; (ii) the target density π^d is bounded and $\log \pi^d$ has up to*

the third bounded partial derivatives; (iii) for $X^d \sim \pi^d$, $\frac{1}{d} \sum_{i=1}^{d} \left(\frac{\partial}{\partial x_i} \log \pi^d(X^d) \right)^2$ converges to a positive constant as $d \to \infty$.

In Section 3.2, we consider optimal scaling via diffusion limits. We prove the new conditions for weak convergence to diffusion limits in Theorem 3.19. We then strengthen this result to consider fixed starting state in Theorem 3.21. Finally, in Section 3.3, we apply our new result on diffusion limits with fixed starting state to obtain complexity bounds for the RMW algorithm, which is given in Corollary 3.23.

Before presenting our main results, we first define a sequence of "sets of typical states".

Definition 3.2. We call $\{F_d\}$ a sequence of "sets of typical states" if $\pi^d(F_d) \to 1$.

Next, we enlarge $\{F_d\}$ in different ways, which will be used later for the new conditions on the target.

Definition 3.3. For a given sequence of "sets of typical states" $\{F_d\}$, we define

$$F_d^{(i)} := \{(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d) : \exists (x_1, \ldots, x_d) \in F_d, \text{such that } |y - x_i| < \sqrt{\log d/d}\}. \tag{9}$$

Furthermore, we define $F_d^+ := \bigcup_{i=1}^{d} F_d^{(i)}$.

Remark 3.4. It is clear from the definitions that $F_d^{(i)}$ is to enlarge the i-th coordinate of $x^d \in F_d$ by covering it with an open interval $(x_i - \sqrt{\log d/d}, x_i + \sqrt{\log d/d})$; F_d^+ is the union of $F_d^{(i)}$, $i = 1, \ldots, d$. Then clearly we have $F_d \subseteq F_d^{(i)} \subseteq F_d^+$. In practice, the difference between F_d^+ and F_d is usually asymptotically ignorable in high dimensions. ◁

Finally, we introduce the idea of "neighborhoods" of a coordinate, which is later used to capture the correlation among different coordinates. We use $\mathcal{H}_i$ to denote a collection of coordinates which are called "neighborhoods" of coordinate i. That is, $\mathcal{H}_i \subseteq \{1, \ldots, d\}$. We also assume $i \in \mathcal{H}_i$. Although the definition of the set $\mathcal{H}_i$ is quite arbitrary, we expect that $j \in \mathcal{H}_i$ implies the coordinates i and j are correlated even conditional on all other coordinates. This idea of "neighborhoods" become clearer if the target distribution comes from a model which can be written as a probabilistic graphical model [Jor04]. For a graphical model, it is convenient to define the "neighborhood" $j \in \mathcal{H}_i$ if there is an edge between nodes i and j. In this definition, clearly $j \notin \mathcal{H}_i$ implies that the two coordinates i and j are conditional independent given all the other $d - 2$ coordinates.

3.1. **Optimal Scaling for Maximizing ESJD.** Suppose $\{F_d\}$ is a sequence of "sets of typical states" and $\{\mathcal{H}_i\}$ are collections of "neighborhoods" for each coordinate. Throughout the paper, we assume $\sup_{i \in \{1, \ldots, d\}} |\mathcal{H}_i| < l_d$ where $l_d = o(d)$.

Remark 3.5. For graphical models, if we define $\mathcal{H}_i$ as the collection of nodes that is directly connected to i by an edge, then $l_d = o(d)$ rules out "dense graphs" for which $l_d \propto d$. ◁

Now we introduce the first assumption A1 on the target π^d.

$$\sup_{(i,j):j \notin \mathcal{H}_i} \sup_{x^d \in F_d^+} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(1), \qquad \sup_{(i,j):j \in \mathcal{H}_i} \sup_{x^d \in F_d^+} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(\sqrt{d/l_d}). \tag{A1}$$

Remark 3.6. For graphical models, if node i is not directly connected to node j, we always have $\frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = 0$. Therefore, in order to make A1 hold, it suffices to check for each edge of the graph, say (i, j), that $\frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(\sqrt{d/l_d})$. Since we have assumed $l_d = o(d)$, this is a very weak condition. For example, A1 holds for all graphical models with bounded second partial derivatives. ◁

Next, we denote the conditional density of the i-th and j-th coordinates, given all the other coordinates fixed, by $\pi_{i,j|-i-j} := \pi^d(x_i, x_j | x_{-i-j})$ where x_{-i-j} with $i < j$ denotes all coordinates of x^d other than the i-th, and j-th coordinates, i.e.

$$x_{-i-j} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d).$$

Note that $\pi_{i,j|-i-j}$ is a probability measure in $\mathbb{R}^2$. Then we introduce the next assumption A2 on the target as follows.

$$\sup_{(i,j):j\notin\mathcal{H}_i} \sup_{\{x_{-i-j}:x^d\in F_d\}} \int \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_i^2} \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_j^2} \frac{1}{\pi_{i,j|-i-j}} \mathrm{d}x_i \mathrm{d}x_j = o(1). \tag{A2}$$

Remark 3.7. The assumption A2 is very weak, since it is only to require that the target has a "flat tail". To see this, consider the target distribution π^d has the special i.i.d. product form of Eq. (1), then A2 reduces to

$$\int \frac{\partial^2 f(x_i)f(x_j)}{\partial x_i^2} \frac{\partial^2 f(x_i)f(x_j)}{\partial x_j^2} \frac{1}{f(x_i)f(x_j)} \mathrm{d}x_i \mathrm{d}x_j = \left(\int \frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} \mathrm{d}x \right)^2 = 0, \tag{10}$$

when f has a "flat tail" so that $\frac{\mathrm{d}f(x)}{\mathrm{d}x} \to 0$ when $|x| \to \infty$. Similarly, for graphical models, if there is no edge between i and j, then when π^d has "flat tail" we have $\int \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_i^2} \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_j^2} \frac{1}{\pi_{i,j|-i-j}} \mathrm{d}x_i \mathrm{d}x_j = 0$. ◁

The next assumption is about conditions on the third partial derivatives.

$$\sup_{(i,j):j\notin\mathcal{H}_i} \sup_{x^d\in\mathbb{R}^d} \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i^2 \partial x_j} = o(1), \quad \sup_{(i,j):j\in\mathcal{H}_i} \sup_{x^d\in\mathbb{R}^d} \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i^2 \partial x_j} = o(d/l_d),$$

$$\sup_{i} \sup_{x^d\in\mathbb{R}^d} \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i^3} = o(d^{1/2}), \quad \sum_{i\neq j\neq k} \left(\sup_{x^d\in\mathbb{R}^d} \left| \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i \partial x_j \partial x_k} \right| \right) = o(d^{3/2}). \tag{A3}$$

Remark 3.8. We consider graphical models that satisfy A3. The first three equations of A3 are similar to A1 and they hold for all graphical models with bounded third partial derivatives. Recall that, in graph theory, a n-clique of a graph is a fully-connected subset of nodes of the graph with cardinality n. The last equation of A3 then involves the number of 3-cliques in the graph. Note that for many realistic hierarchical models, there are no 3-cliques for the corresponding graphs, which implies $\sum_{i\neq j\neq k} \left| \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i \partial x_j \partial x_k} \right| = 0$. Even for the

worst case, considering a graph that has d nodes and each has l_d neighbors, since there are $dl_d/2$ links, the number of 3-cliques is at most $\binom{l_d}{2}d/3 = \mathcal{O}(l_d^2 d)$. Therefore, A3 holds for any graphical model with $l_d = o(d^{1/4})$ and bounded third partial derivatives. ◁

The next assumption is the last assumption before our first main result. We first define a quantity which measures the "roughness" of $\log \pi^d$.

$$I_d(x^d) := \frac{1}{d} \sum_{i=1}^{d} \left(\frac{\partial}{\partial x_i} \log \pi^d(x^d) \right)^2. \tag{11}$$

Similarly, we can consider $I_d(X^d)$ where $X^d \sim \pi^d$ as a random variable. Later we will see that it turns out that $I_d(X^d)$ is a key quantity for optimal scaling results. Assumption A4 is as follows.

There exists α with $0 < \alpha < 1/2$ such that

$$\sup_i \sup_{x^d \in F_d^{(i)}} \frac{\partial \log \pi^d(x^d)}{\partial x_i} = \mathcal{O}(d^\alpha), \quad \sup_{x^d \in F_d^+} \pi^d(x^d) = o(d^{1/2-\alpha}), \quad \sup_{x^d \in F_d^+} 1/I_d(x^d) = \mathcal{O}(d^{\alpha/2}). \tag{A4}$$

Remark 3.9. For A4, the first two conditions do not even require π^d and the first partial derivative of $\log \pi^d$ to be bounded. Thus, they are quite weak. For the last condition, although the mode of π^d is ruled out from F_d^+, the condition can hold as long as $\sup_i \sup_{x^d \in F_d^{(i)}} \frac{\partial \log \pi^d(x^d)}{\partial x_i} = \mathcal{O}(d^{\alpha/2})$ and $I_d(X^d)$ is tight. That is, $\forall 0 < \epsilon < 1$, there exists $K_\epsilon > 0$ such that $\mathbb{P}(I_d(X^d) > K_\epsilon) < 1 - \epsilon$. To see this, one can choose F_d using the tightness such that $\sup_{x^d \in F_d} 1/I_d(x^d) = \mathcal{O}(d^{\alpha/2})$. Then we can replace F_d by F_d^+ since $\inf_{x^d \in F_d} I_d(x^d) - \inf_{x^d \in F_d^+} I_d(x^d) = \mathcal{O}(d^{\alpha/2}(\log d)^{1/2}d^{-1/2}) = o(d^{-1/4}) = o(d^{-\alpha/2})$. Note that $I_d(X^d)$ being tight is a very reasonable assumption, since if $I_d(X^d)$ is not tight, the target π^d becomes "flat" at almost every state x^d. ◁

We are now ready to present our first main result using the assumptions A1, A2, A3, and A4. We establish the following results on asymptotic ESJD and asymptotic acceptance rate.

Theorem 3.10. *(Asymptotic ESJD and acceptance rate) Suppose* π^d *satisfies A1, A2, A3, and A4, then as* $d \to \infty$, *we have*

$$\left| \mathrm{ESJD}(d) - 2\frac{d\ell^2}{d-1}\mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2} \right) \right] \right| \to 0, \tag{12}$$

$$\left| \mathbb{E}_{X^d \sim \pi^d}\mathbb{E}_{Y^d}\left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) - 2\mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2} \right) \right] \right| \to 0, \tag{13}$$

where the expectation over Y^d *is taken for* $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1}I)$ *for given* $X^d = x^d$.

Proof. See Appendix A. □

Since the assumptions required by Theorem 3.10 are very mild, the result of Theorem 3.10 holds for a large class of realistic MCMC models. As an example, we give a class of graphical models that all conditions A1, A2, A3, and A4 hold. Therefore, the asymptotic ESJD and acceptance rate by Theorem 3.10 hold for this class of graphical models. We will further discuss realistic MCMC models later in Section 4.1 and Section 4.2.

We give a simple criterion that the assumptions A1, A2, A3, and A4 hold. More discussions and examples are delayed to Section 4.

Corollary 3.11. *If a graphical model satisfies (i) either each node has at most $l_d = o(d^{1/4})$ links or the number of 3-cliques of the graph is $o(d^{3/2})$; (ii) $I_d(X^d)$ is tight; (iii) π^d has bounded density and $\log \pi^d$ has up to the third bounded partial derivatives, then the assumptions A1, A2, A3, and A4 hold. Therefore, the asymptotic ESJD and acceptance rate results by Theorem 3.10 hold.*

Proof. First, the assumption A1 holds when second partial derivatives of $\log \pi^d$ are bounded. Next, the assumption A2 automatically holds for graphical models. Furthermore, $l_d = o(d^{1/4})$ implies that the number of 3-cliques is $o(d^{3/2})$. Then one can easily verify that the assumption A3 holds using the fact that the third partial derivatives of $\log \pi^d$ are bounded. Finally, the assumption A4 holds since $I_d(X^d)$ is tight. □

Note that Theorem 3.10 suggests that under mild conditions on the target distribution, the expected acceptance rate

$$\mathbb{E}_{X^d \sim \pi^d} \mathbb{E}_{Y^d} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \to 2 \mathbb{E}_{X^d \sim \pi^d} \left[\Phi \left(-\frac{\ell \sqrt{I_d(X^d)}}{2} \right) \right]. \tag{14}$$

Therefore, we can define asymptotic acceptance rate as a function of ℓ as follows.

Definition 3.12. (Asymptotic Acceptance Rate) The asymptotic acceptance rate function is defined by

$$a(\ell) := 2 \mathbb{E}_{X^d \sim \pi^d} \left[\Phi \left(-\frac{\ell \sqrt{I_d(X^d)}}{2} \right) \right]. \tag{15}$$

The next theorem shows that if the target distribution satisfies A1, A2, A3 and A4, then if we maximize the asymptotic ESJD, the resulting asymptotic acceptance rate is no larger than 0.234.

Theorem 3.13. *Defining the optimal parameter for maximizing the asymptotic ESJD by $\hat{\ell}$, i.e.*

$$\hat{\ell} := \arg \max_{\ell} h(\ell), \quad h(\ell) := 2\ell^2 \mathbb{E}_{X^d \sim \pi^d} \left[\Phi \left(-\frac{\ell \sqrt{I_d(X^d)}}{2} \right) \right], \tag{16}$$

then we have $a(\hat{\ell}) \leq 0.234$ (to three decimal places).

Proof. We follow the arguments in [Taw17, Lemma 5.1.4]. First, it can be verified by taking the second derivatives of $h(\ell)$ with respect to ℓ that the maximum of $h(\ell)$ is achieved at ℓ

such that $\frac{\partial h(\ell)}{\partial \ell} = 0$. Therefore, the optimal $\hat{\ell}$ satisfies

$$2\mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)\right] = \mathbb{E}_{X^d \sim \pi^d}\left[\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\Phi'\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)\right]. \quad (17)$$

Therefore, the asymptotic acceptance rate

$$a(\hat{\ell}) = \mathbb{E}_{X^d \sim \pi^d}\left[\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\Phi'\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)\right] = \mathbb{E}_{X^d \sim \pi^d}\left[-\Phi^{-1}(V)\Phi'\left(\Phi^{-1}(V)\right)\right], \quad (18)$$

where $V := \Phi\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)$. By [She06], the function $-\Phi^{-1}(x)\Phi'\left(\Phi^{-1}(x)\right)$ is a concave function for any $x \in (0,1)$. Therefore, we have

$$a(\hat{\ell}) = \mathbb{E}_{X^d \sim \pi^d}\left[-\Phi^{-1}(V)\Phi'\left(\Phi^{-1}(V)\right)\right] \le -\Phi^{-1}[\mathbb{E}_{X^d \sim \pi^d}(V)]\Phi'\left[\Phi^{-1}(\mathbb{E}_{X^d \sim \pi^d}(V))\right]. \quad (19)$$

Defining $m := -\Phi^{-1}[\mathbb{E}_{X^d \sim \pi^d}(V)]$, we can then write $a(\hat{\ell}) = 2\Phi(-m) \le m\Phi'(-m)$. Finally, it suffices to show that $2\Phi(-m) \le m\Phi'(-m)$ implies $2\Phi(-m) \le 0.234$ (to three decimal places). Note that the function $x^2\Phi(-x)$ is maximized at $\hat{m}$ such that $2\Phi(-\hat{m}) = \hat{m}\Phi'(-\hat{m}) \approx 0.234$. By [Taw17, Lemma 5.1.4], the function $2\Phi(-x) - x\Phi'(-x)$ is positive for $x < \hat{m}$ and negative for $x > \hat{m}$. Therefore, $2\Phi(-m) \le m\Phi'(-m)$ implies that $m > \hat{m}$. Since $\Phi(-x)$ is monotonically decreasing with x, we have $a(\hat{\ell}) = 2\Phi(-m) \le 2\Phi(-\hat{m}) \approx 0.234$. $\qquad \square$

The next result is our main result for optimal scaling by maximizing ESJD. Defining the following WLLN condition for the target π^d:

$$I_d(X^d) - \bar{I}_d \to 0 \quad \text{in probability} \qquad (A5)$$

where $X^d \sim \pi^d$ and $\bar{I}_d := \mathbb{E}_{X^d \sim \pi^d}[I_d(X^d)]$, we show that if the target distribution π^d satisfies A1, A2, A3, A4, and the WLLN assumption in A5, then the acceptance rate 0.234 is asymptotically optimal.

Theorem 3.14. *(Optimal Scaling for Maximizing ESJD) Suppose the target distribution π^d satisfies A1, A2, A3, A4, and A5. Then the asymptotic optimal acceptance rate $a(\hat{\ell}) \approx 0.234$ (to three decimal places).*

Proof. By convexity of the function $\Phi(-x)$ when $x \ge 0$, we can immediately obtain a lower bound

$$\ell^2 \mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right] \ge \ell^2 \left[\Phi\left(-\frac{\ell\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]}{2}\right)\right]. \quad (20)$$

Under A5, this lower bound is asymptotically tight. Therefore, as $d \to \infty$, according to [RGG97], we have (to two decimal places)

$$\hat{\ell} \to \frac{2.38}{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]}, \quad h(\hat{\ell}) \to \frac{1.3}{\left(\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]\right)^2}. \quad (21)$$

The acceptance rate which maximizing the asymptotic ESJD is

$$a(\hat{\ell}) = 2\mathbb{E}_{X^d \sim \pi^d} \left[\Phi\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2} \right) \right] \to 2\Phi\left(-\frac{\hat{\ell}\mathbb{E}_{X^d \sim \pi^d}\sqrt{I_d(X^d)}}{2} \right) \tag{22}$$

$$\approx 2\Phi\left(-\frac{2.38}{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]} \frac{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]}{2} \right) = 2\Phi(-1.19) \approx 0.234. \tag{23}$$

$\square$

Remark 3.15. Comparing the results of Theorem 3.13 and Theorem 3.14, it is clear that the "roughness" of π^d, $I_d(X^d)$, is the key quantity which determines the optimal acceptance rate $a(\hat{\ell}) \leq 0.234$ when only the tightness of $I_d(X^d)$ can be verified, or $a(\hat{\ell}) \approx 0.234$ when the concentration of $I_d(X^d)$ as defined in A5 can be verified. We will later demonstrate how to verify A5 for some realistic MCMC models in Section 4.1 and Section 4.2. ◁

3.2. **Optimal Scaling via Diffusion Limits.** In this subsection, we consider sufficient conditions on π^d for establishing weak convergence of diffusion limits. As we discussed before, establishing such results gives stronger guarantee for optimal scaling than maximizing ESJD. However, it also requires stronger conditions on the target distribution. As we will see in the following, we need to strengthen assumptions A2, A3, A4, A5 and add one more assumption A6.

We first strengthen A2 to a new assumption A2+ as follows.

$$\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \int \left(\frac{\partial^2 \pi^d}{\partial x_i^2} \frac{1}{\pi^d} \right) \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d} \right) \left(\frac{\partial^2 \pi^d}{\partial x_k^2} \frac{1}{\pi^d} \right) \pi^d \mathrm{d}x^d = \mathcal{O}(d^{2-\delta}) \tag{A2+}$$

for some $\delta > 0$.

Remark 3.16. The new assumption A2+ is stronger than A2 but is still very mild. To see this, we consider graphical models as examples. For graphical models with d nodes each with $\mathcal{O}(l_d)$ links, there are at most $\mathcal{O}(dl_d^2)$ 3-cliques. Therefore, A2+ holds for any graphical model with $l_d = o(d^{1/2-\delta})$ and bounded second partial derivatives of $\log \pi^d$. Note that this is only for the worst case, as many realistic graphical models do not have 3-cliques. ◁

Next, we slightly strengthen A3 and A4 to A3+ and A4+.

$$\sup_{(i,j):j\notin\mathcal{H}_i} \sup_{x^d\in\mathbb{R}^d} \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i^2 \partial x_j} = o(1), \qquad \sup_{(i,j):j\in\mathcal{H}_i} \sup_{x^d\in\mathbb{R}^d} \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i^2 \partial x_j} = o(\sqrt{d/l_d}),$$

$$\sum_{i\neq j\neq k} \left(\sup_{x^d\in\mathbb{R}^d} \left| \frac{\partial^3 \log \pi^d(x^d)}{\partial x_i \partial x_j \partial x_k} \right| \right) = o(d^{3/2}). \tag{A3+}$$

Suppose exists $0 < \alpha < 1/2$ that

$$\sup_i \sup_{x^d \in F_d^{(i)}} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i^2} = o(d^\alpha), \quad \sup_i \sup_{x^d \in F_d^{(i)}} \frac{\partial \log \pi^d(x^d)}{\partial x_i} = \mathcal{O}(d^{\alpha/2}),$$

$$\sup_{x^d \in F_d^+} \pi^d(x^d) = o(d^{1/2-\alpha}), \quad \sup_{x^d \in F_d^+} 1/I_d(x^d) = \mathcal{O}(d^{\alpha/4}). \tag{A4+}$$

Furthermore, we strengthen the WLLN condition A5 to the following A5+.

$$\sup_{x^d \in F_d^+} \left| I_d(x^d) - \bar{I} \right| \to 0 \tag{A5+}$$

where $\bar{I} := \lim_{d \to \infty} \bar{I}_d$ exists.

Remark 3.17. A3+ is only slightly stronger than A3 on the rates. A4+ also includes a new condition on the rate of $\frac{\partial^2 \log \pi^d(x^d)}{\partial x_i^2}$ which is quite weak. A5+ requires any sequence $(x^1, x^2, \ldots, x^d, \ldots)$ where $x^i \in F_i^+$ converges to the same limit $\bar{I}$, so it is (slightly) stronger than WLLN condition in A5. It will become clear in the proof of Theorem 3.19 that A5+ is to ensure the speed measure of the diffusion process $h(\ell)$ does not depend on the state x^d. ◁

Finally, we define a new assumption A6 on the target distribution. Roughly speaking, the new assumption is to require the first coordinate of π^d is asymptotically independent with the rest.

$$\lim_{d \to \infty} \sup_{x^d \in F_d^+} \left| \frac{\mathrm{d}}{\mathrm{d}x_1} \left[\log \pi^d(x_1 \mid x_{-1}) - \log \tilde{\pi}(x_1) \right] \right| = 0, \tag{A6}$$

where $x_{-1} := (x_2, \ldots, x_d)$, $\tilde{\pi}$ is a one-dimensional density and $(\log \tilde{\pi})'$ is Lipschitz continous.

Remark 3.18. Note that A6 is a strong condition, which may not be satisfied for many realistic MCMC models. However, it might be necessary in order to get a one-dimensional diffusion limit for the first coordinate. In the proof of the optimal scaling via diffusion limits result in Theorem 3.19, the assumption A6 is to ensure the SDE for the first coordinate x_1 doesn't depend on the values of other coordinates. Furthermore, although we do not pursue in this paper, if in A6 we instead assume not just the first component but a finite collection of components are asymptotically independent from the rest, a version of weak convergence to multi-dimensional diffusion limits could be obtained following similar arguments as the proof of the one-dimensional diffusion limit case in Theorem 3.19. ◁

Now we are ready for the main result of optimal scaling via diffusion limits, which is given in Theorem 3.19. We show that, even though different coordinates of the Markov chain are *not* independent *nor* even individually Markovian, the sped-up first-coordinate process converges to a limiting diffusion limit under much more general conditions on the target distribution. Comparing with the assumptions in Theorem 3.14, the new sufficient conditions for diffusion limits include strengthening A2 to A2+, A3 and A4 to A3+ and A4+, A5 to A5+, and adding A6. We also require slightly stronger condition on the sequence of "sets of typical states" $\{F_d\}$.

Theorem 3.19. *(Optimal Scaling via Diffusion Limits) Suppose the sequence $\{F_d\}$ satisfies $\pi^d(F_d^c) = \mathcal{O}(d^{-1-\delta})$ for some $\delta > 0$, the target distribution π^d satisfies A1, A2+, A3+, A4+, A5+, and A6, then for $U^d(t) := X_1^d(\lfloor dt \rfloor)$, as $d \to \infty$, we have $U^d \Rightarrow U$, where $\Rightarrow$ denotes weak convergence in Skorokhod topology, and U satisfies the Langevin SDE*

$$\mathrm{d}U(t) = (h(\ell))^{1/2}\mathrm{d}B(t) + h(\ell)\frac{\tilde{\pi}'(U(t))}{2\tilde{\pi}(U(t))}\mathrm{d}t, \tag{24}$$

where $h(\ell) := 2\ell^2\Phi(-\ell\sqrt{\bar{\bar{I}}}/2)$ is the speed measure for the diffusion process.

Proof. See Appendix C.   □

Remark 3.20. Note that Theorem 3.19 allows dependent coordinates on the target distribution, which is much more general than the product i.i.d. condition. The only strong assumption is A6 which requires the first coordinate is asymptotically independent with other coordinates.   ◁

Next, we present another result with slightly stronger conditions, which allows the RWM algorithm to start at a fixed state. This stronger convergence result later allows us to establish a complexity bound for the RMW algorithm in Section 3.3 Let $X^d = \{X^d(t), t \geq 0\}$ for $d \in \mathbb{N}$ be the RWM processes defined earlier. Without loss of generality, suppose $\{X^d, d = 1, 2, \dots\}$ are defined in a common measurable metric space $(\mathbb{R}^\infty, \mathcal{F}, \rho)$ as independent processes.

Theorem 3.21. *(Optimal Scaling via Diffusion Limits with fixed starting state) Suppose X_1^d converges weakly in the Skorokhod topology as $d \to \infty$ to a càdlàg process X_1^∞. Moreover, assume these processes $\{X^d, d = 1, 2, \dots\}$ all have the same marginal stationary distribution π_1 for the first coordinate and that the first coordinate of X^∞ converges weakly to π_1. Suppose the sequence $\{F_d\}$ satisfies $\pi^d(F_d^c) = \mathcal{O}(d^{-2-\delta})$ for some $\delta > 0$, the target distribution π^d satisfies A1, A3+, A4+, A5+, and A6. We strengthen A2+ to the following condition*

$$\sum_{i,j,k,l,m\in\{2,\dots,d\}} \int \left(\frac{\partial^2\pi_{-1}}{\partial x_i^2} \cdot \frac{\partial^2\pi_{-1}}{\partial x_j^2} \cdot \frac{\partial^2\pi_{-1}}{\partial x_k^2} \cdot \frac{\partial^2\pi_{-1}}{\partial x_l^2} \cdot \frac{\partial^2\pi_{-1}}{\partial x_m^2}\right)\left(\frac{1}{\pi_{-1}}\right)^5 \pi^d\mathrm{d}x^d = \mathcal{O}(d^{3-6\delta}).$$

$$\text{(A2++)}$$

Then as $d \to \infty$, we have $_xU^d \Rightarrow {}_xU$, where $_xU^d(t) := (X_1^d(\lfloor dt \rfloor) \,|\, X_1^d(0) = x)$ is the first coordinate of the RWM algorithm sped up by a factor of d, conditional on starting at the state x, and $_xU$ is the limiting ergodic Langevin diffusion U in Eq. (24) also conditional on starting at x.

Proof. See Appendix D.   □

Remark 3.22. The new assumption A2++ is stronger than A2+ but is still not strong. To see this, for graphical models with d nodes, each with $\mathcal{O}(l_d)$ links, we have at most $\mathcal{O}(dl_d^2)$ 3-cliques. Under flat tail assumptions, at most $\mathcal{O}(d^2l_d^3)$ terms in the summation in A2++ is not zero. Therefore, A2++ holds for any graphical model with $l_d = o(d^{1/3-2\delta})$ and bounded second partial derivatives of $\log\pi^d$. Note that this is only for the worst case, as many realistic graphical models do not have 3-cliques.   ◁

3.3. **Complexity Bounds via Diffusion Limits.** In the following, by combing Theorem 3.21 and Proposition 2.3, we present a complexity bound for the RWM algorithm which holds for much more general target distributions comparing with [RR16]. More specifically, if the target distribution satisfies the conditions given in Theorem 3.21 which allows dependent coordinates of the target distribution, the RWM algorithm in d dimensions takes $\mathcal{O}(d)$ iterations to converge to stationarity.

Corollary 3.23. *(Complexity Bound for RWM Algorithms) Under the conditions of Theorem 3.21, for any $\epsilon > 0$, there exists $D < \infty$ and $T < \infty$, such that*

$$\mathbb{E}_{X_1^d(0) \sim \pi_1} \|\mathcal{L}_{X_1^d(0)}(X_1^d(\lfloor dt \rfloor)) - \pi_1\|_{KR} < \epsilon, \quad \forall t \geq T, d \geq D, \tag{25}$$

where π_1 denotes the marginal stationary distribution of the first coordinate.

Proof. The result directly comes from Proposition 2.3 and Theorem 3.21. $\qquad\square$

4. Examples and Applications

In this section, we further discuss examples and applications of the main results in Section 3. We first discuss in Section 4.1 on verifying the assumptions of Theorem 3.14 for realistic MCMC models. We have explained in Remarks 3.6 to 3.9 that A1, A2, A3, and A4 are typically very weak conditions and they hold for some classes of graphical models. However, as discussed in Remark 3.15, the assumption A5 may need to be verified case by case. Particularly, in order to satisfy A5, we may need to make additional assumptions on the observed data. Fortunately, we show by a simple Gaussian example in Example 4.1 that, in some cases, A5 can be easily verified without any further assumptions. Then, in Section 4.2, we extend the simple Gaussian example in Example 4.1 to a more realistic MCMC model in Example 4.5 and show it satisfies all the assumptions required by Theorem 3.14. Thus, the acceptance rate 0.234 is indeed asymptotically optimal for this realistic MCMC model.

4.1. **Discussions on Theorem 3.14.** The optimal scaling result for maximizing ESJD in Theorem 3.14 requires one to verify that the target distribution satisfies A1, A2, A3, A4, and A5. We discuss how to verify the conditions on the target distribution required by Theorem 3.14 in practice. We explain that A1, A2, A3 and A4 are quite mild and usually easy to be verified. Therefore, we usually only need to focus on the WLLN condition in A5, which might be difficult to check in practice. Throughout this subsection, we demonstrate verification of all the assumptions by a simple Gaussian example, which can be seen as a simplified version of typical Bayesian hierarchical models.

Example 4.1. (A Gaussian Example) Consider a simple Gaussian MCMC model

$$\begin{aligned} Y_{ij} \mid \theta_{ij} &\sim \mathcal{N}(\theta_{ij}, 1), \quad i, j \in \{1, \dots, n\} \\ \theta_{ij} \mid \mu_j &\sim \mathcal{N}(\mu_j, 1), \quad i \in \{1, \dots, n\} \\ \mu_j \mid \nu &\sim \mathcal{N}(\nu, 1) \\ \nu &\sim \text{ flat prior on } \mathbb{R}, \end{aligned} \tag{26}$$

where $\{Y_{ij}\}_{i,j=1}^n$ are the observed data, and $x^d = (\nu, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n)$ are parameters. Note that we have the number of parameters $d = n^2 + n + 1$ in this example. The target distribution (i.e. the posterior distribution) satisfies

$$\pi^d(x^d) = \mathbb{P}(x^d \mid \{Y_{ij}\}_{i,j=1}^n) \propto \prod_{j=1}^n \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu_j-\nu)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\theta_{ij}-\mu_j)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_{ij}-\theta_{ij})^2}{2}}. \quad (27)$$

Note that the hyperparameters ν is conditionally independent given $\{\theta_{ij}\}$. Therefore, ν is only directly dependent with n coordinates $\{\mu_j\}_{j=1}^n$. We can define the "neighborhoods" of ν using the collection of $\mu_j, j = 1, \ldots, n$. Similarly, μ_j is directly dependent with ν and $\{\theta_{ij}\}_{i=1}^n$ and θ_{ij} is directly dependent with μ_j. Therefore, if we choose the directly dependent coordinates as "neighborhoods", we have $l_d = n + 1 = \mathcal{O}(d^{1/2})$. ◁

4.1.1. *Verifying A1 to A4.* First of all, the two conditions for $(i, j) : j \neq \mathcal{H}_i$ in A1 and A3 hold trivially for graphical models. Furthermore, in Example 4.1, the parameter ν is conditional independent with all θ_{ij} and the corresponding conditional posterior distributions all have Gaussian tails, which implies A2 holds for any pair of coordinates (ν, θ_{ij}). Similarly, one can easily verify the assumption holds for other pairs of parameters.

Next, all the conditions on the third partial derivatives of $\log \pi^d$ hold, since there is no 3-cliques. Moreover, in Example 4.1, we have $l_d = \mathcal{O}(d^{1/2})$. The second partial derivative is $\mathcal{O}(1)$, and the density π^d is bounded, so the following conditions hold without the need of choosing $\{F_d\}$:

$$\sup_{(i,j):j\in\mathcal{H}_i} \sup_{x^d\in F_d^+} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(\sqrt{d/l_d}), \quad \sup_{x^d\in F_d^+} \pi^d(x^d) = o(d^{1/2-\alpha}). \quad (28)$$

Finally, the last two conditions are almost immediately true once A5 has been verified:

$$\sup_{i\in\{1,\ldots,d\}} \sup_{x^d\in F_d^+} \frac{\partial \log \pi^d(x^d)}{\partial x_i} = \mathcal{O}(d^\alpha), \quad \sup_{x^d\in F_d^+} 1/I(x^d) = \mathcal{O}(d^{\alpha/2}). \quad (29)$$

To see this, under A5, we have $\frac{1}{d}\sum_{i=1}^d \left(\frac{\partial}{\partial x_i} \log \pi^d(x^d)\right)^2 \to \bar{I}_d$. If $\bar{I}_d \to \bar{I}$ and $\bar{I} > 0$, then we can select constant $K_2 > 0$ small enough such that $\bar{I} > K_2 d^{-\alpha/2} > 0$ then $\bar{I}_d > K_2 d^{-\alpha/2}$ for all large enough d. Next, by choosing the typical set F_d such that for any $x^d \in F_d^+$, we have $\frac{\partial \log \pi^d(x^d)}{\partial x_i} \leq K_1 d^\alpha$, $I_d(x^d) \geq K_2 d^{-\alpha/2}$, where K_1 is a large enough constant. Then it suffices to check if $\{F_d\}$ is a valid sequence of typical sets such that $\pi^d(F_d) \to 1$. For Example 4.1, we have $X^d = (\nu, \{\nu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n)$. We will show later that A5 holds such that under $X^d \sim \pi^d$ we have $\frac{1}{d}\sum_{i=1}^d \left(\frac{\partial}{\partial x_i} \log \pi^d(X^d)\right)^2 \to 3$. For example, we can choose

$K_2 = 0.01$, $K_1 = 100$, and the typical set F_d such that, for any $X^d = x^d \in F_d^+$, we have

$$I_d(x^d) > 0.01n^{-\alpha}, \quad \frac{\partial \log \pi^d}{\partial \nu} = n(\bar{\mu} - \nu) \le 100n^{2\alpha}, \tag{30}$$

$$\frac{\partial \log \pi^d}{\partial \mu_j} = (n+1)\left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right) \le 100n^{2\alpha}, \tag{31}$$

$$\frac{\partial \log \pi^d}{\partial \theta_{ij}} = 2\left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right) \le 100n^{2\alpha}, \tag{32}$$

where $\alpha < 1/2$ can be arbitrarily close to $1/2$. Observing that, under $X^d \sim \pi^d$, we have the following conditional distributions.

$$\theta_{ij} \mid Y_{ij}, \mu_j \quad \sim^{\text{indep.}} \mathcal{N}\left(\frac{\mu_j + Y_{ij}}{2}, \frac{1}{2}\right), \quad i, j \in \{1, \ldots, n\},$$

$$\mu_j \mid \sum_i \theta_{ij}, \nu \sim^{\text{indep.}} \mathcal{N}\left(\frac{\sum_i \theta_{ij} + \nu}{n+1}, \frac{1}{n+1}\right), \quad i \in \{1, \ldots, n\}, \tag{33}$$

$$\nu \mid \bar{\mu} \sim \mathcal{N}\left(\bar{\mu}, \frac{1}{n}\right).$$

Then it can be easily verified that $\pi^d(F_d) \to 1$.

4.1.2. *Verifying A5.* One assumption of Theorem 3.14 that could be difficult to verify in practice is A5. It requires the sequence of random variables $\{I_d(X^d)\}$ converge to a sequence of constants in probability. We feel this assumption has to be checked case by case and it is hard to get general sufficient condition for it to hold. For realistic MCMC models, this may require assumptions on the observed data so that the posterior distribution has certain "concentration" properties as $d \to \infty$.

Fortunately, for Example 4.1, we can verify that A5 holds without any further assumption on the observed data $\{Y_{ij}\}$. Note that in Example 4.1, we have

$$\left(\frac{\partial \log \pi^d}{\partial \nu}\right)^2 = \left(\sum_j (\mu_j - \nu)\right)^2 = n^2 (\bar{\mu} - \nu)^2, \tag{34}$$

$$\left(\frac{\partial \log \pi^d}{\partial \mu_j}\right)^2 = \left(\sum_i (\theta_{ij} - \mu_j) - (\mu_j - \nu)\right)^2 = (n+1)^2 \left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right)^2, \tag{35}$$

$$\left(\frac{\partial \log \pi^d}{\partial \theta_{ij}}\right)^2 = ((Y_{ij} - \theta_{ij}) - (\theta_{ij} - \mu_j))^2 = 4\left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right)^2. \tag{36}$$

Hence, if suffices to show that, under $X^d = (\nu, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n) \sim \pi^d$, the following three terms converges to some constants in probability or in distribution:

$$\frac{1}{d}\left(\frac{\partial \log \pi^d}{\partial \nu}\right)^2 = \frac{n^2}{n^2 + n + 1}(\bar{\mu} - \nu)^2, \tag{37}$$

$$\frac{1}{d}\sum_j \left(\frac{\partial \log \pi^d}{\partial \mu_j}\right)^2 = \frac{(n+1)^2}{n^2 + n + 1}\sum_j \left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right)^2, \tag{38}$$

$$\frac{1}{d}\sum_{ij} \left(\frac{\partial \log \pi^d}{\partial \theta_{ij}}\right)^2 = \frac{4}{d}\sum_{ij} \left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right)^2. \tag{39}$$

We have observed that the target distribution π^d has conditional independence structure in Eq. (33), which immediately leads to

$$(\bar{\mu} - \nu)^2 \to^{\mathbb{P}} 0, \quad \sum_j \left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right)^2 \to^{\mathbb{P}} 1, \quad \frac{1}{d}\sum_{ij} \left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right)^2 \to^{\mathbb{P}} \frac{1}{2}. \tag{40}$$

Therefore, A5 is satisfied.

Overall, we have checked all the assumptions of Theorem 3.14 for our simple Gaussian example. Therefore, by Theorem 3.14, we have the following optimal scaling result for Example 4.1.

Proposition 4.2. *The optimal scaling for Example 4.1 by maximizing ESJD is to choose (to two decimal places) $\hat{\ell} \approx \frac{2.38}{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I(X^d)}]} \to \frac{2.38}{\sqrt{3}} \approx 1.37$ and the corresponding asymptotic acceptance rate is (to three decimal places) 0.234.*

4.2. **Optimal Scaling of a Realistic MCMC Model.** We first discuss sufficient conditions for two more classes of graphical models. In Proposition 4.3, we give sufficient conditions for the first equation of A1, A2, and the first equation of A3 to hold for one particular class of graphical models. In Proposition 4.4, we give sufficient conditions for A5 to hold for one specific class of graphical models.

First, we consider the class of graphical models represented by the factor graphs:

$$\pi^d(x^d) \propto \prod_{k=1}^{K_d} \psi_k(\{x_i : i \in C_k\}), \tag{41}$$

where C_k are cliques, ψ_k are potentials, K_d denotes the number of potentials.

Proposition 4.3. *For the class of graphical models represented by Eq. (41). Let m_d denotes the maximum number of cliques a coordinate can belong to. If all the potentials ψ_k have "flat tails" in the sense that for all k we have $\frac{\partial \psi_k}{\partial x_i} \to 0$ as $|x_i| \to \infty$ for all $i \in C_k$, and the cardinality of C_k satisfies $\sup_k |C_k| = o(d/m_d)$, then the first equation in A1, A2, and the first equation in A3 hold.*

Next, we consider Bayesian hierarchical modeling where K denotes the number of "layers" or "stages" of the model. We use $\theta^{(k)}, k = 1, \ldots, K$ to denote the parameter vector with

length n_k for the k-th layer, where $\theta^{(k)} := (\theta_1^{(k)}, \ldots, \theta_{n_k}^{(k)})$. We consider the special structure of the graphical model such that $\theta^{(k)}$ is only connected to $\theta^{(k-1)}$ and $\theta^{(k+1)}$. Using factor graphs, let $x^d = (\theta^{(1)}, \ldots, \theta^{(K)})$ we can represent the target distribution as

$$\pi^d(x^d) \propto \prod_{k=1}^{K} \psi_k(\theta^{(k-1)}, \theta^{(k)}), \tag{42}$$

where $d = \sum_{k=1}^{K} n_k$, $\{\psi_k\}$ are the potentials, and without loss of generality we assumed $\theta^{(0)}$ to be the observed data.

In the following, we show that A5 hold for the class of graphical models represented by Eq. (42) under certain conditions.

Proposition 4.4. *For the class of graphical models represented by Eq. (42), if $\theta^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{n_k}^{(k)})$ are independent conditional on $\theta^{(k-1)}$ and $\theta^{(k+1)}$ and this holds for all k. Moreover, if under $X^d = (\theta^{(1)}, \ldots, \theta^{(K)}) \sim \pi^d$ all the potentials ψ_k satisfy*

$$\sup_{i \in \{1, \ldots, n_k\}} \left| \frac{\partial \log \psi_k}{\partial \theta_i^{(k)}} \right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{d/n_k}\right), \qquad \sup_{j \in \{1, \ldots, n_{k-1}\}} \left| \frac{\partial \log \psi_k}{\partial \theta_j^{(k-1)}} \right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{d/n_{k-1}}\right) \tag{43}$$

then A5 holds.

Next, we extend the simple Gaussian example in Example 4.1 to a more realistic MCMC model which belongs to both classes of graphical models in Eqs. (41) and (42) and show that all the assumptions for the optimal scaling result in Theorem 3.14 hold.

Example 4.5. (A Realistic MCMC Model) Consider a realistic MCMC model

$$\begin{aligned}
Y_{ij} \mid \theta_{ij} \quad &\sim \mathcal{N}(\theta_{ij}, W), \quad i, j \in \{1, \ldots, n\} \\
\theta_{ij} \mid \mu_j &\sim \mathcal{N}(\mu_j, V), \quad i \in \{1, \ldots, n\} \\
\mu_j \mid \nu &\sim \mathcal{N}(\nu, A) \\
\nu &\sim \text{ flat prior on } \mathbb{R}, \\
A &\sim \mathbf{IG}(a, b),
\end{aligned} \tag{44}$$

where $x^d = (\nu, A, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n)$ are parameters, $\{Y_{ij}\}$ are the observed data, and a, b, W, V are known constants. ◁

We further assume that the observed data $\{Y_{ij}\}$ is not abnormal so that the posterior of the hyperparameter A concentrates to some unknown constant.

Assumption. *The posterior of the hyperparameter A in Example 4.5 concentrates to some unknown constant $A_0 > 0$ as $n \to \infty$.*

Note that this is a very reasonable assumption which implies the MCMC model is not seriously misspecified. We do not discuss sufficient conditions on the observed data $\{Y_{ij}\}_{i,j=1}^n$ for concentration of posterior distribution of A here since it is not the focus of this paper. Next, we show that, under this assumption, the realistic MCMC model satisfies all the conditions required for optimal scaling in Theorem 3.14. Therefore, the acceptance rate

0.234 is indeed asymptotically optimal for this MCMC model in the sense of maximizing ESJD.

Proposition 4.6. *Under the above assumption, the optimal asymptotic acceptance rate for the realistic MCMC model in Example 4.5 is (to three decimal places)* 0.234.

Proof. See Appendix E. □

Acknowledgment

References

[ARR11]   Y. F. Atchadé, G. O. Roberts, and J. S. Rosenthal. "Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo". *Statistics and Computing* 21.4 (2011), pp. 555–568.

[AT08]   C. Andrieu and J. Thoms. "A tutorial on adaptive MCMC". *Statistics and Computing* 18.4 (2008), pp. 343–373.

[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.

[BDM12]   M. Bédard, R. Douc, and E. Moulines. "Scaling analysis of multiple-try MCMC methods". *Stochastic Processes and their Applications* 122.3 (2012), pp. 758–786.

[BGJM11]   S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[BPRSS+13] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, and A. Stuart. "Optimal tuning of the hybrid Monte Carlo algorithm". *Bernoulli* 19.5A (2013), pp. 1501–1534.

[BPS04]   L. A. Breyer, M. Piccioni, and S. Scarlatti. "Optimal scaling of MALA for nonlinear regression". *The Annals of Applied Probability* 14.3 (2004), pp. 1479–1505.

[BR00]   L. A. Breyer and G. O. Roberts. "From Metropolis to diffusions: Gibbs states and optimal scaling". *Stochastic Processes and their Applications* 90.2 (2000), pp. 181–206.

[BR08]   M. Bédard and J. S. Rosenthal. "Optimal scaling of Metropolis algorithms: Heading toward general target distributions". *Canadian Journal of Statistics* 36.4 (2008), pp. 483–503.

[BREZ18]   N. Bou-Rabee, A. Eberle, and R. Zimmer. "Coupling and Convergence for Hamiltonian Monte Carlo" (2018). arXiv:1805.00452.

[BRS09]   A. Beskos, G. Roberts, and A. Stuart. "Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions". *The Annals of Applied Probability* 19.3 (2009), pp. 863–898.

[BRSV08]   A. Beskos, G. Roberts, A. Stuart, and J. Voss. "MCMC methods for diffusion bridges". *Stochastics and Dynamics* 8.03 (2008), pp. 319–350.

[Béd07]   M. Bédard. "Weak convergence of Metropolis algorithms for non-i.i.d. target distributions". *The Annals of Applied Probability* 17.4 (Aug. 2007), pp. 1222–1244.

[Béd08]   M. Bédard. "Optimal acceptance rates for Metropolis algorithms: moving beyond 0.234". *Stochastic Processes and their Applications* 118.12 (2008), pp. 2198–2222.

[Béd19]   M. Bédard. "Hierarchical Models and Tuning of Random Walk Metropolis Algorithms". *Journal of Probability and Statistics* 2019 (2019).

[CDPS18]   V. Chen, M. M. Dunlop, O. Papaspiliopoulos, and A. M. Stuart. "Dimension-Robust MCMC in Bayesian Inverse Problems" (Mar. 9, 2018).

[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.

[CRR05]   O. F. Christensen, G. O. Roberts, and J. S. Rosenthal. "Scaling limits for the transient phase of local Metropolis–Hastings algorithms". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 253–268.

[CRSW13]   S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. "MCMC methods for functions: modifying old algorithms to make them faster". *Statistical Science* (2013), pp. 424–446.

[Dal17]   A. S. Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

[DCWY18]   R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. "Log-concave sampling: Metropolis-Hastings algorithms are fast!" In: *Conference On Learning Theory*. 2018, pp. 793–797.

[DK19]   A. S. Dalalyan and A. Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". *Stochastic Processes and their Applications* (2019).

[DLCMR17] A. Durmus, S. Le Corff, E. Moulines, and G. O. Roberts. "Optimal scaling of the random walk Metropolis algorithm under L^p mean differentiability". *Journal of Applied Probability* 54.4 (2017), pp. 1233–1260.

[EK86]   S. N. Ethier and T. G. Kurtz. *Markov processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Characterization and convergence. John Wiley & Sons, Inc., New York, 1986, pp. x+534. ISBN: 0-471-08186-8.

[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.

[Has70]   W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika* 57.1 (1970), pp. 97–109.

[HSV11]   M. Hairer, A. Stuart, and J. Voss. "Signal processing problems on function space: Bayesian formulation, stochastic PDEs and effective MCMC methods". In: *The Oxford handbook of nonlinear filtering*. Oxford Univ. Press, Oxford, 2011, pp. 833–873.

[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.

[JLM14]   B. Jourdain, T. Lelièvre, and B. a. Miasojedow. "Optimal scaling for the transient phase of Metropolis Hastings algorithms: the longtime behavior". *Bernoulli* 20.4 (2014), pp. 1930–1978.

[JLM15]   B. Jourdain, T. Lelièvre, and B. Miasojedow. "Optimal scaling for the transient phase of the random walk Metropolis algorithm: The mean-field limit". *The Annals of Applied Probability* 25.4 (2015), pp. 2263–2300.

[Jor04]   M. I. Jordan. "Graphical Models". *Statistical Science* 19.1 (2004), pp. 140–155.

[KOS18]   J. Kuntz, M. Ottobre, and A. M. Stuart. "Non-stationary phase of the MALA algorithm". *Stochastics and Partial Differential Equations: Analysis and Computations* 6.3 (2018), pp. 446–499.

[KOS19]   J. Kuntz, M. Ottobre, and A. M. Stuart. "Diffusion limit for the random walk Metropolis algorithm out of stationarity". *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 55.3 (Aug. 2019), pp. 1599–1648.

[Liu08]   J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[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.

[MPS12]   J. C. Mattingly, N. S. Pillai, and A. M. Stuart. "Diffusion limits of the random walk Metropolis algorithm in high dimensions". *The Annals of Applied Probability* 22.3 (2012), pp. 881–930.

[MRRT+53]   N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of state calculations by fast computing machines". *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[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.

[MT96]   K. L. Mengersen and R. L. Tweedie. "Rates of convergence of the Hastings and Metropolis algorithms". *The Annals of Statistics* 24.1 (1996), pp. 101–121.

[NR06]   P. Neal and G. Roberts. "Optimal scaling for partially updating MCMC algorithms". *The Annals of Applied Probability* 16.2 (2006), pp. 475–515.

[NR08]   P. Neal and G. Roberts. "Optimal scaling for random walk Metropolis on spherically constrained target densities". *Methodology and Computing in Applied Probability* 10.2 (2008), pp. 277–297.

[NR11]   P. Neal and G. Roberts. "Optimal scaling of random walk Metropolis algorithms with non-Gaussian proposals". *Methodology and Computing in Applied Probability* 13.3 (2011), pp. 583–601.

[NRY12]   P. Neal, G. Roberts, and W. K. Yuen. "Optimal scaling of random walk Metropolis algorithms with discontinuous target densities". *The Annals of Applied Probability* 22.5 (2012), pp. 1880–1927.

[PG10]   C. Pasarica and A. Gelman. "Adaptively Scaling the Metropolis Algorithm using Expected Square Jumped Distance". *Statistica Sinica* 20.1 (2010), pp. 343–364.

[PST12]   N. S. Pillai, A. M. Stuart, and A. H. Thiéry. "Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions". *The Annals of Applied Probability* 22.6 (Dec. 2012), pp. 2320–2356.

[RC04]   C. P. Robert and G. Casella. "Monte Carlo Statistical Methods". *Springer, New York* (2004).

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

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

[Ros11]   J. S. Rosenthal. "Optimal proposal distributions and adaptive MCMC". *Handbook of Markov Chain Monte Carlo* 4 (2011).

[Ros95]   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.

[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.

[RR01]   G. O. Roberts and J. S. Rosenthal. "Optimal scaling for various Metropolis–Hastings algorithms". *Statistical science* 16.4 (2001), pp. 351–367.

[RR14]   G. O. Roberts and J. S. Rosenthal. "Minimising MCMC variance via diffusion limits, with an application to simulated tempering". *The Annals of Applied Probability* 24.1 (2014), pp. 131–149.

[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:1508.00947* (2015).

[RS94]   G. O. Roberts and A. F. Smith. "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms". *Stochastic processes and their applications* 49.2 (1994), pp. 207–216.

[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.

[SFR10]   C. Sherlock, P. Fearnhead, and G. O. Roberts. "The random walk Metropolis: linking theory and practice through a case study". *Statistical Science* (2010), pp. 172–190.

[She06]   C. Sherlock. "Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis". PhD thesis. Lancaster University, 2006.

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

[SR09]   C. Sherlock and G. Roberts. "Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets". *Bernoulli* (2009), pp. 774–798.

[STRR15]   C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. "On the efficiency of pseudo-marginal random walk Metropolis algorithms". *The Annals of Statistics* 43.1 (2015), pp. 238–275.

[Stu10]   A. M. Stuart. "Inverse problems: a Bayesian perspective". *Acta Numerica* 19 (2010), pp. 451–559.

[Taw17]   N. Tawn. "Towards Optimality of the Parallel Tempering Algorithm". PhD thesis. University of Warwick, 2017.

[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.

[YR17]   J. Yang and J. S. Rosenthal. "Complexity Results for MCMC derived from Quantitative Bounds". *arXiv:1708.00829* (2017).

[ZBK17]   G. Zanella, M. Bédard, and W. S. Kendall. "A Dirichlet form approach to MCMC optimal scaling". *Stochastic Processes and their Applications* 127.12 (2017), pp. 4053 –4082.

A. Proof of Theorem 3.10

Throughout the proof, for simplicity, we assume the coordinates are linear ordered. The "neighborhoods" of a coordinate is defined by $\mathcal{H}_i := \{j : |i - j| < l_d\}$. Therefore $\sup_{(i,j):j\in\mathcal{H}_i}$ can be simplified to $\sup_{|i-j|<l_d}$ and $\sup_{(i,j):j\notin\mathcal{H}_i}$ can be simplified to $\sup_{|i-j|\geq l_d}$. Note that the use of linear ordering is only for simplifying notations. It is straightforward to extend the proof to the cases of general ordering.

For Theorem 3.10, we only prove

$$\left| \text{ESJD}(d) - 2\frac{d\ell^2}{d-1}\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2} \right) \right] \right| \to 0, \tag{45}$$

since the proof of

$$\left| \mathbb{E}_{X^d\sim\pi^d}\mathbb{E}_{Y^d}\left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) - 2\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2} \right) \right] \right| \to 0 \tag{46}$$

follows similarly.

First, we write ESJD as $\text{ESJD}(d) =: \sum_{i=1}^d \text{ESJD}_i(d)$, where

$$\text{ESJD}_i(d) := \mathbb{E}_{X^d\sim\pi^d}\mathbb{E}_{Y^d}\left[(Y_i - X_i)^2 \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right]. \tag{47}$$

Then it suffices to show that

$$\sup_{i\in\{1,\dots,d\}} \left| \text{ESJD}_i(d) - \frac{2\ell^2}{d-1}\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2} \right) \right] \right| = o(d^{-1}). \tag{48}$$

Writing $\text{ESJD}_i(d) = \mathbb{E}_{X^d\sim\pi^d}\mathbb{E}_{Y_i}\left[(Y_i - X_i)^2 \mathbb{E}_{Y_{-i}}\left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right]$, it suffices to show that uniformly over $i \in \{1, \dots, d\}$

$$\mathbb{E}_{X^d\sim\pi^d} \left| \mathbb{E}_{Y_i}\left[(Y_i - X_i)^2 \mathbb{E}_{Y_{-i}}\left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right] - \frac{2\ell^2}{d-1}\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2} \right) \right| \tag{49}$$

$$= \mathbb{E}_{X^d\sim\pi^d} \left| \mathbb{E}_{Y_i}\left\{ (Y_i - X_i)^2 \left[\mathbb{E}_{Y_{-i}}\left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) - 2\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2} \right) \right] \right\} \right| \tag{50}$$

$$= o(d^{-1}). \tag{51}$$

It then suffices to show

$$\sup_{x^d \in F_d} \left| \mathbb{E}_{Y_i} \left\{ (Y_i - x_i)^2 \mathbb{1}_{y^d(i) \in F_d^{(i)}} \left[\mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right) - 2\Phi \left(-\frac{\ell \sqrt{I_d(x^d)}}{2} \right) \right] \right\} \right| \quad (52)$$

$$\leq \mathbb{E}_{Y_i} \left\{ (Y_i - x_i)^2 \sup_{y^d(i) \in F_d^{(i)}, x^d \in F_d} \left| \mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right) - 2\Phi \left(-\frac{\ell \sqrt{I_d(x^d)}}{2} \right) \right| \right\} = o(d^{-1}), \quad (53)$$

where $y^d(i) := (x_1, \ldots, x_{i-1}, Y_i, x_{i+1}, \ldots, x_d)$. Defining $M_{x^d}^{(i)}(Y_i) := \mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right)$, since

$$\log \frac{\pi^d(Y^d)}{\pi^d(x^d)} = \log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(Y_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid x_i)} \quad (54)$$

$$= \left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid x_i)} \right) + \log \frac{\pi_{-i}(Y_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid Y_i)}, \quad (55)$$

we can write

$$M_{x^d}^{(i)}(Y_i) = \mathbb{E}_{Y_{-i}} \left[1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right] = \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp \left(\log \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right) \right] \quad (56)$$

$$= \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp \left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid x_i)} + \log \frac{\pi_{-i}(Y_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid Y_i)} \right) \right]. \quad (57)$$

Note that the expectation is taken over Y_{-i} and only the last term, $\log \frac{\pi_{-i}(Y_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid Y_i)}$, involves Y_{-i}.

In the following, we then first focus on approximating $\log \frac{\pi_{-i}(Y_{-i} \mid x_i)}{\pi_{-i}(x_{-i} \mid x_i)}$ for given $x^d \in F_d^+$. Since $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1} I)$, we first approximate $\log \frac{\pi_{-i}(Y_{-i} \mid x_i)}{\pi_{-i}(x_{-i} \mid x_i)}$ by the first two terms of its Taylor expansion.

Define

$$m_1^{(i)}(Y_{-i}, x^d) := (\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) + \frac{1}{2} (Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}] (Y_{-i} - x_{-i}), \quad (58)$$

where

$$(\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) := \sum_{j \in \{1, \ldots, d\}, j \neq i} \frac{\partial \log \pi_{-i}(x_{-i} \mid x_i)}{\partial x_j} (Y_j - x_j) \quad (59)$$

and $[\nabla^2 \log \pi_{-i}]$ denotes the $(d-1) \times (d-1)$ matrix with elements

$$\left\{ \frac{\partial^2 \log \pi_{-i}(x_{-i} \mid x_i)}{\partial x_j \partial x_k} \right\}_{j,k \in \{1, \ldots, d\}, j \neq i, k \neq i}.$$

Then, we have the following result.

Lemma A.1. *Uniformly over $i \in \{1, \ldots, d\}$, we have*

$$\sup_{x^d \in F_d^+} \mathbb{E}_{Y_{-i}} \left[\left| m_1^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \mid x_i)}{\pi_{-i}(x_{-i} \mid x_i)} \right| \right] \to 0. \tag{60}$$

Proof. See Appendix B.1. □

Next, we approximate the second order term of the Taylor approximation $\frac{1}{2}(Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i})$ by a non-random term $\frac{1}{2} \frac{\ell^2}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2}$.

Lemma A.2. *Uniformly over $i \in \{1, \ldots, d\}$, we have*

$$\sup_{x^d \in F_d^+} \mathbb{E}_{Y_{-i}} \left[\left| (Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i}) - \frac{\ell^2}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2} \right| \right] \to 0. \tag{61}$$

Proof. See Appendix B.2. □

Defining

$$m_2^{(i)}(Y_{-i}, x^d) := (\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) + \frac{1}{2} \frac{\ell^2}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2}, \tag{62}$$

we have

$$m_2^{(i)}(Y_{-i}, x^d) \sim \mathcal{N} \left(\ell^2 S_d^{(i)}/2, \ell^2 R_d^{(i)} \right), \tag{63}$$

where

$$R_d^{(i)} := \frac{1}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \mid x_i)}{\partial x_j} \right)^2, \quad S_d^{(i)} := \frac{1}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}(x_{-i} \mid x_i)}{\partial x_j^2}. \tag{64}$$

Next, we show we can approximate $S_d^{(i)}$ by $-R_d^{(i)}$.

Lemma A.3. *There exists a sequence of subsets of states $\{F_d'\}$, such that $\pi^d(F_d') \to 1$ and*

$$\sup_{i \in \{1, \ldots, d\}} \sup_{x^d \in F_d'} \left| R_d^{(i)} + S_d^{(i)} \right| \to 0. \tag{65}$$

Proof. See Appendix B.3. □

Now defining

$$m_3^{(i)}(Y_{-i}, x^d) := (\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) + \frac{1}{2} \frac{\ell^2}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \mid x_i)}{\partial x_j} \right)^2, \tag{66}$$

we have

$$m_3^{(i)}(Y_{-i}, x^d) \sim \mathcal{N} \left(-\ell^2 R_d^{(i)}/2, \ell^2 R_d^{(i)} \right). \tag{67}$$

By triangle inequality, we can write

$$\left| m_3^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \mid x_i)}{\pi_{-i}(x_{-i} \mid x_i)} \right| \leq \left| m_1^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \mid x_i)}{\pi_{-i}(x_{-i} \mid x_i)} \right| \tag{68}$$

$$+ \left| m_2^{(i)}(Y_{-i}, x^d) - m_1^{(i)}(Y_{-i}, x^d) \right| \tag{69}$$

$$+ \left| m_3^{(i)}(Y_{-i}, x^d) - m_2^{(i)}(Y_{-i}, x^d) \right|. \tag{70}$$

Therefore, using Lemmas A.1 to A.3, we get

$$\sup_{i \in \{1,\ldots,d\}} \sup_{x^d \in F_d^+ \cap F_d'} \mathbb{E}_{Y_{-i}} \left[\left| m_3^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \mid x_i)}{\pi_{-i}(x_{-i} \mid x_i)} \right| \right] \to 0. \tag{71}$$

Next, we abuse the notation a little bit by defining

$$R_d^{(i)}(y) := \frac{1}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \mid x_i = y)}{\partial x_j} \right)^2. \tag{72}$$

Then by the definition of $m_3^{(i)}$, we replace x^d by $y^d(i) = (x_1, \ldots, x_{i-1}, Y_i, x_{i+1}, \ldots, x_d)$, which yields

$$m_3^{(i)}(Y_{-i}, y^d(i)) = (\nabla \log \pi_{-i}(x_{-i} \mid Y_i))^T (Y_{-i} - x_{-i}) \tag{73}$$

$$+ \frac{1}{2} \frac{\ell^2}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \mid Y_i)}{\partial x_j} \right)^2. \tag{74}$$

Then, we have

$$m_3^{(i)}(Y_{-i}, y^d(i)) \sim \mathcal{N} \left(-\ell^2 R_d^{(i)}(Y_i)/2, \ell^2 R_d^{(i)}(Y_i) \right). \tag{75}$$

Recall that $M_{x^d}^{(i)}(Y_i) = \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp \left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid x_i)} + \log \frac{\pi_{-i}(Y_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid Y_i)} \right) \right]$, defining

$$\hat{M}_{x^d}^{(i)}(Y_i) = \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp \left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \mid Y_i)}{\pi_{-i}(x_{-i} \mid x_i)} + m_3^{(i)}(Y_{-i}, y^d(i)) \right) \right], \tag{76}$$

we next apply the following two lemmas from [RGG97].

Lemma A.4. ([RGG97, Proposition 2.2]) The function $g(x) = 1 \wedge e^x$ is Lipschitz such that

$$|g(x) - g(y)| \leq |x - y|, \quad \forall x, y. \tag{77}$$

Lemma A.5. ([RGG97, Proposition 2.4]) If $z \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbb{E}(1 \wedge e^z) = \Phi(\mu/\sigma) + \exp(\mu + \sigma^2/2)\Phi(-\sigma - \mu/\sigma). \tag{78}$$

By Lemma A.4 and Eq. (71), we have that uniformly over $i \in \{1, \ldots, d\}$

$$\sup_{y^d(i) \in F_d^+ \cap F_d'} \left| M_{x^d}^{(i)}(Y_i) - \hat{M}_{x^d}^{(i)}(Y_i) \right| \to 0. \tag{79}$$

Applying Lemma A.5 to $\hat{M}_{x^d}^{(i)}(Y_i)$ yields

$$\hat{M}_{x^d}^{(i)}(Y_i) = \Phi\left(R_d^{(i)}(Y_i)^{-1/2}\left(\ell^{-1}\log\frac{\pi^d(y^d(i))}{\pi^d(x^d)} - \ell R_d^{(i)}(Y_i)/2\right)\right) \tag{80}$$

$$+ \exp\left(\log\frac{\pi^d(y^d(i))}{\pi^d(x^d)}\right)\Phi\left(-\ell R_d^{(i)}(Y_i)^{1/2}/2 - \log\frac{\pi^d(y^d(i))}{\pi^d(x^d)}R_d^{(i)}(Y_i)^{-1/2}\ell^{-1}\right). \tag{81}$$

Note that it is easy to check that $\hat{M}_{x^d}^{(i)}(x_i) = 2\Phi\left(-\frac{\ell\sqrt{R_d^{(i)}}}{2}\right)$. We then show $\hat{M}_{x^d}^{(i)}(x_i)$

converges to $2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2}\right)$.

Lemma A.6.

$$\sup_{i\in\{1,\dots,d\}}\sup_{x^d\in F_d^+}\left|2\Phi\left(-\frac{\ell\sqrt{R_d^{(i)}}}{2}\right) - 2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2}\right)\right| \to 0. \tag{82}$$

Proof. See Appendix B.4. □

Finally, using Taylor expansion together with $\mathbb{E}_{Y_i}(Y_i - x_i)^2 = \ell^2/(d-1)$ and $\mathbb{E}_{Y_i}|Y_i - x_i|^3 = \mathcal{O}(d^{-3/2})$, we have

$$\mathbb{E}_{Y_i}\left\{(Y_i - x_i)^2\sup_{y^d(i)\in F_d^+}\left|\hat{M}_{x^d}^{(i)}(Y_i) - 2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2}\right)\right|\right\} \tag{83}$$

$$\leq \frac{\ell^2}{d-1}\sup_{x^d\in F_d^+}\left|2\Phi\left(-\frac{\ell\sqrt{R_d^{(i)}}}{2}\right) - 2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2}\right)\right| \tag{84}$$

$$+ \mathcal{O}(d^{-3/2})\sup_{y^d(i)\in F_d^+}\left|\frac{\mathrm{d}\hat{M}_{x^d}^{(i)}(y_i)}{\mathrm{d}y_i}(Y_i)\right|. \tag{85}$$

For the last term, we have the following lemma.

Lemma A.7.

$$\sup_{i\in\{1,\dots,d\}}\sup_{y^d(i)\in F_d^+}\left|\frac{\mathrm{d}\hat{M}_{x^d}^{(i)}(y_i)}{\mathrm{d}y_i}(Y_i)\right| = o\left(d^{1/2}\right). \tag{86}$$

Proof. See Appendix B.5. □

The proof of Theorem 3.10 is completed by applying Lemma A.6 and Lemma A.7.

B. Proof of Lemmas in Appendix A

B.1. **Proof of Lemma A.1.** For $x^d \in F_d^+$, by Taylor expansion and mean value theorem, we have

$$|\log \pi_{-i}(Y_{-i} \,|\, x_i) - \log \pi_{-i}(x_{-i} \,|\, x_i) - m_1(Y_{-i}, x^d)| \tag{87}$$

$$\leq \sup_{\tilde{x}^d \in \mathbb{R}^d} \left| \frac{1}{6} \sum_{j,k,l \neq i} \frac{\partial^3 \log \pi^d(\tilde{x}^d)}{\partial x_j \partial x_k \partial x_l} (Y_j - x_j)(Y_k - x_k)(Y_l - x_l) \right|. \tag{88}$$

In the above summation, the summation over the cases of $j = k = l$ equals to

$$\sup_{\tilde{x}^d \in \mathbb{R}^d} \left| \frac{\partial^3 \log \pi^d(\tilde{x}^d)}{\partial x_j^3} \right| \mathcal{O}(d\mathbb{E}|Y_j - x_j|^3) = o(d^{1/2})\mathcal{O}\left(d(\sqrt{\ell^2/(d-1)})^3 \right) = o(1). \tag{89}$$

For the cases of $j = k \neq l$, we have

$$\sum_{j=k \neq l} \frac{\partial^3 \log \pi_{-i}(\tilde{x}^d)}{\partial x_j^2 \partial x_l} (Y_j - x_j)^2 (Y_l - x_l) = \sum_j (Y_j - x_j)^2 \sum_{l \neq k} \frac{\partial^3 \log \pi_{-i}(\tilde{x}^d)}{\partial x_j^2 \partial x_l} (Y_l - x_l). \tag{90}$$

By Assumption A3, we have $\mathbb{E}\left| \sum_{j \neq l} \frac{\partial^3 \log \pi_{-i}(\tilde{x}^d)}{\partial x_j^2 \partial x_l} (Y_l - x_l) \right| = \mathcal{O}(l_d/d)o(d/l_d) = o(1)$ since $\frac{\partial^3 \log \pi^d(\tilde{x}^d)}{\partial x_j^2 \partial x_l}$ goes to zero when $|k - i| > l_d$. Then, by $\mathbb{E}|Y_j - x_j|^2 = \mathcal{O}(1/d)$, the summation over all cases of $j = k \neq l$ equals to $d\mathcal{O}_{\mathbb{P}}(1/d)o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$.

Finally, for $j \neq k \neq l$, it suffices to show

$$\sup_{\tilde{x}^d \in \mathbb{R}^d} \left| \sum_{j \neq k \neq l \neq i} \frac{\partial^3 \log \pi_{-i}(\tilde{x}^d)}{\partial x_j \partial x_k \partial x_l} (Y_j - x_j)(Y_k - x_k)(Y_l - x_l) \right| \tag{91}$$

$$\leq \sum_{i \neq j \neq k \neq l} \left(\sup_{\tilde{x}^d \in \mathbb{R}^d} \left| \frac{\partial^3 \log \pi_{-i}(\tilde{x}^d)}{\partial x_j \partial x_k \partial x_l} \right| \right) |(Y_j - x_j)(Y_k - x_k)(Y_l - x_l)| = o_{\mathbb{P}}(1). \tag{92}$$

Note that $\{|(Y_j - x_j)(Y_k - x_k)(Y_l - x_l)|\}_{j \neq k \neq l}$ are independent random variables which don't depend on the values of x_j, x_k, x_l, and

$$|(Y_j - x_j)(Y_k - x_k)(Y_l - x_l)| = \mathcal{O}_{\mathbb{P}}\left((\sqrt{\ell^2/(d-1)})^3 \right) = \mathcal{O}_{\mathbb{P}}(d^{-3/2}). \tag{93}$$

Therefore, the summation for cases $j \neq k \neq l$ is $o_{\mathbb{P}}(1)$ under Assumption A3. We have proven the result for fixed i. Finally, it is easy to check the proof holds uniformly over $i \in \{1, \ldots, d\}$.

B.2. **Proof of Lemma A.2.**

Lemma B.1. *(Quadratic Form of Gaussian Random Vector) If $z^d \sim \mathcal{N}_d(\mu, \Sigma)$, then*

$$\mathbb{E}(z^T A z) = \operatorname{tr}(A\Sigma) + \mu^T A \mu, \quad \operatorname{var}(z^T A z) = 2\operatorname{tr}(A\Sigma A\Sigma) + 4\mu^T A\Sigma A\mu. \tag{94}$$

Note that $Y_{-i} \sim \mathcal{N}_{d-1}(x_{-i}, \frac{\ell^2}{d-1}I)$ and $(Y_{-i} - x_{-i})^T[\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i})$ is a quadratic form of Gaussian random vector. By Lemma B.1,

$$\mathbb{E}\left[(Y_{-i} - x_{-i})^T[\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i})\right] = \frac{\ell^2}{d-1}\sum_{j\neq i}\frac{\partial^2 \log \pi_{-i}}{\partial x_j^2}. \tag{95}$$

Therefore, it suffices to show the variance of the quadratic form goes to zero. Using the assumptions, the variance satisfies

$$\frac{2\ell^4}{(d-1)^2}\operatorname{tr}\left([\nabla^2 \log \pi_{-i}][\nabla^2 \log \pi_{-i}]\right) \tag{96}$$

$$= \frac{2\ell^4}{(d-1)^2}\sum_{j\neq i}\sum_{k\neq i}\left(\frac{\partial^2 \log \pi_{-i}}{\partial x_j \partial x_k}\right)^2 \tag{97}$$

$$\leq \frac{2\ell^4}{(d-1)^2}\sum_{l=0}^{d-1}\sum_{\{j,k:|j-k|=l\}}\left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{98}$$

$$= \frac{2\ell^4}{(d-1)^2}\sum_{l\leq l_d}\sum_{\{j,k:|j-k|=l\}}\left(\frac{\partial^2 \log \pi^d}{\partial x_i \partial x_j}\right)^2 \tag{99}$$

$$+ \frac{2\ell^4}{(d-1)^2}\sum_{l>l_d}\sum_{\{j,k:|j-k|=l\}}\left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{100}$$

$$\leq \frac{2\ell^4}{(d-1)^2}(d-1)l_d \sup_{|j-k|\leq l_d}\sup_{x^d\in F_d^+}\left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{101}$$

$$+ \frac{2\ell^4}{(d-1)^2}(d-1)^2 \sup_{|j-k|>l_d}\sup_{x^d\in F_d^+}\left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{102}$$

$$= \mathcal{O}(l_d/d)o(d/l_d) + o(1) = o(1), \tag{103}$$

where we have used $\sup_{x^d\in F_d^+}\sup_{|j-k|\leq l_d}\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k} = o(\sqrt{d/l_d})$ from Assumption A1.

B.3. **Proof of Lemma A.3.** Note that

$$R_d^{(i)} + S_d^{(i)} = \frac{1}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}}{\partial x_j} \right)^2 + \frac{1}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2} \tag{104}$$

$$= \frac{1}{d-1} \sum_{j \neq i} \left\{ \left(\frac{\partial \log \pi^d}{\partial x_j} \right)^2 + \frac{\partial^2 \log \pi^d}{\partial x_j^2} \right\} \tag{105}$$

$$= \frac{1}{d-1} \sum_{j \neq i} \left\{ \frac{1}{(\pi^d)^2} \left(\frac{\partial \pi^d}{\partial x_j} \right)^2 + \frac{\partial}{\partial x_j} \left(\frac{\partial \log \pi^d}{\partial x_j} \right) \right\} \tag{106}$$

$$= \frac{1}{d-1} \sum_{j \neq i} \left\{ \frac{1}{(\pi^d)^2} \left(\frac{\partial \pi^d}{\partial x_j} \right)^2 + \frac{\partial}{\partial x_j} \left(\frac{1}{\pi^d} \frac{\partial \pi^d}{\partial x_j} \right) \right\} \tag{107}$$

$$= \frac{1}{d-1} \sum_{j \neq i} \left\{ \frac{1}{(\pi^d)^2} \left(\frac{\partial \pi^d}{\partial x_j} \right)^2 + \frac{\pi^d \frac{\partial^2 \pi^d}{\partial x_j^2} - \left(\frac{\partial \pi^d}{\partial x_j} \right)^2}{(\pi^d)^2} \right\} \tag{108}$$

$$= \frac{1}{(d-1)} \sum_{j \neq i} \frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d}. \tag{109}$$

Next, we show $\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2 \right]$ converges to 0. To prove this, consider writing $\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2 \right]$ as sum of $(d-1)^2$ terms

$$\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2 \right] = \frac{1}{(d-1)^2} \int \sup_i \sum_{j \neq i} \sum_{k \neq i} \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d} \right) \left(\frac{\partial^2 \pi^d}{\partial x_k^2} \frac{1}{\pi^d} \right) \pi^d \mathrm{d}x^d \tag{110}$$

$$\leq \frac{1}{(d-1)^2} \sum_{j=1}^{d} \sum_{k=1}^{d} \int \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d} \right) \left(\frac{\partial^2 \pi^d}{\partial x_k^2} \frac{1}{\pi^d} \right) \pi^d \mathrm{d}x^d - \frac{2}{(d-1)^2} \int \inf_i \sum_{j \neq i} \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d} \right) \pi^d \mathrm{d}x^d \tag{111}$$

$$= \frac{1}{(d-1)^2} \sum_{j=1}^{d} \sum_{k=1}^{d} \int \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d} \right) \left(\frac{\partial^2 \pi^d}{\partial x_k^2} \frac{1}{\pi^d} \right) \pi^d \mathrm{d}x^d + o(1), \tag{112}$$

where the last equality follows from

$$\frac{2}{(d-1)^2} \int \inf_i \sum_{j \neq i} \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d} \right) \pi^d \mathrm{d}x^d \geq \frac{2}{(d-1)^2} \int \inf_i \sum_{j \neq i} \left(\frac{\partial^2 \log \pi_{-i}}{\partial x_j^2} \right) \pi^d \mathrm{d}x^d \tag{113}$$

$$= \frac{2}{(d-1)^2} o(d\sqrt{d/l_d}) = o(1). \tag{114}$$

When $|j - k| \geq l_d$, by Assumption A2, we have

$$\int \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \frac{1}{\pi^d} \right) \left(\frac{\partial^2 \pi^d}{\partial x_k^2} \frac{1}{\pi^d} \right) \pi^d \mathrm{d}x^d \tag{115}$$

$$= \int \left(\frac{\partial^2 \pi^d}{\partial x_j^2} \right) \left(\frac{\partial^2 \pi^d}{\partial x_k^2} \right) \frac{1}{\pi^d} \mathrm{d}x^d \tag{116}$$

$$= \int \left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_j^2} \right) \left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_k^2} \right) \frac{1}{\pi_{j,k|-j-k}} \pi_{-j-k} \mathrm{d}x_{-j-k} \mathrm{d}x_j \mathrm{d}x_k \tag{117}$$

$$\leq \int \left[\sup_{x^d \in F_d} \int \left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_j^2} \right) \left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_k^2} \right) \frac{1}{\pi_{j,k|-j-k}} \mathrm{d}x_j \mathrm{d}x_k \right] \pi_{-j-k} \mathrm{d}x_{-j-k} \tag{118}$$

$$\rightarrow 0. \tag{119}$$

This implies $\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2 \right] = \frac{\mathcal{O}(d\,l_d) + (d - l_d)^2 o(1)}{(d-1)^2} + o(1) \rightarrow 0$. Therefore, uniformly over i, $R_d^{(i)} + S_d^{(i)} \rightarrow 0$ in probability, then there exists a sequence $\{F_d'\}$ such that $\mathbb{P}(R_d^{(i)} + S_d^{(i)} \in F_d', \forall i) \rightarrow 1$ and the following holds

$$\sup_i \sup_{x^d \in F_d'} \left| R_d^{(i)} + S_d^{(i)} \right| \rightarrow 0. \tag{120}$$

B.4. **Proof of Lemma A.6.** Note that Assumption A4 implies

$$\sup_{i \in \{1,\dots,d\}} \sup_{x^d \in F_d^+} \frac{\partial}{\partial x_i} \log \pi^d(x^d) = o\left(d^{1/2} \right). \tag{121}$$

Then, by the definitions of $R_d^{(i)}$ and $I_d(x^d)$, we have

$$R_d^{(i)} - I_d(x^d) = \frac{1}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \mid x_i)}{\partial x_j} \right)^2 - \frac{1}{d} \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} \log \pi^d(x^d) \right)^2 \tag{122}$$

$$= \frac{1}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi^d(x^d)}{\partial x_j} \right)^2 - \frac{1}{d} \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} \log \pi^d(x^d) \right)^2 \tag{123}$$

$$= \frac{1}{d} R_d^{(i)} - \frac{1}{d} \left(\frac{\partial}{\partial x_i} \log \pi^d(x^d) \right)^2 \rightarrow 0. \tag{124}$$

B.5. **Proof of Lemma A.7.** Recall that we have shown

$$\hat{M}_{x^d}^{(i)}(Y_i) = \Phi \left(R_d^{(i)}(Y_i)^{-1/2} \left(\ell^{-1} \log \frac{\pi^d(y^d(i))}{\pi^d(x^d)} - \ell R_d^{(i)}(Y_i)/2 \right) \right) \tag{125}$$

$$+ \exp \left(\log \frac{\pi^d(y^d(i))}{\pi^d(x^d)} \right) \Phi \left(-\ell R_d^{(i)}(Y_i)^{1/2}/2 - \log \frac{\pi^d(y^d(i))}{\pi^d(x^d)} R_d^{(i)}(Y_i)^{-1/2} \ell^{-1} \right). \tag{126}$$

For notational simplicity, we omit the index i and write $R_d^{(i)}$ by R_d. To simplify the derivation, we note that $\hat{M}_{x^d}^{(i)}(y)$ has the following form

$$M(y) = \Phi\left(f(y)g(y) - \frac{1}{2}f^{-1}(y)\right) + \exp(g(y))\Phi\left(-\frac{1}{2}f^{-1}(y) - f(y)g(y)\right), \quad (127)$$

where $f^{-1}(y) := \ell R_d^{1/2}(y)$ and $g(y) = \log \pi^d(y^d(i)) - \log \pi^d(x^d)$. Taking the derivative with respect to y, we get

$$\frac{dM(y)}{dy} = \Phi'(fg - f^{-1}/2)\frac{d}{dy}(fg - f^{-1}/2) \quad (128)$$

$$+ \exp(g)\Phi'(-f^{-1}/2 - fg)\frac{d}{dy}(-fg - f^{-1}/2) \quad (129)$$

$$+ \exp(g)\left(\frac{d}{dy}g\right)\Phi(-fg - f^{-1}/2) \quad (130)$$

$$\leq \|\Phi'\|_\infty\left|\frac{df}{dy}g + \frac{dg}{dy}f - \frac{1}{2}\frac{df^{-1}}{dy}\right| \quad (131)$$

$$+ \exp(g)\|\Phi'\|_\infty\left|\frac{df}{dy}g + \frac{dg}{dy}f + \frac{1}{2}\frac{df^{-1}}{dy}\right| \quad (132)$$

$$+ \exp(g)\left|\frac{dg}{dy}\right|\|\Phi\|_\infty \quad (133)$$

Note that both Φ and Φ' are bounded functions. It then suffices to show

$$\exp(g)\left|\frac{dg}{dy}\right| = o(d^{1/2}), \quad \exp(g)\left|\frac{df}{dy}g\right| = o(d^{1/2}), \quad (134)$$

$$\exp(g)\left|\frac{dg}{dy}f\right| = o(d^{1/2}), \quad \exp(g)\left|\frac{df^{-1}}{dy}\right| = o(d^{1/2}). \quad (135)$$

Observing that $\frac{df^{-1}}{dy} = \frac{1}{2}\ell R_d'/R_d^{1/2}$ and $\frac{df}{dy} = -\frac{1}{2\ell}\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}$, if we can show

$$\sup_{i\in\{1,\dots,d\}}\frac{dR_d^{(i)}(y)}{dy}\frac{1}{[R_d^{(i)}(y)]^{1/2}} = o(1), \quad (136)$$

then we can get $\frac{\mathrm{d}f^{-1}}{\mathrm{d}y} = o(1)$ and $\frac{\mathrm{d}f}{\mathrm{d}y} = o(1/R_d)$. Using $R_d^{(i)} \to I_d(x^d)$ from Appendix B.4, it suffices to show

$$\left(\sup_{x^d \in F_d^+} \pi^d(x^d) \right) \left(\sup_i \sup_{x^d \in F_d^{(i)}} \frac{\partial \log \pi^d(x^d)}{\partial x_i} \right) = o(d^{1/2}), \tag{137}$$

$$\left(\sup_{x^d \in F_d^+} \pi^d(x^d) \right) \left(\sup_{x^d \in F_d^+} \left| \log(\pi^d(x^d))/I_d(x^d) \right| \right) = o(d^{1/2}), \tag{138}$$

$$\left(\sup_{x^d \in F_d^+} \pi^d(x^d) \right) \left(\sup_i \sup_{x^d \in F_d^{(i)}} \left| \frac{\partial \log \pi^d(x^d)}{\partial x_i} / \sqrt{I_d(x^d)} \right| \right) = o(d^{1/2}). \tag{139}$$

One can easily verify that the above equations hold under Assumption A4.

Finally, we complete the proof by showing Eq. (136). Recall that

$$R_d^{(i)}(y) = \frac{1}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \mid x_i = y)}{\partial x_j} \right)^2. \tag{140}$$

For notational simplicity, we write

$$R_d^{(i)}(y) = \frac{1}{d-1} \sum_{j \neq i} f_j^2(y), \tag{141}$$

where $f_j(y) := \frac{\partial \log \pi_{-i}(x_{-i} \mid x_i = y)}{\partial x_j}$. Then, by Cauchy–Schwartz inequality

$$\frac{\partial R_d^{(i)}(y)}{\mathrm{d}y} = \frac{2}{d-1} \sum_{j \neq i} f_j(y) f_j'(y) \leq \frac{2}{d-1} \sqrt{\sum_{j \neq i} f_j^2(y) \sum_{j \neq i} |f_j'(y)|^2}. \tag{142}$$

Note that by A1, if $|i - j| > l_d$ then $f_j'(y) \leq \sup_{x^d \in F_d} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} \to 0$. Hence, we have

$$\sup_{i \in \{1,\ldots,d\}} \frac{\mathrm{d}R_d^{(i)}(y)}{\mathrm{d}y} \frac{1}{[R_d^{(i)}(y)]^{1/2}} \leq \sup_i \frac{\frac{2}{d-1} \sqrt{\sum_{j \neq i} f_j^2(y) \sum_{j \neq i} |f_j'(y)|^2}}{\sqrt{\frac{1}{d-1} \sum_{j \neq i} f_j^2(y)}} \tag{143}$$

$$= 2 \sup_i \sqrt{\frac{1}{d-1} \sum_{j \neq i} |f_j'(y)|^2} \leq 2 \sqrt{\frac{1}{d-1} \sum_{j=1}^d |f_j'(y)|^2} = o\left(\sqrt{\frac{l_d}{d}} (\sqrt{d/l_d})^2 \right) = o(1). \tag{144}$$

C. Proof of Theorem 3.19

Similar to Appendix A, we assume the coordinates are linear ordered for simplicity. The proof follows the framework of [RGG97] using the generator approach [EK86].

Define the (discrete time) generator of x^d by

$$(G_d f)(x^d) := d\mathbb{E}_{Y^d} \left\{ [f(Y^d) - f(x^d)] \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right) \right\}, \tag{145}$$

for any function f for which this definition makes sense. In the Skorokhod topology, it doesn't cause any problem to treat G_d as a continuous time generator. We shall restrict attention to test functions such that $f(x^d) = f(x_1)$. We show uniform convergence of G_d to G, the generator of the limiting (one-dimensional) Langevin diffusion, for a suitable large class of real-valued functions f, where, for some fixed function $h(\ell)$,

$$(Gf)(x_1) := h(\ell) \left\{ \frac{1}{2} f''(x_1) + \frac{1}{2} [(\log \tilde{\pi})'(x_1)] \, f'(x_1) \right\}, \tag{146}$$

in which $\tilde{\pi}$ is a one-dimensional density of the first coordinate of π^d. Since we have assumed in A6 that $(\log \tilde{\pi})'$ is Lipschitz, by [EK86, Thm 2.1 in Ch.8], a core for the generator has domain C_c^∞, which is the class of continuous functions with compact support such that all orders of derivatives exist. This enable us to restrict attentions to functions $f_c \in C_c^\infty$ such that $f_c(x^d) = f_c(x_1)$.

Note that using Assumption A2+, and the assumption $\pi^d(F_d^c) = \mathcal{O}(d^{-1-\delta})$, following the arguments in the proof of Lemma A.3 we can get a stronger version of Lemma A.3 for $F_d' := \{x^d : \sup_i |R_d^{(i)} + S_d^{(i)}| \le d^{-\delta}\}$. Then using a union bound yields

$$\mathbb{P}(X^d(\lfloor ds \rfloor) \notin F_d \cap F_d', \exists 0 \le s \le t) \to 0. \tag{147}$$

Therefore, for any fixed t, if $d \to \infty$ then the probability of all $X^d(\lfloor ds \rfloor), 0 \le s \le t$ are in $F_d \cap F_d'$ goes to 1. Since $F_d \cap F_d' \subseteq F_d^+ \cap F_d' \subseteq F_d^+$, it suffices to consider $x^d \in F_d^+$.

Note that $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1} I)$, we can write

$$(G_d f_c)(x^d) = d \mathbb{E}_{Y_1} \left\{ [f_c(Y_1) - f_c(x_1)] \mathbb{E}_{Y_{-1}} \left[1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right] \right\}, \tag{148}$$

where $\mathbb{E}_{Y_{-1}}[\cdot]$ is short for $\mathbb{E}_{Y_2,\ldots,Y_d \mid Y_1}[\cdot]$ and π^d denotes the target distribution in d-dimension. The goal is then to prove $(G_d f_c)$ converges to $(G f_c)$.

Recall the definition Eq. (56), we omit the index to write $M_{x^d}^{(1)}$ as M_{x^d}, which is defined by

$$M_{x^d}(Y_1) = \mathbb{E}_{Y_{-1}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right). \tag{149}$$

Then we have previously shown in Eq. (79) that $M_{x^d}(Y_1)$ can be approximated by

$$\hat{M}_{x^d}(Y_1) = \Phi \left(R_d(Y_1)^{-1/2} \left(\ell^{-1} \log \frac{\pi^d(Y_1, x_{-1})}{\pi^d(x^d)} - \ell R_d(Y_1)/2 \right) \right) \tag{150}$$

$$+ \exp \left(\log \frac{\pi^d(Y_1, x_{-1})}{\pi^d(x^d)} \right) \Phi \left(-\ell R_d(Y_1)^{1/2}/2 - \log \frac{\pi^d(Y_1, x_{-1})}{\pi^d(x^d)} R_d(Y_1)^{-1/2} \ell^{-1} \right) \tag{151}$$

For $x^d \in F_d^+$, some properties of $\hat{M}_{x^d}$ is given as follows.

Lemma C.1. *For $\hat{M}_{x^d}$, we have*

$$\hat{M}_{x^d}(x_1) = 2\Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right), \tag{152}$$

$$\hat{M}'_{x^d}(x_1) = \Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\frac{d[\log \pi_1(x) + \log \pi_{-1}(x_{-1}\,|x)]}{dx}(x_1) + o(1), \tag{153}$$

$$\hat{M}'_{x^d}(x_1) = o(d^{1/2}), \quad \sup_{x^d \in F_d^+} \hat{M}''_{x^d} = o(d^{1/2}). \tag{154}$$

Proof. See Appendix C.1. $\qquad\square$

Since $f_c(Y_1) - f_c(x_1)$ is bounded, it suffices to show

$$\mathbb{E}_{Y_1}\left\{d[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\} \to (Gf_c)(x_1). \tag{155}$$

Now using mean value theorem and Taylor expansion of $\mathbb{E}_{Y_1}\left\{[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\}$ at $(Y_1 - x_1)$ yields

$$[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1) \tag{156}$$

$$= \left[f'_c(x_1)(Y_1 - x_1) + \frac{1}{2}f''_c(x_1)(Y_1 - x_1)^2 + K(Y_1 - x_1)^3\right] \tag{157}$$

$$\cdot \left[\hat{M}_{x^d}(x_1) + \hat{M}'_{x^d}(x_1)(Y_1 - x_1) + \frac{1}{2}\hat{M}''_{x^d}(x')(Y_1 - x_1)^2\right] \tag{158}$$

$$= f'_c(x_1)\hat{M}_{x^d}(x_1)(Y_1 - x_1) + \left[\frac{1}{2}f''_c(x_1)\hat{M}_{x^d}(x_1) + f'_c(x_1)\hat{M}'_{x^d}(x_1)\right](Y_1 - x_1)^2 \tag{159}$$

$$+ \left[K\hat{M}_{x^d}(x_1) + \frac{1}{2}f''_c(x_1)\hat{M}'_{x^d}(x_1) + \frac{1}{2}\hat{M}''_{x^d}(x')f'_c(x_1)\right](Y_1 - x_1)^3 \tag{160}$$

$$+ \left[\frac{1}{4}\hat{M}''_{x^d}(x')f''_c(x_1) + K\hat{M}'_{x^d}(x_1)\right](Y_1 - x_1)^4 + \frac{1}{2}\hat{M}''_{x^d}(x')K(Y_1 - x_1)^5, \tag{161}$$

where K is a constant since f_c has bounded third derivative. Note that both $f'_c(x_1)$ and $f''_c(x_1)$ are bounded as well. Therefore, taking expectation over Y_1 and using $\hat{M}'_{x^d}(x_1) = o(d^{1/2}), \sup_{x^d} \hat{M}''_{x^d} = o(d^{1/2})$ in Lemma C.1, we have

$$\mathbb{E}_{Y_1}\left\{[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\} = \left[\frac{1}{2}f''_c(x_1)\hat{M}_{x^d}(x_1) + f'_c(x_1)\hat{M}'_{x^d}(x_1)\right]\frac{\ell^2}{d-1} + o(d^{-1}). \tag{162}$$

Finally, by Assumption A6, we have

$$f_c'(x_1)\hat{M}'_{x^d}(x_1) + \frac{1}{2}f_c''(x_1)\hat{M}_{x^d}(x_1) \tag{163}$$

$$= 2\Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}[\log\pi_1(x) + \log\pi_{-1}(x_{-1}\,|\,x)]}{\mathrm{d}x}(x_1)\right) \tag{164}$$

$$= 2\Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}\log\pi_{1|-1}(x\,|\,x_{-1})}{\mathrm{d}x}(x_1)\right) \tag{165}$$

$$\to 2\Phi\left(-\frac{\ell I(x^d)^{1/2}}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}\log\tilde{\pi}(x)}{\mathrm{d}x}(x_1)\right) \tag{166}$$

$$\to 2\Phi\left(-\frac{\ell\bar{I}^{1/2}}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}\log\tilde{\pi}(x)}{\mathrm{d}x}(x_1)\right), \tag{167}$$

which implies that $\mathbb{E}_{Y_1}\left\{d[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\} \to (Gf_c)(x_1)$ where $h(\ell) := 2\ell^2\Phi(-\ell\sqrt{\bar{I}}/2)$.

C.1. **Proof of Lemma C.1.** The proof is quite tedious. In order to simplify the notations, we first introduce the following lemma.

Lemma C.2. *For the function $M(y)$ defined by*

$$M(y) = \Phi\left(f(y)g(y) - \frac{1}{2}f^{-1}(y)\right) + e^{g(y)}\Phi\left(-\frac{1}{2}f^{-1}(y) - f(y)g(y)\right), \tag{168}$$

we have

$$\frac{\mathrm{d}M(y)}{\mathrm{d}y} = \Phi'(fg - f^{-1}/2)\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2) \tag{169}$$

$$+ e^g\Phi'(-f^{-1}/2 - fg)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \tag{170}$$

$$+ e^g\left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)\Phi(-fg - f^{-1}/2). \tag{171}$$

$$\frac{\mathrm{d}^2 M(y)}{\mathrm{d}y^2} = \Phi''(fg - f^{-1}/2) \left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2) \right]^2 + \Phi'(fg - f^{-1}/2) \frac{\mathrm{d}^2}{\mathrm{d}y^2}(fg - f^{-1}/2) \quad (172)$$

$$+ e^g \left(\frac{\mathrm{d}}{\mathrm{d}y}g \right) \Phi'(-f^{-1}/2 - fg) \frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \quad (173)$$

$$+ e^g \left\{ \Phi''(-fg - f^{-1}/2) \left[\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \right]^2 + \Phi'(-fg - f^{-1}/2) \frac{\mathrm{d}^2}{\mathrm{d}y^2}(-fg - f^{-1}/2) \right\} \quad (174)$$

$$+ e^g \left(\frac{\mathrm{d}}{\mathrm{d}y}g \right) \Phi'(-fg - f^{-1}/2) \frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \quad (175)$$

$$+ \Phi(-fg - f^{-1}/2) \left[e^g \left(\frac{\mathrm{d}^2}{\mathrm{d}y^2}g \right) + e^g \left(\frac{\mathrm{d}}{\mathrm{d}y}g \right)^2 \right]. \quad (176)$$

Furthermore, if $g(x_1) = 0$, *then we have*

$$\frac{\mathrm{d}M(y)}{\mathrm{d}y}(x_1) = \left(\Phi'(-f^{-1}/2) \frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2) \right. \quad (177)$$

$$+ \Phi'(-f^{-1}/2) \frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \quad (178)$$

$$\left. + \left(\frac{\mathrm{d}}{\mathrm{d}y}g \right) \Phi(-f^{-1}/2) \right)(x_1) \quad (179)$$

$$= \left(\Phi'(-f^{-1}/2) \frac{\mathrm{d}}{\mathrm{d}y}(-f^{-1}) + \left(\frac{\mathrm{d}}{\mathrm{d}y}g \right) \Phi(-f^{-1}/2) \right)(x_1) \quad (180)$$

$$= -\Phi' \left(-\frac{f^{-1}(x_1)}{2} \right) \frac{\mathrm{d}f^{-1}(y)}{\mathrm{d}y}(x_1) + \frac{\mathrm{d}g(y)}{\mathrm{d}y}(x_1)\Phi \left(-\frac{f^{-1}(x_1)}{2} \right). \quad (181)$$

Remark C.3. Let $g(y) = \log \frac{\pi^d(y, x_{-1})}{\pi^d(x^d)}$ and $f^{-1}(y) = \ell R_d^{1/2}(y)$ then $\hat{M}_{x^d}(y) = M(y)$. ◁

Now substituting $g(y) = \log \frac{\pi^d(y, x_{-1})}{\pi^d(x^d)}$ and $f^{-1}(y) = \ell R_d^{1/2}(y)$ to Lemma C.2, we have

$$\hat{M}_{x^d}(x_1) = 2\Phi \left(-\frac{\ell R_d^{1/2}(x_1)}{2} \right), \quad (182)$$

and

$$\hat{M}'_{x^d}(x_1) = \frac{\mathrm{d}\hat{M}_{x^d}(y)}{\mathrm{d}y}(x_1) \quad (183)$$

$$= \Phi \left(-\frac{\ell R_d^{1/2}(x_1)}{2} \right) \frac{\mathrm{d}[\log \pi_1(x) + \log \pi_{-1}(x_{-1}|x)]}{\mathrm{d}x}(x_1) \quad (184)$$

$$- \Phi' \left(-\frac{\ell R_d^{1/2}(x_1)}{2} \right) \frac{\ell}{2R_d^{1/2}(x_1)} R'_d(x_1). \quad (185)$$

Since Φ' is bounded and by Eq. (136), $R_d'(x_1)/R_d^{1/2}(x_1) \to 0$, therefore

$$\Phi'\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\frac{\ell}{2R_d^{1/2}(x_1)}R_d'(x_1) = o(1). \tag{186}$$

Also, $\hat{M}'_{x^d}(x_1) = o(d^{1/2})$ since $\frac{\partial \log \pi^d}{\partial x_i} = \mathcal{O}(d^{\alpha/2}) = o(d^{1/2})$.

Now we prove $\sup_{x^d} \hat{M}''_{x^d} = o(d^{1/2})$. For simplicity, we keep the notations of f and g (recall that $g(y) = \log \frac{\pi^d(y, x_{-1})}{\pi^d(x^d)}$ and $f^{-1}(y) = \ell R_d^{1/2}(y)$) and use the results in Appendix B.5. Since Φ, Φ', Φ'' are bounded, it suffices to bound all the following terms to be $o(d^{1/2})$:

$$\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2, \quad \frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2), \quad \exp(g)\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2), \tag{187}$$

$$\exp(g)\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2, \quad \exp(g)\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2), \quad \exp(g)\left(\frac{\mathrm{d}^2g}{\mathrm{d}y^2}\right), \quad \exp(g)\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)^2. \tag{188}$$

Next, we show that most of them can be verified using Assumption A4+, and the results in Appendix B.5:

$$\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2 = \mathcal{O}\left[\left(\sup_{x^d \in F_d^+} \log \pi^d(x^d)\mathcal{O}(d^{\alpha/4}) + \sup_{x^d \in F_d^+} \frac{\partial \log \pi^d}{\partial x_1}\right)^2\right] \tag{189}$$

$$= \mathcal{O}\left[(d^{\alpha/4}\log d + d^{\alpha/2})^2\right] = o(d^{1/2}), \tag{190}$$

$$\left|e^g\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2)\right| = \mathcal{O}\left[\sup_{x^d \in F_d^+} \pi^d(x^d)\sup_{x^d \in F_d^+} \frac{\partial \log \pi^d}{\partial x_1}\left(d^{\alpha/4}\log d + d^{\alpha/2}\right)\right] \tag{191}$$

$$= o(d^{1/2-\alpha}d^{\alpha/2}(d^{\alpha/4}\log d + d^{\alpha/2})) = o(d^{1/2}), \tag{192}$$

$$\left|\exp(g)\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2\right| = o(d^{1/2-\alpha}d^\alpha) = o(d^{1/2}), \tag{193}$$

$$\left|\exp(g)\left(\frac{\mathrm{d}^2g}{\mathrm{d}y^2}\right)\right| = \mathcal{O}\left[\sup_{x^d \in F_d^+} \pi^d(x^d)\sup_{x^d \in F_d^+} \frac{\partial^2 \log \pi^d}{\partial x_1^2}\right] = o(d^{1/2-\alpha})\mathcal{O}(d^\alpha) = o(d^{1/2}), \tag{194}$$

$$\left|\exp(g)\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)^2\right| = \mathcal{O}\left[\sup_{x^d \in F_d^+} \pi^d(x^d)\sup_{x^d \in F_d^+} \left(\frac{\partial \log \pi^d}{\partial x_1^2}\right)^2\right] = o(d^{1/2-\alpha})\mathcal{O}(d^{\alpha/2})^2 = o(d^{1/2}). \tag{195}$$

The only terms left are $\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2)$ and $\exp(g)\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2)$. Therefore, it suffices to show

$$\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2) = \mathcal{O}(d^\alpha). \tag{196}$$

Note that

$$\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2) = \frac{\mathrm{d}}{\mathrm{d}y}(f'g + g'f - \frac{1}{2}\mathrm{d}f^{-1}) \tag{197}$$

$$= \frac{\mathrm{d}}{\mathrm{d}y}\left[\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}g + \frac{1}{R_d^{1/2}}g' - \frac{1}{2}\frac{R_d'}{R_d^{1/2}}\right] \tag{198}$$

$$= \frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}g' + \left(\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}\right)'g + \frac{1}{R_d^{1/2}}g'' + \left(\frac{1}{R_d^{1/2}}\right)'g' - \frac{1}{2}\left(\frac{R_d'}{R_d^{1/2}}\right)'. \tag{199}$$

Note that we have shown $R_d' = o(R_d^{1/2})$ in Appendix B.5. Similarly, we also can show using Assumption A3+ that

$$R_d'' = \frac{1}{d-1}(\sum_{j\neq1} f_j f_j')' = \frac{1}{d-1}\sum_{j\neq1}(f_j')^2 + \frac{1}{d-1}\sum_{j\neq1} f_j f_j'' \tag{200}$$

$$\leq \frac{1}{d-1}\sum_{j\neq1}(f_j')^2 + \sqrt{\frac{1}{d-1}\sum_{j\neq1} f_j^2}\sqrt{\frac{1}{d-1}\sum_{j\neq1}(f_j'')^2} \tag{201}$$

$$= \mathcal{O}(l_d/d)o((\sqrt{d/l_d})^2) + o(R_d^{1/2}\sqrt{l_d/d(\sqrt{d/l_d})^2}) = o(R_d^{1/2}), \tag{202}$$

where $f_j(x) := \frac{\partial \log \pi_{-1}(x_{-1}\,|\,x_1=x)}{\partial x_j}$. Therefore $R_d'' = o(R_d^{1/2})$ as well. Finally, we can complete the proof by verifying Eq. (196) using Assumption A4+ as follows.

$$\left|\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}g'\right| = \mathcal{O}\left(\frac{1}{R_d}\right)o(1)\mathcal{O}\left(\sup_{x^d\in F_d^+}\frac{\partial \log \pi^d}{\partial x_1}\right) = \mathcal{O}(d^{\alpha/4})o(d^{\alpha/2}) = o(d^\alpha), \tag{203}$$

$$\left|\left(\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}\right)'g\right| = \mathcal{O}\left[\frac{R_d''R_d^{3/2} + 3/2(R_d')^2R_d^{1/2}}{R_d^3}g\right] = \mathcal{O}\left[\frac{1}{R_d^{3/2}}(R_d''g)\right] \tag{204}$$

$$= \mathcal{O}(d^{\alpha/4})o(1)\mathcal{O}(d^{\alpha/2}) = o(d^\alpha), \tag{205}$$

$$\left|\frac{1}{R_d^{1/2}}g''\right| = \mathcal{O}\left(\sup_{x^d\in F_d^+}\frac{\partial^2 \log \pi^d}{\partial x_1^2}\right) = o(d^\alpha), \tag{206}$$

$$\left|\left(\frac{1}{R_d^{1/2}}\right)'g'\right| = \mathcal{O}\left(\frac{1}{2}\frac{1}{R_d^{3/2}}R_d'g'\right) = o(1/R_d)\mathcal{O}\left(\sup_{x^d\in F_d^+}\frac{\partial \log \pi^d}{\partial x_1}\right) = \mathcal{O}(d^{\alpha/4})o(d^{\alpha/2}) = o(d^\alpha), \tag{207}$$

$$\left|\left(\frac{R_d'}{R_d^{1/2}}\right)'\right| = \left|\frac{R_d''R_d^{1/2} - \frac{1}{2}(R_d')^2\frac{1}{R_d^{1/2}}}{R_d}\right| = \mathcal{O}\left(R_d''/R_d^{1/2}\right) = o(1) = o(d^\alpha). \tag{208}$$

D. Proof of Theorem 3.21

We follow the same approach as in the proof of [RR16, Proposition 3]. The idea is to follow the proof of Theorem 3.19 except in the proof of Eq. (79), we need a stronger version of Lemma A.3 to determine the sequence of "typical sets" $\{F'_d\}$.

Given fixed time t, considering the sequence of "typical sets" $\{F'_d\}$ defined by

$$F'_d := \{x^d : |R_d + S_d| \leq d^{-\delta}\}, \tag{209}$$

where $\delta > 0$ and we used R_d and S_d to denote $R_d^{(1)}$ and $S_d^{(1)}$ for simplicity. We need to guarantee that when d is large enough, we always have $X^d(\lfloor ds \rfloor) \in F_d \cap F'_d, \forall 0 \leq s \leq t$ and this happens for almost all starting state $X_1^d(0) = x$. That is, defining

$$p(d, x) := \mathbb{P}(X(\lfloor ds \rfloor) \notin F_d \cap F'_d, \exists 0 \leq s \leq t \mid X_1^d(0) = x), \tag{210}$$

letting π_1 denote the marginal stationary distribution for the first coordinate, we want to show that for any given $\epsilon > 0$, as $d \to \infty$

$$\mathbb{P}_{x \sim \pi_1}[p(d, x) \geq \epsilon, \text{infinite often}] = 0. \tag{211}$$

We prove it using Borel–Cantelli Lemma. Note that the application of Borel–Cantelli lemma is valid since we have assumed all of the processes are jointly defined on the same probability space as independent processes. First, note that

$$\mathbb{E}_{x \sim \pi_1}[p(d, x)] = dt\mathbb{P}_{\pi^d}((F_d \cap F'_d)^c) = dt\mathbb{P}_{\pi^d}(F_d^c \cup (F'_d)^c) \leq dt\mathbb{P}_{\pi^d}(F_d^c) + dt\mathbb{P}_{\pi^d}((F'_d)^c). \tag{212}$$

For any given $\epsilon > 0$, we have

$$\sum_{d=2}^{\infty} \mathbb{P}(p(x, d) \geq \epsilon) \leq \sum_{d=2}^{\infty} \frac{\mathbb{E}_{x \sim \pi_1}[p(d, x)]}{\epsilon} \tag{213}$$

$$\leq \frac{dt}{\epsilon} \sum_{d=2}^{\infty} \mathbb{P}_{\pi^d}(|R_d + S_d| > d^{-\delta}) + \frac{dt}{\epsilon} \sum_{d=2}^{\infty} \mathbb{P}(X^d \notin F_d). \tag{214}$$

By $\pi^d(F_d^c) = \mathcal{O}(d^{-2-\delta})$, we have $dt \sum_{d=2}^{\infty} \mathbb{P}(X^d \notin F_d) < \infty$. Now in order to use Borel–Cantelli Lemma, the condition we need is that for some number of moments m such that

$$\mathbb{P}_{\pi^d}(|R_d + S_d| > d^{-\delta}) \leq \frac{\mathbb{E}|R_d + S_d|^m}{d^{-m\delta}} = d^{m\delta}\mathbb{E}|R_d + S_d|^m = \mathcal{O}(d^{-2-\delta}), \tag{215}$$

which leads to $\sum_{d=2}^{\infty} \mathbb{P}(p(x, d) \geq \epsilon) < \infty$. In order to obtain non-trivial conditions, we let $m = 5$ and Assumption A2++ implies $\mathbb{E}|R_d + S_d|^5 = \mathcal{O}(d^{-2-6\delta})$. We can then use this sequence of typical sets $\{F'_d\}$ in the proof of Theorem 3.19 to replace the sequence of $\{F'_d\}$ used in Lemma A.3. The residual proof follows the same as Theorem 3.19.

E. Proof of Proposition 4.6

Note that we have the number of parameters $d = n^2 + n + 2$ in this example. The target distribution (i.e. the posterior distribution) satisfies

$$
\begin{aligned}
\pi^d(x^d) &= \mathbb{P}(x^d \mid \{Y_{ij}\}_{i,j=1}^n) \\
&\propto \frac{b^a}{\Gamma(a)} A^{-a-1} e^{-b/A} \prod_{j=1}^n \frac{1}{\sqrt{2\pi A}} e^{-\frac{(\mu_j - \nu)^2}{2A}} \prod_{i=1}^n \frac{1}{\sqrt{2\pi V}} e^{-\frac{(\theta_{ij} - \mu_j)^2}{2V}} \frac{1}{\sqrt{2\pi W}} e^{-\frac{(Y_{ij} - \theta_{ij})^2}{2W}}.
\end{aligned}
\tag{216}
$$

Clearly, this model can be represented by the graphical model in Eq. (41). It can be easily checked that the maximum number cliques any coordinate belongs to is $n + 1$ and the cardinality of cliques is bounded by constant 2, so $\sup_k |C_k| = o(d/m_d) = o(n)$. Furthermore, the target distribution clearly satisfies "flat tail" condition required by Proposition 4.3 since all the conditional distributions are standard distributions. Therefore, the first equation in A1, the first equation in A3, and A2 hold by Proposition 4.3.

Next, we verify A5 using Proposition 4.4. Note that this model can be represented by the graphical model in Eq. (42) using $K = 3$ layers. In order to check the conditions in Proposition 4.4, note that

$$
\log \pi^d \propto \left(-a - 1 - \frac{n}{2}\right) \log A - \frac{b}{A} - \frac{\sum_j (\mu_j - \nu)^2}{2A} - \frac{\sum_{i,j} (\theta_{ij} - \mu_j)^2}{2V} - \frac{\sum_{i,j} (Y_{ij} - \theta_{ij})^2}{2W}.
\tag{217}
$$

Observing that, under $X^d = (\nu, A, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n) \sim \pi^d$, we have

$$
\theta_{ij} \mid Y_{ij}, \mu_j \quad \sim^{\text{indep.}} \mathcal{N}\left(\frac{W\mu_j + VY_{ij}}{W + V}, \frac{VW}{W + V}\right), \quad i,j \in \{1, \ldots, n\},
\tag{218}
$$

$$
\mu_j \mid \sum_i \theta_{ij}, \nu, A \sim^{\text{indep.}} \mathcal{N}\left(\frac{\sum_i A\theta_{ij} + V\nu}{nA + V}, \frac{AV}{nA + V}\right), \quad i \in \{1, \ldots, n\},
\tag{219}
$$

$$
\nu \mid \bar{\mu}, A \sim \mathcal{N}\left(\bar{\mu}, \frac{A}{n}\right),
\tag{220}
$$

$$
A \mid \{\mu_j\}, \nu \sim \mathbf{IG}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_j (\mu_j - \nu)^2\right).
\tag{221}
$$

Therefore, we have

$$
\left|\frac{\partial \log \pi^d}{\partial A}\right| = \left|\frac{b + \frac{1}{2}\sum_j (\mu_j - \nu)^2}{A^2} - \frac{a + 1 + \frac{n}{2}}{A}\right| = \mathcal{O}_{\mathbb{P}}(d^{1/2}).
\tag{222}
$$

since $\frac{a+1+n/2}{A} \to_{\mathbb{P}} \frac{a+1+n/2}{A_0} = \mathcal{O}(d^{1/2})$ and $\sum_j (\mu_j - \nu)^2 \to_{\mathbb{P}} \sum_j (\mu_j - \bar{\mu})^2 + \frac{A_0}{n} = \mathcal{O}_{\mathbb{P}}(d^{1/2})$. Other coordinates can also be verified, which are shown as follows.

$$\left(\frac{\partial \log \pi^d}{\partial \nu} \right)^2 = \left(\frac{n(\bar{\mu} - \nu)}{A} \right)^2 = \mathcal{O}_{\mathbb{P}}\left(\frac{n}{A} \right) = \mathcal{O}_{\mathbb{P}}(d/n), \tag{223}$$

$$\left(\frac{\partial \log \pi^d}{\partial \mu_j} \right)^2 = \left(\frac{\sum_i (\theta_{ij} - \mu_j)}{V} - \frac{\mu_j - \nu}{A} \right)^2 = (nA + V)^2 \left(\frac{A \sum_i \theta_{ij} + V\nu}{nA + V} - \mu_j \right)^2 \tag{224}$$

$$= \mathcal{O}_{\mathbb{P}}\left[(nA + V)^2 \frac{AV}{nA + V} \right] = \mathcal{O}_{\mathbb{P}}(d/n), \tag{225}$$

$$\left(\frac{\partial \log \pi^d}{\partial \theta_{ij}} \right)^2 = \left(\frac{Y_{ij} - \theta_{ij}}{V} - \frac{\theta_{ij} - \mu_j}{W} \right)^2 = (W + V)^2 \left(\frac{VY_{ij} + W\mu_j}{W + V} - \theta_{ij} \right)^2 = \mathcal{O}_{\mathbb{P}}(d/n^2). \tag{226}$$

$$\tag{227}$$

Therefore, A5 holds by Proposition 4.4. Finally, all the other conditions in A1, A3, and A4 can be verified in a similar way as in Section 4.1 for Example 4.1.