On some asymptotic properties of the Expectation-Maximization Algorithm and the Metropolis-Hastings Algorithm (EMアルゴリズムとメトロポリス-ヘイスティングスアルゴリズムの漸近的性質)

鎌谷 研吾

Contents

Preface

The present paper is concerned with some asymptotic properties for two kinds of Monte Carlo iteration methods: the Expectation-Maximization (EM) algorithm and the Metropolis-Hastings (MH) algorithm.

The term, *asymptotic* has two meanings in this paper. In most works related to the asymptotic properties for those iteration methods, the meaning of asymptotic is in the sense that the number of iterations goes to infinity. In Chapter 3, we address this type of asymptotic property. On the other hand, in Chapter 1 and 2, the meaning of asymptotic is that not only the number of iterations, but also the number of observations goes to infinity. Therefore, in the second meaning, we assume a large sample.

In the first two chapters, we make a framework for the asymptotic theory for the EM algorithm and the Gibbs sampler, which is a popular sub class of the MH algorithm. There are three motivations.

First, using the framework, we can validate the convergence of the EM algorithm and the Gibbs sampler. For the EM algorithm, we prove that the sequence generated by the algorithm converges to the maximum likelihood estimator. This type of convergence is hard to show in finite sample size. For the Gibbs sampler, several previous works have already addressed the validation issues for the convergence to the Bayesian estimator in finite sample size. We give another sense of the convergence property. The former convergence property is concerned with the behavior of the Gibbs sampler in the region far from the true parameter of the parameter space. On the other hand, the latter is concerned with the behavior around the true parameter.

Second, using the framework, we can validate some speed up methods of the EM algorithm and the Gibbs sampler. There are a lot of speed up methods for the EM algorithm and the Gibbs sampler. We validate these speed up methods in the framework of the asymptotic theory.

Third, the framework may be beneficial for more complicated Monte Carlo methods. We approximate the traditional algorithms by simple algorithms. This approximation may be useful for the Monte Carlo EM algorithm or some adaptive Monte Carlo methods.

In Chapter 3, we develop the results on polynomial ergodicity of Markov chains and apply to the MH algorithms based on a Langevin diffusion. When a prescribed distribution p has heavy tails, the MH algorithms based on a Langevin diffusion do not converge to p at any geometric rate. However those Langevin based algorithms behave like the diffusion itself in the tail area, and using this fact, we provide sufficient conditions of a polynomial rate convergence. By the feature in the tail area, our results can be applied to a large class of distributions to which p belongs. Then we show that the convergence rate can be improved by a transformation. We also prove central limit theorems for those algorithms.

I am grateful to Prof. Nakahiro Yoshida for his helpful comments and corrected a lot of errors. He also suggested me to construct my result on regular statistical experiments not on independent and identically distributed observations.

This work is in part supported by Grant-in-Aid for JSPS Fellows (19-3140) and the 21st Century COE Program.

Common Notation

Let a and b be real numbers.

$a \vee b := \max a, b$

$a \wedge b := \min a, b$

$a^+ := \max a, 0$

$B_\epsilon(x)$: open ball with radius ϵ centered at x in a metric space

n: sample size

$\mathbf{N} := \{1, 2, \ldots, \}$

$\mathbf{N}_0 := \{0, 1, 2, \ldots\}$

$\mathbf{R}$: real line

$\mathbf{R}^d$: d-dimensional Euclidean space

$x_{l:m} = (x_l, x_{l+1}, \ldots, x_m)$: subsequence of $x = (x_0, \ldots, x_k)$ for any $0 \leq l \leq m \leq k$.

$\phi(x; \mu, \Sigma)$: density of normal distribution with mean $\mu \in \mathbf{R}^d$, and the covariance matrix $\Sigma \in \mathbf{R}^{d \times d}$

For any signed measure ν on a measurable space $(\mathcal{Z}, \mathcal{C})$, let

$$\|\nu\|_{\mathrm{TV}} := \sup_{\{f; |f| \leq 1\}} \nu(f) = \sup_{\{f; |f| \leq 1\}} \int_{\mathcal{Z}} f(z) \nu(dz).$$

If $(\mathcal{Z}_n, \mathcal{C}_n, \mathcal{P}_n)$ be a sequence of probability spaces, and $(Z_n : \mathcal{Z}_n \to \mathbf{R}^d)$ is a sequence of random variables, we write $Z_n = o_{P_n}(1)$ to mean

$$\lim_{n \to \infty} P_n(|Z_n| > \epsilon) = 0$$

for any $\epsilon > 0$.

Chapter 1

Asymptotic Properties for the EM algorithm

1.1 Introduction

In this chapter, we are concerned with the asymptotic properties for some EM algorithms in the large sample framework. Our meaning of *asymptotic* is that both the number of iteration and the number of observation tend to infinity. The main results in this chapter are convergence theorems, a validation of the rate matrix and its application to independent and identically distributed observations.

In the finite sample theory, the convergence properties had already been established. The most important property is the existence of a monotone convergence theorem ([43] and [4]. See also [24]). However, the theorem does not tell us whether the sequence generated by the EM algorithm does converge to the maximum likelihood estimator (MLE). It may converge to a local maxima or a local minima of the likelihood function, and it may not converge to any point (see Section 3.6 of [25]). On the other hand, in the framework of the large sample theory, we can show that the sequence generated by the EM algorithm converges to the MLE, if we assume the sequence starts from an estimator T_n such that $n^{1/2}(T_n - \theta_0)$ is tight with respect to $P_{\theta_0}^{(n)}$ when θ_0 is the true value.

The rate matrix is used to measure the convergence rate of the EM algorithms (for example, [7], [28] and [26]). Unfortunately, in our meaning of convergence, we can not find any validation for the convergence rate in previous works. On the other hand, in the large sample framework, it is clear that the rate matrix determine the convergence rate.

As we mentioned, we assume that the initial point of the sequence of the EM algorithm T_n satisfies above tightness condition. This is similar to the case of the one-step estimator [5]. It is well known that the choice of the initial point is very important for the performance of the algorithm. Without this assumption, the sequence may not converge to the MLE. An estimator satisfying the tightness condition exists in general (for example, see [5]). In many cases, the moment estimator works well.

1.2 Matrix Algebra

In this section, we review some key elements of matrix algebra which will be used in a later section. Consider the space $\mathbf{C}^d$ with the inner product $\langle u, v\rangle$, that is,

$$\langle u, v\rangle = \sum_{i=1}^{d} u_i \overline{v_i} \ (u = (u_1, \ldots, u_d)^T, v = (v_1, \ldots, v_d)^T),$$

and $|u|^2 = \langle u, u\rangle$.

Lemma 1.1. *Let $A, C \in \mathbf{R}^{d\times d}$ be positive definite matrices such that $C - A$ is nonnegative definite. Let $K = C^{-1}(C - A)$. Then K and $L = C^{-1/2}(C - A)C^{-1/2}$ have the same eigenvalues and the same algebraic multiplicity for each eigenvalue. Moreover, K is diagonalizable and each eigenvalue λ of K is $0 \le \lambda < 1$.*

Proof. Let λ be an eigenvalue of K and u be one of its eigenvectors. Then, $Ku = \lambda u$ and we have $(C - A)u = \lambda Cu$. It is easy to see that the sets of eigenvalues of K and L are the same. Therefore, since L is nonnegative, λ is real and $\lambda \ge 0$. Since A is positive definite, if $\lambda \ne 0$, then $\langle u, (C - A)u\rangle = \lambda\langle u, Cu\rangle > \lambda\langle u, (C - A)u\rangle$ and hence $\lambda \in (0, 1)$. Therefore, $\lambda \in [0, 1)$ and the claim follows.

□

Let $d_1, \ldots, d_k$ be integers such that $\sum_{i=1}^{k} d_i = d$. Then, we divide any $d \times d$-matrix into k^2 partitions, such that

$$M = \begin{pmatrix} M_{1,1} & \ldots & M_{1,k} \\ \vdots & \ddots & \vdots \\ M_{k,1} & \ldots & M_{k,k} \end{pmatrix}$$

where $M_{i,j}$ is a $d_i \times d_j$-matrix. Let M be denoted by $(M_{i,j}; i, j = 1, \ldots, k) = (M_{i,j})$. We define some matrices related to the matrix M. Let $\operatorname{diag}(M)$

denote a $d \times d$-matrix $S = (S_{i,j})$ such that $S_{i,j} = 1_{\{i=j\}} M_{i,j}$, and M^l and M^u denote $T = (T_{i,j})$ and $U = (U_{i,j})$ such that $T_{i,j} = 1_{\{i \geq j\}} M_{i,j}$ and $U_{i,j} = 1_{\{i \leq j\}} M_{i,j}$.

For any $d \times d$-matrix K, let $\|K\|_2 = \sup_{|h|=1} |Kh|$.

Lemma 1.2. *Let $A, C \in \mathbf{R}^{d \times d}$ be positive definite matrices such that $C - A$ is nonnegative definite. Let $K = (C^l)^{-1}(C^l - A)$. Then there exists $r \in [0,1)$ and that $|\lambda| \leq r$ for any eigenvalue λ of K, and for any $\epsilon > 0$, there exist a nonsingular matrix P and a lower triangular matrix Λ such that $K = P^{-1}\Lambda P$ and $\|\Lambda\|_2 \leq r + \epsilon$.*

Proof. Let λ be an eigenvalue of K and u be one of its eigenvectors. Then, $Ku = \lambda u$ and we have $(C^l - A)u = \lambda C^l u$. We have $\langle u, (C^l - A)u \rangle = \overline{\lambda}\langle u, C^l u\rangle$, and its transpose $\langle u, (C^u - A)u\rangle = \lambda \langle u, C^u u\rangle$. Then we have

$$(1 - \overline{\lambda})\langle u, C^l u\rangle = \langle u, Au\rangle, \text{ and } (1-\lambda)\langle u, C^u u\rangle = \langle u, Au\rangle.$$

Multiplying above equations by $(1-\lambda)$ or $(1-\overline{\lambda})$, we obtain

$$|1-\lambda|^2 \langle u, C^l u\rangle = (1-\lambda)\langle u, Au\rangle, \text{ and } |1-\lambda|^2\langle u, C^u u\rangle = (1-\overline{\lambda})\langle u, Au\rangle.$$

The sum of the two equations yields

$$0 = |1-\lambda|^2 \langle u, (D + C)u\rangle - (2 - \lambda - \overline{\lambda})\langle u, Au\rangle,$$

where $D = \operatorname{diag}(C)$. Since D is a positive definite matrix and $C - A$ is nonnegative definite matrix, we have

$$0 > |1-\lambda|^2\langle u, Au\rangle - (2-\lambda-\overline{\lambda})\langle u, Au\rangle = (|\lambda|^2 - 1)\langle u, Au\rangle.$$

Since A is positive definite matrix, there exists $0 \leq r < 1$ such that for any eigenvalue λ of K, we have $|\lambda| < r$. Therefore, there exist a nonsingular matrix $P_1 \in \mathbf{C}^{d\times d}$ and a lower triangular matrix $\Lambda_1 = (\lambda_{1,i,j}; i,j = 1, \ldots d) \in \mathbf{C}^{d \times d}$ such that $K = P_1^{-1}\Lambda_1 P_1$ and $|\lambda_{1,i,i}| \leq r$ (for example, see [12]). Fix any $\epsilon \in (0,1)$. Let $s = \max_{i \neq j} |\lambda_{1,i,j}| \vee 1$. Take $D = \operatorname{diag}((\epsilon/s), (\epsilon/s)^2, \ldots, (\epsilon/s)^d)$. Then we have

$$(D\Lambda_1 D^{-1})_{i,j} = (\frac{\epsilon}{s})^{i-j}\lambda_{i,j}.$$

Let $P = DP_1$ and $\Lambda = (\lambda_{i,j}; i,j = 1, \ldots, k) = PKP^{-1} = D\Lambda_1 D^{-1}$. Then Λ is also a lower triangular matrix, and $|\lambda_{i,i}| = |\lambda_{1,i,i}| \leq r$, $|\lambda_{i,j}| \leq \epsilon$ $(i > j)$, and $\lambda_{j,i} = 0$ $(i < j)$. For any $u = (u_1, \dots, u_d)^T \in \mathbf{C}^d$ such that $|u| = 1$, we have

$$\begin{aligned}
|\Lambda u|^2 &= \sum_{i=1}^{d} |\sum_{j=1}^{i} \lambda_{i,j} u_j|^2 \leq \sum_{i=1}^{d} (r|u_i| + \epsilon \sum_{j=1}^{i-1} |u_j|)^2 \\
&= \sum_{i=1}^{d} (r^2|u_i|^2 + 2r\epsilon \sum_{j=1}^{i-1} |u_i||u_j| + \epsilon^2 \sum_{l,m=1}^{i-1} |u_l||u_m|) \\
&\leq \sum_{i=1}^{d} (r^2|u_i|^2 + r\epsilon \sum_{j=1}^{i-1} (|u_i|^2 + |u_j|^2) + \frac{\epsilon^2}{2} \sum_{l,m=1}^{i-1} (|u_l|^2 + |u_m|^2)) \\
&\leq \sum_{i=1}^{d} ((r^2 + dr\epsilon)|u_i|^2 + \sum_{j=1}^{i-1} (r\epsilon + d\epsilon^2)|u_j|^2) \\
&\leq r^2 + dr\epsilon + d(r\epsilon + d\epsilon^2) \leq r^2 + 2d\epsilon + d^2\epsilon^2.
\end{aligned}$$

Therefore, $\|\Lambda\|_2^2 \leq r^2 + 2d\epsilon + d^2\epsilon^2$. Since $\epsilon > 0$ is arbitrary, the conclusion follows. □

1.3 Regular Statistical Experiments

Suppose Θ is an open subset of $\mathbf{R}^d$. The parameter space Θ is equipped with its Borel σ-algebra $\mathcal{F}$. Fix any $\theta_0 \in \Theta$. The element θ_0 will be used as a true value of the following model. Consider a family of statistical experiments $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{h,n}; h \in H_n)$, where $H_n = n^{1/2}(\Theta - \theta_0)$. We set $\mathcal{H}_n = \sqrt{n}(\mathcal{F} - \theta_0)$.

1.3.1 Two-Stage EM Algorithm

Suppose a function $Q_{\cdot,\cdot,\cdot,n} : \mathcal{X}_n \times H_n \times H_n \to [-\infty, \infty)$ is $\mathcal{A}_n \times \mathcal{H}_n \times \mathcal{H}_n$-measurable. Let $I(\theta_0)$ and $I_{2|1}(\theta_0)$ be $d \times d$-matrices, and $I_{1,2}(\theta_0) = I(\theta_0) + I_{2|1}(\theta_0)$. Let $Z_n : \mathcal{X}_n \to \mathbf{R}^d$ be a sequence of random variables.

Assumption 1.3.1. *The matrix $I(\theta_0)$ is positive definite and $I_{2|1}(\theta_0)$ is nonnegative definite, and $(Z_n; n \in \mathbf{N})$ is $P_{0,n}$-tight.*

Assumption 1.3.2. *There exists $M_n \to \infty$ such that*

$$\sup_{|h|,|g| \leq M_n} |Q_{x,g,h,n} - \overline{Q}_{x,g,h,n}| = o_{P_{0,n}}(1),$$

where $\overline{Q}_{x,g,h,n}$ *is*

$$\langle h, Z_n + I_{2|1}(\theta_0)g\rangle - \frac{1}{2}\langle h, I_{1,2}(\theta_0)h\rangle + \frac{1}{2}\langle g, (I(\theta_0) - I_{2|1}(\theta_0))g\rangle - \langle g, Z_n\rangle.$$

We note a few comments about the assumptions. The matrix $I_{1,2}(\theta_0)^{-1}I_{2|1}(\theta_0)$ plays an important role in the asymptotic theory. Suppose Assumption 1.3.1, then $I_{1,2}(\theta_0)^{-1}I_{2|1}(\theta_0)$ is diagonalizable and all eigenvalues are nonnegative and smaller than 1. See Lemma 1.1.

We are going to define the *Expectation-Maximization (EM)* algorithm, which was formulated in [7]. Fix any $n \in \mathbf{N}$. When we have an observation x from $P_{0,n}$, we define the EM algorithm starting from $h \in H_n$ as follows. First, set $h_0 = h$, and go to step 1.

Step i The step consists of further two minor steps, *Expectation-step (E-step)* and *Maximization-step (M-step)*.

E-step Calculate the value $Q_{x,h_{i-1},\cdot,n} : H_n \to \mathbf{R} \cup \{-\infty\}$.

M-step Maximize $Q_{x,h_{i-1},h,n}$ with respect to h and set the maximizer as h_i. If there are more than two maximizers, select one of the nearest maximizer from h_{i-1}. Go to Step $i + 1$.

We call $(h_i = h_{x,i,n}; i \in \mathbf{N})$ the *sequence of the EM algorithm* (starting from h). Sometimes, the M-step is replaced by finding one of the roots of $\partial Q_{x,h_{i-1},h,n}/\partial h = 0$. This minor change does not affect our results, so even in this case, we also call it the sequence of the EM algorithm.

Lemma 1.3. *Let* $(\mathcal{Z}_n, \mathcal{C}_n, P_n)$ $(n = 1, 2, \ldots)$ *be a sequence of probability spaces. Let* $(W_n; n \in \mathbf{N})$ *be a sequence of* P_n*-tight random variables on* $\mathbf{R}^d$, *and let* $K = P^{-1}\Lambda P$ *be a* $d \times d$*-matrix, where* P *is a nonsingular matrix and* $\|\Lambda\|_2 = r$ *for some* $r < 1$. *Let* $f_n : \mathcal{Z}_n \times \mathbf{R}^d \to \mathbf{R}^d$ *be a random function, and let* $f_n^{(0)}(z, h) = h$ *and* $f_n^{(i+1)}(z, h) = f_n(z, f_n^{(i)}(z, h))$ $(i = 0, 1, 2, \ldots)$. *If there exist some* $M_n \to \infty$ *such that*

$$\sup_{|h| \le M_n} |(f_n(z, h) - W_n(z)) - K(h - W_n(z))| = o_{P_n}(1)$$

then for some $M'_n \to \infty$, *we have*

$$\sup_{|h| < M'_n} \sup_{i \in \mathbf{N}} |(f_n^{(i)}(z, h) - W_n(z)) - K^i(h - W_n(z))| = o_{P_n}(1).$$

Proof. Fix $\epsilon > 0$ and assume that n is large enough to be $\epsilon < M_n$. Taking a set $A_n \in \mathcal{C}_n$ as

$$\{z; \sup_{|h|\le M_n} |(f_n(z,h)-W_n(z))-K(h-W_n(z))| \le \frac{\epsilon(1-r)}{4c}, |W_n(z)| \le \frac{M_n}{4(1\vee c)}\}$$

then $P_n(A_n) \to 1$, where $c = \|P\|_2\|P^{-1}\|_2$. Let $\epsilon_{i,n}(z,h) = (f_n^{(i)}(z,h) - W_n(z)) - K(f_n^{(i-1)}(z,h) - W_n(z))$, then we have

$$\begin{aligned}
&|(f_n^{(i)}(z,h) - W_n(z)) - K^i(h - W_n(z))| \\
&= |\sum_{j=0}^{i-1} K^j((f_n^{(i-j)}(z,h) - W_n) - K(f_n^{(i-j-1)}(z,h) - W_n))| \\
&= |\sum_{j=0}^{i-1} K^j\epsilon_{i-j,n}(z,h)| \le \frac{c}{1-r} \sup_{j=1,\dots,i} |\epsilon_{j,n}(z,h)|.
\end{aligned}$$

If $|h| \le M_n/4(1\vee c)$ and $z \in A_n$, we show that using induction, we have $|f_n^{(i)}(z,h)| \le M_n$ for all $i \in \mathbf{N}$. When $i = 0$, $|f_n^{(0)}(z,h)| = |h| \le M_n$ and if $|f_n^{(j)}(z,h)| \le M_n$ for $j = 1,\dots,i-1$, then

$$\begin{aligned}
|f_n^{(i)}(z,h)| &\le |(f_n^{(i)}(z,h) - W_n(z)) - K^i(h - W_n(z))| + |W_n(z)| + cr^i|h - W_n(z)| \\
&\le \frac{\epsilon}{4} + \frac{M_n}{4} + \frac{M_n}{2} \le M_n.
\end{aligned}$$

Hence the claim of the lemma follows for $M_n' = M_n/4(1\vee c)$. □

We note a few comments about the lemma. The matrix K is called the *rate matrix.* The rate matrix K has the form $K = P^{-1}\Lambda P$ for all algorithms in this paper, where P is a nonsingular matrix, and $\|\Lambda\|_2 = r < 1$. Then, $\|K^i\| \le \|P\|\|P^{-1}\|r^i$. Therefore, r is considered to be the upper bound of the convergence rate of $f_n^{(i)}(z,h)$. This r is called the *rate of convergence* of f_n.

Consider we want to compare two random functions $f_{n,1}$ and $f_{n,2}$ with the rate matrices $K_1 = P_1^{-1}\Lambda_1 P_1$ and $K_2 = P_2^{-1}\Lambda_2 P_2$, where P_i is a nonsingular matrix, and $\|\Lambda_i\|_2 < 1$ for $i = 1, 2$. The random function $f_{n,2}$ is preferable if $\|\Lambda_1\|_2 > \|\Lambda_2\|_2$, since the convergence rate of $f_{n,2}$ is better than that of $f_{n,1}$.

Theorem 1.1. *Let Assumptions 1.3.1 and 1.3.2 be satisfied. Then for some $M_n \to \infty$, for $f_n(x,g) := \arg\max_{|h|\le M_n} Q_{x,g,h,n}$, we have*

$$\sup_{|h|\le M_n} |(f_n(x,h) - I(\theta_0)^{-1}Z_n(x)) - (I_{1,2}^{-1}(\theta_0)I_{2|1}(\theta_0))(h - I(\theta_0)^{-1}Z_n(x))| = o_{P_{0,n}}(1).$$

In particular, for $f_n^{(0)}(x,h) = h$ *and* $f_n^{(i+1)}(x,h) = f_n(x, f_n^{(i)}(x,h))$ $(i = 0,1,2,\ldots)$, *the following value tends in probability to* 0 *for some* $M_n' \to \infty$*:*

$$\sup_{|h|<M_n'} \sup_{i\in\mathbf{N}} |(f_n^{(i)}(x,h) - I(\theta_0)^{-1}Z_n(x)) - (I_{1,2}^{-1}(\theta_0)I_{2|1}(\theta_0))^i(h - I(\theta_0)^{-1}Z_n(x))|.$$

Proof. Let $\mu_n(g) = I_{1,2}(\theta_0)^{-1}(I_{2|1}(\theta_0)g + Z_n)$. Then $\overline{Q}_{x,g,h,n}$ can be written in the form:

$$-\frac{1}{2}\langle h - \mu_n(g), I_{1,2}(\theta_0)(h - \mu_n(g))\rangle + C_n(x,g),$$

where $C_n(x,g)$ is a constant which is not depend on h. Let

$$I_{\epsilon,n}(g) = \{h \in \mathbf{R}^d; \frac{1}{2}\langle h - \mu_n(g), I_{1,2}(\theta_0)(h - \mu_n(g))\rangle \leq \epsilon\}.$$

Let $r = \|I_{1,2}(\theta_0)^{-1}I_{2|1}(\theta_0)\|_2 \vee 1$ and $R_{x,n} = \sup_{|g|,|h|\leq 2rM_n} |Q_{x,g,h,n} - \overline{Q}_{x,g,h,n}|$. If $R_{x,n} < \epsilon$, and $|I_{1,2}^{-1}(\theta_0)Z_n| \leq rM_n$, then $f_n(x,g) \in I_{2\epsilon,n}(g)$ for any $|g| \leq M_n$, since if $h = f_n(x,g) \in I_{2\epsilon,n}(g)^c \cap \overline{B_{M_n}(0)}$, then

$$Q_{x,g,\mu_n(g),n} > \overline{Q}_{x,g,\mu_n(g),n} - \epsilon = C_n(x,g) - \epsilon > \overline{Q}_{x,g,h,n} + \epsilon > Q_{x,g,h,n}.$$

Let s be the smallest eigenvalue of $I_{1,2}(\theta_0)$ which is not 0. If $h \in I_{2\epsilon.n}(g)$, we have $|h - \mu_n(g)|^2 < 2\epsilon/s$. Since $\epsilon > 0$ is arbitrary, the first claim follows. Using Lemma 1.1, the second claim is easy corollary of Lemma 1.3. □

We state few comments about the theorem. The above sequence $f_n^{(i)}(x,h)$ $i = 1,\ldots$ are equal to the sequence of EM algorithm starting from h in $P_{0,n}$-probability. Therefore, we can replace $f_n(x,g)$ by the one step of the EM algorithm from g.

The local convergence rate of the EM algorithm is determined by $J(\theta_0) = J_0(\theta_0) = I_{1,2}^{-1}(\theta_0)I_{2|1}(\theta_0)$, which was already suggested by [7] and this fact was used in a lot of papers, though we can not find any theoretical validation. The matrix $I_{1,2}^{-1}(\theta_0)I_{2|1}(\theta_0)$ is the rate matrix for the algorithm.

1.3.2 Multi-Stage EM Algorithm

There are a number of multi-stage EM algorithms. We consider one of them, the *expectation-conditional maximization (ECM)* algorithm, which was introduced in [27].

Let $\Theta = \otimes_{i=1}^k \Theta_i$, where each Θ_i is an open subset of $\mathbf{R}^{d_i}$ and $\sum_{i=1}^k d_i = d$. Using this representation, we write $\theta_0 = (\theta_{0,1},\ldots,\theta_{0,k})$ and $H_n = \otimes_{i=1}^k H_{n,i}$.

We also divide $d \times d$-matrix M into k^2 partitions and define $\mathrm{diag}(M)$, M^l and M^u as the definition after Lemma 1.1.

For simplicity, we use the notation $h_{l:m} = (h_l, h_{l+1}, \ldots, h_m)$ for a sequence $h = (h_0, h_1, \ldots)$, and for $l, m \in \mathbf{N}_0$, $l \leq m$.

The algorithm is defined as follows. Assume that we have an observation x from $P_{0,n}$. First, fix any $h^0 = (h_1^0, \ldots, h_k^0) \in H_n$, and go to step 1.

Step i The step consists of further $k+1$ minor steps, one E-step and k *Conditional Maximization-steps (CM-steps).*

E-step Calculate the value $Q_{x,h^{i-1},\cdot,n} : H_n \to \mathbf{R}$. Go to 1st CM-step.

jth CM-step Find $f \in H_{n,j}$ that maximize

$$Q_{x,h^{i-1},(h^i_{1:j-1},f,h^{i-1}_{j+1:k}),n}.$$

and set $h_j^i = f$. If there are more than two maximizers, select one of the nearest maximizer from h^{i-1}. Go to $j+1$th CM-step if $j < k$ and go to step $i+1$ when $j = k$.

We call $(h_i = h_{x,i,n}; i \in \mathbf{N})$ the *sequence of the ECM algorithm* (starting from h^0).

Note that, by definition, we have

$$\begin{aligned}
&\overline{Q}_{x,g,(h_{1:i-1},f,g_{i+1:k}),n} \\
&= -\frac{1}{2}\langle f, I_{1,2}(\theta_0)_{i,i} f\rangle + C \\
&+ \langle f, Z_{n,i} + \sum_{j=1}^{k} I_{2|1}(\theta_0)_{i,j} g_j - \sum_{j<i} I_{1,2}(\theta_0)_{i,j} h_j - \sum_{j>i} I_{1,2}(\theta_0)_{i,j} g_j \rangle,
\end{aligned}$$

where C is a term which does not depend on f. The above function with respect to f takes the maximum at $f = (I_{1,2}(\theta_0)_{i,i})^{-1}(Z_{n,i} - \sum_{j<i} I_{1,2}(\theta_0)_{i,j} h_j - \sum_{j>i} I_{1,2}(\theta_0)_{i,j} g_j + \sum_{j=1}^{k} I_{2|1}(\theta_0)_{i,j} g_j)$. We will show that each maximization of $Q_{x,h^{i-1},(h^i_{1:j-1},f,h^{i-1}_{j+1:k}),n}$ with respect to f is almost equal to the maximization of $\overline{Q}_{x,h^{i-1},(h^i_{1:j-1},f,h^{i-1}_{j+1:k}),n}$. Therefore, starting from $g \in \mathbf{R}^d$, after one E-step and k CM-steps, we have $h \in \mathbf{R}^d$ as follows as the result of one step iteration:

$$h \sim E^{-1}(Z_n - (C^l - E)h - (C^u - E)g + Bg),$$

where $B = I_{2|1}(\theta_0)$, $C = I_{1,2}(\theta_0)$ and $E = \mathrm{diag}(I_{1,2}(\theta_0))$. We have $C^l h \sim Z_n + (B+E-C^u)g = Z_n + (C^l - A)g$, and hence, $h - I(\theta_0)^{-1}Z_n \sim J_1(\theta_0)(g - I(\theta_0)^{-1}Z_n)$, where

$$J_1(\theta_0) = (I_{1,2}(\theta_0)^l)^{-1}(I_{1,2}(\theta_0)^l - I(\theta_0)).$$

This is the rate matrix of the algorithm.

Corollary 1.1. *Let Assumptions 1.3.1 and 1.3.2 be satisfied. For some $M_n \to \infty$, and for $f_n : \mathcal{X}_n \times H_n \to \mathbf{R}^d$ such that for $h = (h_1, \ldots, h_d) = f_n(x,g)$,*

$$h_i = \mathrm{argmax}_{|f| \le M_n} Q_{x,g,(h_{1:i-1},f,g_{i+1:k}),n} \ (i = 1, \ldots, d),$$

we have

$$\sup_{|g| \le M_n} |(f_n(x,g) - I(\theta_0)^{-1}Z_n(x)) - J_1(\theta_0)(g - I(\theta_0)^{-1}Z_n(x))| = o_{P_{0,n}}(1).$$

In particular, $f_n^{(0)}(x,h) = h$ and $f_n^{(i+1)}(x,h) = f_n(x, f_n^{(i)}(x,h))$ $(i = 0, 1, 2, \ldots)$. the following value tends in probability to 0 for some $M_n' \to \infty$:

$$\sup_{|h| \le M_n} \sup_{i \in \mathbf{N}} |(f_n^{(i)}(x,h) - I(\theta_0)^{-1}Z_n(x)) - J_1(\theta_0)^i(h - I(\theta_0)^{-1}Z_n(x))|.$$

Proof. For any $i = 1, \ldots, k$ and $|g| \le M_n$, let

$$f_{n,i}(x, g, h_{1:i-1}) = \arg\max_{|h_i| \le M_n} Q_{x,g,(h_{1:i-1},h_i,g_{i+1:k}),n}.$$

Using the same argument as Theorem 1.1, taking

$$g_{n,i}(x,g,h_{1:i-1}) = (I_{1,2}(\theta_0)_{i,i})^{-1}(Z_{n,i} - \sum_{j<i} I_{1,2}(\theta_0)_{i,j}h_j - \sum_{j>i} I_{1,2}(\theta_0)_{i,j}g_j + \sum_{j=1}^{k} I_{2|1}(\theta_0)_{i,j}g_j),$$

the following value tends in $P_{0,n}$-probability to 0:

$$\sup_{|g_l|,|h_m| \le M_n, 1 \le l \le k, 1 \le m \le i-1} |(f_{n,i}(x,g,h_{1:i-1}) - g_{n,i}(x,g,h_{1:i-1})| = o_{P_{0,n}}(1).$$

Therefore, we have

$$\sup_{|g| \le M_n} |(f_n(x,g) - I(\theta_0)^{-1}Z_n) - J_1(\theta_0)(g - I(\theta_0)^{-1}Z_n)| = o_{P_{0,n}}(1).$$

Let $C = I_{1,2}(\theta_0)$ and $A = I(\theta_0)$. Using Lemma 1.2, the second claim is easy corollary of Lemma 1.3. □

We note a few comments about the corollary. We can insert E-step after each CM-step in ECM algorithm, which is a special case of *Multi Cycle ECM (MCECM)*. Let

$$J_2(\theta_0) = (\mathrm{diag}(I_{2|1}(\theta_0)) + I(\theta_0)^l)^{-1}(\mathrm{diag}(I_{2|1}(\theta_0)) + I(\theta_0)^l - I(\theta_0)).$$

Let $C = \mathrm{diag}(I_{2|1}(\theta_0)) + I(\theta_0)$ and $A = I(\theta_0)$. Then, using Lemma 1.2, we can prove the same conclusion for the algorithm as Corollary 1.1 if we replace $J_1(\theta_0)$ by $J_2(\theta_0)$.

We can compare these algorithms by the largest eigenvalue in absolute value of the rate matrix. The order of the largest eigenvalues vary with a change in the matrices, $I(\theta_0)$ and $I_{2|1}(\theta_0)$ (see [26]). We do not treat this comparison in detail in this paper.

1.3.3 Point Estimation

We assume an estimator $T_n : \mathcal{X}_n \to H_n$ to hold the following tightness property: for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} P_{0,n}(T_n \in B_\delta(0)^c) \leq \epsilon.$$

If the estimator is non-localized, that is, $T_n : \mathcal{X}_n \to \Theta$, then assume the following: for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} P_{0,n}(n^{1/2}(T_n - \theta_0) \in B_\delta(\theta_0)^c) \leq \epsilon. \tag{1.1}$$

Corollary 1.2. *Let Assumption 1.3.1 be satisfied. Let $K = P^{-1}\Lambda P$ be a $d \times d$-matrix, where P is a nonsingular matrix and $\|\Lambda\|_2 = r$ for some $r < 1$. Assume (1.1) holds for the maximum likelihood estimator $T_n = \hat{\theta}_n$ and $T_n = \theta_{x,0,n} = \theta_0 + h_{x,0,n}n^{-1/2}$. Assume there exists a random function $f_n : \mathcal{X}_n \times H_n \to \mathbf{R}^d$, such that for some $M_n \to \infty$ we have*

$$\sup_{|h|\leq M_n} \sup_{i\in\mathbf{N}} |(f_n^{(i)}(x,h) - I(\theta_0)^{-1}Z_n) - K^i(h - I(\theta_0)^{-1}Z_n)| = o_{P_{0,n}}(1),$$

where $f_n^{(0)}(x,h) = h$ and $f_n^{(i+1)}(x,h) = f_n(x, f^{(i)}(x,h))$. If $\theta_{x,i,n} = \theta_0 + f_n^{(i)}(x, h_{x,0,n})n^{-1/2}$, then for any $m_n \to \infty$, $n^{1/2}(\theta_{x,m_n,n} - \hat{\theta}_n) = o_{P_{0,n}}(1)$.

Proof. Easy. □

1.3.4 Some Extensions

There are a number of methods which speed up usual EM algorithm. These speed up methods are divided into two groups. One group consists of those methods which have a monotone convergence property (for example, [28]), and the other consists of those which have not (for example, [20]). Even for our large sample framework, which we only consider local property around the true parameter, there are large differences between them. With the existence of monotone convergence property, our framework may be applicable even in the case that we can not find a good initial guess $h_{x,0,n}$, since the algorithm will modify the value to some good values as the number of the iteration goes to infinity. In this subsection, we only consider one of the former methods.

In this subsection, we assume that the parameter spaces Θ, H_n can be divided into two components $\Theta^{(a)}, H_n^{(a)} \in \mathbf{R}^{d_a}, \Theta^{(b)}, H_n^{(b)} \in \mathbf{R}^{d_b}$ and $\Theta = \Theta^{(a)} \times \Theta^{(b)}, H_n = H_n^{(a)} \times H_n^{(b)}$. The parameter space $H_n^{(b)}$ is a dummy space, that is, $P_{(h_a,h_b),n} = P_{(h_a,g_b),n}$, where $h_a \in H_n^{(a)}$ and $h_b, g_b \in H_n^{(b)}$. The parameter space $\Theta^{(a)}$ is Θ and $\Theta^{(b)}$ has only one element in the previous subsections. Therefore, the parameter space Θ in the EM algorithm we considered should be replaced by $\Theta^{(a)}$ in their notation.

Let $\theta_0 = (\theta_0^{(a)}, \theta_0^{(b)})$. The *PX-EM algorithm* which is defined in [22], is exactly the same procedure as in Subsection 1.3.1. In the notation of the present subsection, the values in the space $\Theta^{(b)}$ is fixed throughout the iteration in the EM algorithm. Note that in the PX-EM algorithm, the user define the value $\theta_0^{(b)}$.

Let $I(\theta_0)$ and $I_{2|1}(\theta_0)$ be $d \times d$-matrices and $I_{1,2}(\theta_0) = I(\theta_0) + I_{2|1}(\theta_0)$. Let $(Z_n; \mathcal{X} \to \mathbf{R}^{d_a})$ be a sequence of random variables.

For any d-dimensional vector h, we divide h into two parts, $h_a \in \mathbf{R}^{d_a}$ and $h_b \in \mathbf{R}^{d_b}$ such that $h = (h_a, h_b)^T$. For any $d \times d$ matrix M such as $I(\theta_0), I_{1,2}(\theta_0)$, we divide M into 4 small matrices $M_{i,j}$ $(i, j = a, b)$, where $M_{i,j}$ is a $d_i \times d_j$ matrix.

Assumption 1.3.3. *The matrix $I(\theta_0)_{a,a}$ is a positive definite $d_a \times d_a$-matrix, and $I(\theta_0)_{a,b}, I(\theta_0)_{b,a}, I(\theta_0)_{b,b}$ is 0. The matrix $I_{2|1}(\theta_0)$ is a positive definite $d \times d$-matrix, and $I_{1,2}(\theta_0) = I(\theta_0) + I_{2|1}(\theta_0)$. A d_a-dimensional sequence of random variables $(Z_n; n \in \mathbf{N})$ is $P_{0,n}$-tight.*

Let $\mu_{x,n} = (I(\theta_0)_{a,a}^{-1} Z_n, (I(\theta_0)_{1,2}^{-1})_{b,a} Z_n)^T$. This value behaves as if it were a MLE in the following corollary.

Corollary 1.3. *Let Assumptions 1.3.2 and 1.3.3 be satisfied. Then for some*

$M_n \to \infty$, for $f_n(x, g) := \arg\max_{|h| \le M_n} Q_{x,g,h,n}$, we have

$$\sup_{|h| \le M_n} |(f_n(x, h) - \mu_{x,n}) - (I_{1,2}^{-1}(\theta_0) I_{2|1}(\theta_0))(h - \mu_{x,n})| = o_{P_{0,n}}(1).$$

In particular, $f_n^{(0)}(x, h) = h$ and $f_n^{(i+1)}(x, h) = f_n(x, f_n^{(i)}(x, h))$ $(i = 0, 1, 2, \ldots)$. the following value tends in probability to 0 for some $M'_n \to \infty$:

$$\sup_{|h| < M'_n} \sup_{i \in \mathbf{N}} |(f_n^{(i)}(x, h) - \mu_{x,n}) - (I_{1,2}^{-1}(\theta_0) I_{2|1}(\theta_0))^i (h - \mu_{x,n})|.$$

We note a few comments about the corollary. Since $D = I_{1,2}^{-1}(\theta_0) I_{2|1}(\theta_0)$ has a form such that $D_{a,a} = (A + E)^{-1} E$, $D_{a,b} = 0$, $D_{b,a} = -(B^{-1})_{b,a} A D_{a,a}$ and $D_{b,b} = I$, where $A = I(\theta_0)_{a,a}$, $B = I_{2|1}(\theta_0)$ and $E = B_{a,a} - B_{a,b} B_{b,b}^{-1} B_{b,a}$, the following value tends in $P_{0,n}$-probability to 0:

$$\sup_{|h| < M'_n} \sup_{i \in \mathbf{N}} |(f_n^{(i)}(x, h)_a - I(\theta_0)_{a,a}^{-1} Z_n) - D_{a,a}^i (h_a - I(\theta_0)_{a,a}^{-1} Z_n)|,$$

where $f_n^{(i)}(x, h) = (f_n^{(i)}(x, h)_a, f_n^{(i)}(x, h)_b)$. Therefore, if we concentrate on Θ_a, the rate matrix of the model is $D_{a,a}$. The model we considered in Subsection 1.3.1 can be considered to be a restriction of the model to Θ_a. Then, the rate matrix of the EM algorithm in Subsection 1.3.1 corresponds to $(A + B_{a,a})^{-1} B_{a,a}$. Then, the largest eigenvalue of the rate matrix of PX-EM algorithm is smaller than that of the EM, since $A^{-1/2} B_{a,a} A^{-1/2} - A^{-1/2} E A^{-1/2}$ is positive definite. It means that in this sense, the PX-EM algorithm is more efficient than the EM algorithm.

1.4 Independent and Identically Distributed Observations

We consider the case of independent and identically distributed observations. We take two parametric families $(\mathcal{X}, \mathcal{A}, P_\theta; \theta \in \Theta)$ and $(\mathcal{Y}, \mathcal{B}, P_{x,\theta}^{2|1}; \theta \in \Theta, x \in \mathcal{X})$. Let $\theta_0 \in \Theta$ and $H_n = n^{1/2}(\Theta - \theta_0)$. We also take an experiment $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{h,n}; h \in H_n)$, where $\mathcal{X}_n = \mathcal{X}^n, \mathcal{A}_n = \mathcal{A}^n$ and $P_{h,n} = P_{\theta_0 + hn^{-1/2}}^n$.

We assume some properties for the parametric families.

Assumption 1.4.1. *There exist a σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$ and a family of σ-finite measures $(\mu_x^{2|1}; x \in \mathcal{X})$ on $(\mathcal{Y}, \mathcal{B})$. We have $P_\theta \ll \mu$, and P_θ has a density $dP_\theta/d\mu(x) = p_\theta(x)$. We have $P_{x,\theta}^{2|1} \ll \mu_x^{2|1}$, and $P_{x,\theta}^{2|1}$ has a density $dP_{x,\theta}^{2|1}/d\mu_x^{2|1}(y) = p_{x,\theta}^{2|1}(y)$.*

Assumption 1.4.2. *For any $B \in \mathcal{B}$, $\mu_{\cdot}^{2|1}(B) : \mathcal{X} \to [0,1]$ is $\mathcal{A}$-measurable and $p^{2|1}_{\cdot,\theta}(\cdot)$ is $\mathcal{A} \times \mathcal{B}$-measurable for any $\theta \in \Theta$.*

Under Assumptions 1.4.1 and 1.4.2, let $p^{1,2}_{s,t}(x,y) = p_s(x)p^{2|1}_{x,t}(y)$ and $p^{1,2}_s = p^{1,2}_{s,s}$, and let

$$Q_{s,t}(x) = \int_{\mathcal{Y}} \log \frac{p^{1,2}_t(x,y)}{p^{1,2}_s(x,y)} dP^{2|1}_{x,s}(dy).$$

We define $\mu^{1,2}(dxdy) = \mu(dx)\mu^{2|1}_x(dy)$.

Assumption 1.4.3. *For $\theta = \theta_0$, there exist functions $\eta_\theta : \mathcal{X} \to \mathbf{R}^d$ and $\eta^{2|1}_{x,\theta} : \mathcal{Y} \to \mathbf{R}^d$ $(x \in \mathcal{X})$ such that, $\eta^{2|1}_{\cdot,\theta}(\cdot)$ is $\mathcal{A} \times \mathcal{B}$-measurable and*

$$\int_{\mathcal{X}} \frac{|\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \langle \eta_\theta(x), h\rangle|^2}{|h|^2} \mu(dx) \to 0,$$
$$\int_{\mathcal{X}\times\mathcal{Y}} \frac{|\sqrt{p^{1,2}_{\theta,\theta+h}(x,y)} - \sqrt{p^{1,2}_\theta(x,y)} - \langle \eta^{2|1}_{x,\theta}(y), h\rangle|^2}{|h|^2} \mu^{1,2}(dxdy) \to 0 \ (h \to 0).$$

Under Assumption 1.4.3, we define matrices $I(\theta) = (I(\theta)_{i,j}; i,j = 1,\dots,d)$ and $I_{2|1}(\theta) = (I_{2|1}(\theta)_{i,j}; i,j = 1,\dots,d)$ such as

$$I(\theta)_{i,j} = 4\int_{\mathcal{X}} \eta_{\theta,i}(x)\eta_{\theta,j}(x)\mu(dx), \tag{1.2}$$
$$I_{2|1}(\theta)_{i,j} = 4\int_{\mathcal{X}\times\mathcal{Y}} \eta^{2|1}_{x,\theta,i}(y)\eta^{2|1}_{x,\theta,j}(y)\mu^{1,2}(dxdy),$$

where $\eta_\theta(x) = (\eta_{\theta,1}(x),\dots,\eta_{\theta,d}(x))^T$ and $\eta^{2|1}_{x,\theta}(y) = (\eta^{2|1}_{x,\theta,1}(y),\dots,\eta^{2|1}_{x,\theta,d}(y))^T$.

The following result is due to [15].

Lemma 1.4 (Hajek). *Let Assumption 1.4.1 be satisfied, and for μ-almost all x, $p_\theta(x)$ be continuously differentiable around θ_0. If $p'_\theta(x)$ exists and $p_\theta(x) > 0$, let $\eta_\theta(x) = p'_\theta(x)/2p_\theta(x)^{1/2}$, and $\eta_\theta(x) = 0$ otherwise. Assume the Fisher information matrix (1.2) exists and continuous around θ_0. Then the first condition of Assumption 1.4.3 follows.*

The following lemma is a simple modification of the above lemma.

Lemma 1.5. *Let Assumption 1.4.1 be satisfied, and for $\mu^{1,2}$-almost all (x,y), let $p^{1,2}_{\theta_0,\theta}(x,y)$ be continuously differentiable around $\theta = \theta_0$. If $p^{1,2}_{\theta_0,\theta}(x,y)'$*

exists and $p^{1,2}_{\theta_0,\theta}(x,y) > 0$, *let* $\eta^{2|1}_{x,\theta_0,\theta}(y) = (p^{1,2}_{\theta_0,\theta}(x,y))'/2(p^{1,2}_{\theta_0,\theta}(x,y))^{1/2}$, *and let* $\eta^{2|1}_{x,\theta_0,\theta}(y) = 0$ *otherwise. Let* $\eta^{2|1}_{x,\theta_0} = \eta^{2|1}_{x,\theta_0,\theta_0}$. *Assume for* $\mu^{1,2}$*-almost all* (x,y), *the Fisher information matrix* $I_{2|1}(\theta_0,\theta) = (I_{2|1}(\theta_0,\theta)_{i,j}; i,j = 1,\dots,d)$ *such that*

$$4\int_{\mathcal{X}\times\mathcal{Y}} \eta^{2|1}_{x,\theta_0,\theta,i}(y)\eta^{2|1}_{x,\theta_0,\theta,j}(y)\mu^{1,2}(dxdy)$$

exists and continuous around $\theta = \theta_0$, *where* $\eta^{2|1}_{x,\theta_0,\theta} = (\eta^{2|1}_{x,\theta_0,\theta,1},\dots,\eta^{2|1}_{x,\theta_0,\theta,k})$. *Then the second condition of Assumption 1.4.3 follows.*

Assumption 1.4.4. *There exist some* $\epsilon > 0$, *and some* $M \in L^2(P_{\theta_0})$ *such that for any* $s,t,u,v \in B_\epsilon(\theta_0)$, *we have*

$$|Q_{s,u}(x) - Q_{t,v}(x)| \le M(x)(|s-t|^2 + |u-v|^2)^{1/2}.$$

In the following lemma, we use

$$Q^{2|1}_{s,t}(x) = \int_{\mathcal{Y}} \log\frac{p^{2|1}_{x,t}(y)}{p^{2|1}_{x,s}(y)} dP^{2|1}_{x,s}(dy).$$

We also define

$$Q_{x^{(n)},g,h,n} = \sum_{i=1}^n Q_{\theta_0+gn^{-1/2},\theta_0+hn^{-1/2}}(x_i),$$
$$Q^{2|1}_{x^{(n)},g,h,n} = \sum_{i=1}^n Q^{2|1}_{\theta_0+gn^{-1/2},\theta_0+hn^{-1/2}}(x_i),$$

where $x^{(n)} = (x_1,\dots,x_n) \in \mathcal{X}_n$.

Lemma 1.6. *Let Assumptions 1.4.1, 1.4.2 and the second condition of Assumption 1.4.3 be satisfied. Then for any* $g,h \in \mathbf{R}^d$,

$$\sum_{i=1}^n \int \left(\sqrt{p^{2|1}_{x_i,\theta_0+gn^{-1/2}}(y)} - \sqrt{p^{2|1}_{x_i,\theta_0+hn^{-1/2}}(y)}\right)^2 \mu^{2|1}_{x_i}(dy) \tag{1.3}$$

tends in $P_{0,n}$*-probability to* $\langle g-h, I_{2|1}(\theta_0)(g-h)\rangle/4$.

Proof. If $p_{\theta_0}(x) \ne 0$, then let

$$s^{(1)}_n(x,y) = \sqrt{p^{2|1}_{x_i,\theta_0+gn^{-1/2}}(y)} - \sqrt{p^{2|1}_{x_i,\theta_0+hn^{-1/2}}(y)} - n^{-1/2}\langle g-h, \frac{\eta^{2|1}_{x,\theta_0}(y)}{p^{1/2}_{\theta_0}(x)}\rangle,$$

and $s_n^{(2)}(x,y) = n^{-1/2}\langle g-h, \eta^{2|1}_{x,\theta_0}(y)/p^{1/2}_{\theta_0}(x)\rangle$. If $p_{\theta_0}(x) = 0$, then let $s_n^{(1)} = s_n^{(2)} = 0$. Then (1.3) is $P_{0,n}$-almost surely

$$\sum_{i=1}^{n}\int s_n^{(1)}(x_i,y)^2\mu^{2|1}_{x_i}(dy) - 2\int s_n^{(1)}(x_i,y)s_n^{(2)}(x_i,y)\mu^{2|1}_{x_i}(dy) + \int s_n^{(2)}(x_i,y)^2\mu^{2|1}_{x_i}(dy).$$

The first term tends in $P_{0,n}$-probability to 0 by the second condition of Assumption 1.4.3. The second term also tends to 0 by the Schwarz inequality. The last term tends to

$$\int \langle g-h, \eta^{2|1}_{x,\theta_0}(y)\rangle^2 \mu^{1,2}(dxdy)$$

in $P_{0,n}$-almost surely by the law of large numbers, and it is equal to $\langle g-h, I_{2|1}(\theta_0)(g-h)\rangle/4$ by the second condition of Assumption 1.4.3. □

Proposition 1.1. *Let Assumptions 1.4.1-1.4.3 be satisfied. Then, as* $n \to \infty$, $Q_{x,g,h,n} = \overline{Q}_{x,g,h,n} + o_{P_{0,n}}(1)$, *where* $Z_n(x) = n^{-1/2}\sum_{i=1}^{n}\eta_{\theta_0}(x_i)/p^{1/2}_{\theta_0}(x_i)$ *if* $p^{1/2}_{\theta_0}(x_i) \neq 0$ $(i = 1,\ldots,n)$ *and* $Z_n(x) = 0$ *otherwise. Moreover, under* $P_{0,n}$, $Z_n \Rightarrow N(0, I(\theta_0))$, *and* $\lim_{n\to\infty} P_{0,n}(Q_{x,g,h,n}) - P_{0,n}(\overline{Q}_{x,g,h,n}) = 0$.

Proof. By definition, $Q_{x,g,h,n} = Q^{2|1}_{x,g,h,n} + (\log L_{h,n}(x) - \log L_{g,n}(x))$, where $L_{h,n}$ is the likelihood ratio, that is, $\log L_{h,n}(x) = \sum_{i=1}^{n}\log(p_{\theta_0+hn^{-1/2}}(x_i)/p_{\theta_0}(x_i))$. By the expansion of the likelihood ratio, such as Theorem 12.2.3 of [21], the second term is

$$\log\frac{L_{h,n}}{L_{g,n}} = \langle h-g, Z_n\rangle - \frac{1}{2}\langle h, I(\theta_0)h\rangle + \frac{1}{2}\langle g, I(\theta_0)g\rangle + o_{P_{0,n}}(1)$$

and Z_n tends to $N(0, I(\theta_0))$ in $P_{0,n}$-distribution. We consider an expansion of $Q^{2|1}_{x,g,h,n}$. Let

$$t_{g,h,n}(x,y) = \frac{\sqrt{p^{2|1}_{x,\theta_0+hn^{-1/2}}(y)}}{\sqrt{p^{2|1}_{x,\theta_0+gn^{-1/2}}(y)}} - 1,$$

then $\log(p^{2|1}_{x,\theta_0+hn^{-1/2}}(y)/p^{2|1}_{x,\theta_0+gn^{-1/2}}(y)) = 2\log(t_{g,h,n}(x,y)+1)$. Using a Taylor expansion, we have

$$\log(y+1) = y - \frac{1}{2}y^2 + y^2 r(y)$$

where r is $0 \leq r(z) \leq 1/2$ and $r(z) \to 0$ as $z \to 0$. Therefore, $Q^{2|1}_{x,g,h,n}$ is

$$2(\sum_{i=1}^{n} s^{(1)}_{g,h,n}(x_i) - \frac{1}{2}s^{(2)}_{g,h,n}(x_i) + s^{(3)}_{g,h,n}(x_i)),$$

where

$$\begin{aligned} s^{(1)}_{g,h,n}(x) &= \int t_{g,h,n}(x,y) P^{2|1}_{x,\theta_0+gn^{-1/2}}(dy) \\ s^{(2)}_{g,h,n}(x) &= \int t^2_{g,h,n}(x,y) P^{2|1}_{x,\theta_0+gn^{-1/2}}(dy) \\ s^{(3)}_{g,h,n}(x) &= \int r(t_{g,h,n}(x,y)) t^2_{g,h,n}(x,y) P^{2|1}_{x,\theta_0+gn^{-1/2}}(dy). \end{aligned}$$

By Lemma 1.6, we have

$$-2\sum_{i=1}^{n} s^{(1)}_{g,h,n}(x_i) = \sum_{i=1}^{n} s^{(2)}_{g,h,n}(x_i) \to \frac{1}{4}\langle h-g, I_{2|1}(\theta_0)(h-g)\rangle$$

in $P_{0,n}$-probability. We show that $\sum_{i=1}^{n} s^{(3)}_{g,h,n}(x_i)$ also in $P_{0,n}$-probability to 0. The integral $\int_{\mathcal{X}_n} P_{0,n}(dx)|\sum_{i=1}^{n} s^{(3)}_{g,h,n}(x_i)|$ is bounded above by

$$n\int_{\mathcal{X}\times\mathcal{Y}} r(t_{g,h,n}(x,y))\Big(\sqrt{p^{1,2}_{\theta_0,\theta_0+hn^{-1/2}}(x,y)} - \sqrt{p^{1,2}_{\theta_0,\theta_0+gn^{-1/2}}(x,y)}\Big)^2 \mu^{1,2}(dxdy),$$

and the integrand is also bounded above by the sum of the following three terms:

$$\begin{aligned} &\frac{n}{2}\Big(\sqrt{p^{1,2}_{\theta_0,\theta_0+hn^{-1/2}}(x,y)} - \sqrt{p^{1,2}_{\theta_0,\theta_0+gn^{-1/2}}(x,y)} - n^{-1/2}\langle h-g, \eta^{2|1}_{x,\theta_0}(y)\rangle\Big)^2, \\ &\frac{n^{1/2}}{2}\Big(\sqrt{p^{1,2}_{\theta_0,\theta_0+hn^{-1/2}}(x,y)} - \sqrt{p^{1,2}_{\theta_0,\theta_0+gn^{-1/2}}(x,y)} - n^{-1/2}\langle h-g, \eta^{2|1}_{x,\theta_0}(y)\rangle\Big) \\ &\times \langle h-g, \eta^{2|1}_{x,\theta_0}(y)\rangle, \\ &r(t_{g,h,n}(x,y))\langle h-g, \eta^{2|1}_{x,\theta_0}(y)\rangle^2. \end{aligned}$$

The integral of the first term tends to 0 by Assumption 1.4.3, and the second term also tends to 0 by the the Schwarz inequality. The convergence of the last term is by the Lebesgue's dominated convergence theorem since $r(t_{g,h,n}(x,y))$ tends in $P^{1,2}_{\theta_0,\theta_0+hn^{-1/2}}$-probability, hence in $P^{1,2}_{\theta_0}$-probability to 0. Therefore, $Q_{x,g,h,n} - \log L_{h,n}/L_{g,n} \to -\langle (h-g), I_{2|1}(\theta_0)(h-g)\rangle/2$ in $P_{0,n}$-probability and hence the first claim follows.

The second claim about Z_n is obvious. We consider $P_{0,n}(Q_{x,g,h,n} - \log L_{h,n}/L_{g,n}) \to -\langle (h-g), I_{2|1}(\theta_0)(h-g)\rangle/2$. This proof is almost identical, and we omit the detail. The convergence $P_{0,n}(\log L_{h,n}/L_{g,n}) \to -\langle h, I(\theta_0)h\rangle/2 + \langle g, I(\theta_0)g\rangle/2$ is obtained by a likelihood expansion theory, hence the third claim follows, since $P_{0,n}(Z_n) = 0$. □

Next we show that under Assumptions 1.4.1-1.4.4, Assumption 1.3.2 holds. We use the following maximal inequality to prove the fact. The maximal inequality was studied for example, in [32], and this version of the following lemma is from [41].

Consider the space $L^2(\mathcal{Z}, \mathcal{C}, P) = L^2(P)$ with the L^2-norm $\|f\|_{L^2} = (\int_{\mathcal{Z}} f(z)^2 P(dz))^{1/2}$ ($f \in L^2(P)$). The bracket $[f,g]$ is a subset of $L^2(P)$ which is defined as $[f,g] = \{h \in L^2(P); f(z) \leq h(z) \leq g(z)\}$. We call the bracket $[f,g]$, ϵ-bracket if $\|f-g\|_{L^2} \leq \epsilon$. For any subset $\mathcal{F}$ of $L^2(P)$, and for any $\delta > 0$, the bracket number $N_{[\ ]}(\delta, \mathcal{F})$ is the smallest number of δ-brackets needed to cover $\mathcal{F}$. The bracket integral $J_{[\ ]}(\delta, \mathcal{F})$ is defined as

$$J_{[\ ]}(\delta, \mathcal{F}) = \int_0^\delta (\log N_{[\ ]}(\epsilon, \mathcal{F}))^{1/2} d\epsilon.$$

Lemma 1.7 (Maximal Inequality). *Let $\mathcal{F} \subset L^2(\mathcal{Z}, \mathcal{C}, P)$. Assume that for some $\delta > 0$, there exists $M \in L^2(P)$ such that $|f| \leq \delta M$ ($f \in \mathcal{F}$). Then we have for some $C > 0$,*

$$\begin{aligned} &n^{-1/2} \int_{\mathcal{X}^n} P^n(d(x_1, \dots, x_n)) \sup_{f \in \mathcal{F}} |\sum_{i=1}^n f(x_i) - P(f)| \\ &\leq C(J_{[\ ]}(\delta, \mathcal{F}) + \frac{\|M 1_{\{M > n^{1/2} a(\delta)\}}\|_{L^2}^2}{a(\delta)}), \end{aligned}$$

where P^n and $\mathcal{X}^n$ denote the n product of P and $\mathcal{X}$ and $a(\delta) = \delta/(\log(N_{[\ ]}(\delta, \mathcal{F})) \vee 1)^{1/2}$.

The following proposition is a simple modification of Theorem 5.39 of [41].

Proposition 1.2. *Under Assumptions 1.4.1-1.4.4, Assumption 1.3.2 holds.*

Proof. Fix any compact set set $K \subset \mathbf{R}^{2d}$. Let $Q_{x,t,n} = Q_{x,t_1,t_2,n}$ and $\overline{Q}_{x,t,n} = \overline{Q}_{x,t_1,t_2,n}$ ($t = (t_1, t_2) \in K$). Note that we have

$$Q_{x,t,n} - \overline{Q}_{x,t,n} - P_{0,n}(Q_{x,t,n} - \overline{Q}_{x,t,n}) = o_{P_{0,n}}(1) \qquad (1.4)$$

for any fixed $t \in K$, by Proposition 1.1. In fact, the left hand side of (1.4) tends in $P_{0,n}$-probability to 0 uniformly in $t \in K$. It is enough to show the tightness of the left hand side of (1.4). Since the tightness of $\overline{Q}_{x,\cdot,n} - P_{0,n}(\overline{Q}_{x,\cdot,n})$ is clear, we show that for any $\eta, \epsilon > 0$, there exists δ such that

$$\limsup_{n\to\infty} P_{0,n}(\sup_{|s-t|\leq\delta, s,t\in K} |Q_{x,s,n} - Q_{x,t,n} - P_{0,n}(Q_{x,s,n} - Q_{x,t,n})| > \epsilon) \leq \eta. \tag{1.5}$$

Let $g_{n,t}(x) = n^{1/2}Q_{\theta_0+t_1n^{-1/2},\theta_0+t_2n^{-1/2}}(x)$ for any $x \in \mathcal{X}$ and $t = (t_1, t_2) \in K$. Let $\mathcal{F}_n^\delta = \{g_{n,s} - g_{n,t}; t, s \in K, |t - s| \leq \delta\}$. Note that for any $f \in \mathcal{F}_n^\delta$, we have $|f| \leq \delta M$. By Example 19.7 of [41], we have

$$N_{[\,]}(\delta\epsilon\|M\|_{L^2}, \mathcal{F}_n^\delta) \leq C(\delta\epsilon^{-1})^{4d} \ (0 < \epsilon < \delta).$$

Therefore, we have

$$\begin{aligned} J_{[\,]}(\delta_1, \mathcal{F}_n^\delta) &= \delta\|M\|_{L^2} \int_0^{\frac{\delta_1}{\delta\|M\|_{L^2}}} \sqrt{N_{[\,]}(\delta\epsilon\|M\|_{L^2}, \mathcal{F}_n^\delta)} d\epsilon \\ &\leq \delta\|M\|_{L^2} \int_0^{\frac{\delta_1}{\delta\|M\|_{L^2}}} \sqrt{-4d\log\epsilon + \log C\delta^{4d}} d\epsilon \end{aligned}$$

and the right hand side is bounded above. For any $\epsilon_1 > 0$, we can choose δ_1 such as $J_{[\,]}(\delta_1, \mathcal{F}_n^\delta) \leq \epsilon_1$.

By Lemma 1.7, we have

$$\begin{aligned} & P_{0,n}(\sup_{\mathcal{F}_n^\delta} n^{-1/2}|\sum_{i=1}^n f(x_i) - P_{\theta_0}(f)| > \epsilon) \\ & \leq \epsilon^{-1}n^{-1/2} \int_{\mathcal{X}^n} P_{0,n}(d(x_1, x_2, \ldots, x_n)) \sup_{\mathcal{F}_n^\delta} |\sum_{i=1}^n f(x_i) - P_{\theta_0}(f)| \\ & \leq \epsilon^{-1}C(J_{[\,]}(\delta_1, \mathcal{F}_n^\delta) + \frac{\|M1_{\{M>n^{1/2}a(\delta_1)\}}\|_{L^2}^2}{a(\delta_1)}). \end{aligned}$$

The second term tends to 0 in the limit $n \to \infty$ and the first term is bounded above by $C\epsilon^{-1}\epsilon_1$. The real number ϵ_1 is arbitrary, hence the above value tends to 0, and (1.5) holds. □

Chapter 2

Asymptotic Properties for the Gibbs Sampler

2.1 Introduction

In this chapter, we address some asymptotic properties for the Gibbs sampler. The main results in this chapter are convergence theorem, validation of a speed up method and its application to independent and identically distributed observations.

The convergence property for finite sample size has already been established, which is based on the ergodicity of Markov chain (see [39], [23]). The Markov chain made by the Gibbs sampler is Harris recurrent under fairly general assumptions, and moreover, the chain is sometimes geometrically ergodic. This convergence property is concerned with the behavior of the Gibbs sampler in the region far from the true parameter of the parameter space. This meaning of convergence can be considered to be a *global* convergence.

On the other hand, in the present paper, we address some properties of the behavior of the Gibbs sampler around the true parameter. This convergence property is a *local* convergence property. Therefore, these two kinds of convergence are different properties for the Gibbs sampler and both of which are useful.

The relative merits of the local convergence are as follows. In the present study, we approximate the Gibbs sampler by a simple transition kernel, which is defined by the score statistic and matrices which are related to the Fisher information matrix. Therefore, we can measure the convergence rate of the Gibbs sampler by simple statistics, and we can compare different kinds of Gibbs samplers by these simple transition kernels.

The most important assumption of our results is that, the initial guess of the sequence of the Gibbs sampler is $\theta_0 \sim \nu_{x,n}$ such that $\nu_{x,n}(B_{M_n}(\theta_0))$ tends in probability to 1 when $n^{-1/2}M_n \to \infty$ for some $M_n \to \infty$.

In this chapter, both probability measure and transition kernel may depend on the observations. We will assume the following measurability conditions for those. Let $(\mathcal{Z}, \mathcal{C})$ and $(T, \mathcal{T})$ be measurable spaces. In this chapter, we assume that any family of probability measures $\nu = (\nu_z(S), z \in \mathcal{Z}, S \in \mathcal{T})$ satisfies the following:

1. For any $S \in \mathcal{T}$, $\nu_\cdot(S) : \mathcal{Z} \to [0,1]$ is $\mathcal{C}$-measurable.
2. For any $z \in \mathcal{Z}$, ν_z is a probability measure on $(T, \mathcal{T})$.

A probability transition kernel $K = (K_{z,s}(S); z \in \mathcal{Z}, s \in T, S \in \mathcal{T})$ is a function such that:

1. For any $S \in \mathcal{T}$, $K_{\cdot,\cdot}(S) : \mathcal{Z} \times T \to [0,1]$ is $\mathcal{C} \times \mathcal{T}$-measurable.
2. For any $z \in \mathcal{Z}$ and $s \in T$, $K_{z,s}$ is a probability measure on $(T, \mathcal{T})$.

We use the following notation about the kernel. If K has an invariant probability distribution, we write it as K_z. Fix $z \in \mathcal{Z}$. Let $s_0, s_1, \ldots$ be a sequence of the Markov chain from $K_{z,\cdot}(\cdot)$. Then we write the joint distribution of $(s_k, \ldots, s_l)$ depending on the distribution of s_0 as follows:

$$\begin{aligned} &K_{z,\nu}^{(k:l)} \text{ if } s_0 \sim \nu_z(dt), \\ &K_{z,s}^{(k:l)} \text{ if } s_0 \sim \delta_s(dt), \\ &K_z^{(k:l)} \text{ if } s_0 \sim K_z. \end{aligned}$$

If $k = l$, we write them as $K_{z,\nu}^k$, $K_{z,s}^k$ and K_z^k respectively. If K and ν does not depend on z, we drop z from the above symbols. Note that we have $K_{z,\nu}^0 = \nu_z$ by definition.

2.2 Regular Statistical Experiments

Suppose Θ is an open subset of $\mathbf{R}^d$, and $\theta_0 \in \Theta$. The parameter space Θ is equipped with its Borel σ-algebra $\mathcal{F}$. The element θ_0 will be used as the true value of the following model. Consider a family of statistical experiments $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{h,n}; h \in H_n)$, where $H_n = n^{1/2}(\Theta - \theta_0)$. We set $\mathcal{H}_n = n^{1/2}(\mathcal{F} - \theta_0)$.

We also consider other two families of statistical experiments $\mathcal{E}_n^{2|1} = (\mathcal{Y}_n, \mathcal{B}_n, P_{x,h,n}^{2|1}; x \in \mathcal{X}_n, h \in H_n)$ and $\mathcal{E}_n^{1,2} = (\mathcal{X}_n \times \mathcal{Y}_n, \mathcal{A}_n \times \mathcal{B}_n, P_{g,h,n}^{1,2}; g, h \in H_n)$, where $P_{g,h,n}^{1,2}(dx, dy) = P_{g,n}(dx)P_{x,h,n}^{2|1}(dy)$. Let $P_{h,n}^{1,2}$ denote $P_{h,h,n}^{1,2}$.

Let $(F_{x,n}; x \in \mathcal{X}_n)$, and $(F_{x,y,n}^{1|2}; x \in \mathcal{X}_n, y \in \mathcal{Y}_n)$ be families of probability measures such that each $F_{x,n}$ and $F_{x,y,n}^{1|2}$ are probability measures on $(H_n, \mathcal{H}_n)$. In Section 2.3, these families will be families of posterior distributions.

We assume the following measurability condition.

Assumption 2.2.1. *For any $n \in \mathbf{N}$ and $B \in \mathcal{B}_n$, $P_{\cdot,\cdot,n}^{2|1}(B)$ is $\mathcal{A}_n \times \mathcal{H}_n$-measurable. For any $H \in \mathcal{H}_n$, $F_{\cdot,n}(H)$ is $\mathcal{A}_n$-measurable, and $F_{\cdot,\cdot,n}^{1|2}(H)$ is $\mathcal{A}_n \times \mathcal{B}_n$-measurable.*

2.2.1 Two-Stage Gibbs Sampler

We define the simplest Gibbs sampler. Assume we have an observation x from $P_{0,n}$.

Step 0 Set $h_0 \in H_n$, then, go to Step 1.

Step i Generate y_i from $P_{h_{i-1},x,n}^{2|1}$. Then, generate h_i from $F_{x,y_i,n}^{1|2}$ and go to Step $i+1$.

This procedure defines a Markov chain $h_0, h_1, h_2, \ldots$. Let $F_{x,\cdot,n}(\cdot) = F_{0,x,\cdot,n}(\cdot) = (F_{x,h,n}(A); h \in H_n, A \in \mathcal{H}_n)$ denote its transition kernel.

Let $I(\theta_0)$ and $I_{2|1}(\theta_0)$ be $d \times d$- matrices and $(Z_n : \mathcal{X}_n \to \mathbf{R}^d; n \in \mathbf{N})$ be a sequence of random variables. Let $I_{1,2}(\theta_0) = I(\theta_0) + I_{2|1}(\theta_0)$. We assume the following condition.

Assumption 2.2.2. *The matrices $I(\theta_0)$ and $I_{2|1}(\theta_0)$ are positive definite and $(Z_n; n \in \mathbf{N})$ is tight with respect to $P_{0,n}$.*

We are going to show that the transition kernel $F_{x,\cdot,n}$ tends to a simple transition kernel of the following Markov chain $h_0, h_1, h_2, \ldots$.

Step 0 Set $h_0 \in \mathbf{R}^d$, then, go to Step 1.

Step i Generate g_i from $G_{Z_n,h_{i-1}}^{2|1}$ where

$$G_{z,h}^{2|1} = N(I_{2|1}(\theta_0)h + z, I_{2|1}(\theta_0)).$$

Then, generate h_i from $G_{g_i}^{1|2} = N(I_{1,2}(\theta_0)^{-1}g_i, I_{1,2}(\theta_0)^{-1})$.

This procedure defines an AR process such that

$$h_i - \mu_n = J(\theta_0)(h_{i-1} - \mu_n) + \epsilon_n,$$

where $\mu_n = I(\theta_0)^{-1}Z_n$, $J(\theta_0) = I_{1,2}(\theta_0)^{-1}I_{2|1}(\theta_0)$ and $(\epsilon_i; i \in \mathbf{N})$ is a sequence random variables of independent and identically distributed as $N(0, I_{1,2}(\theta_0)^{-1}I_{2|1}(\theta_0)I_{1,2}(\theta_0)^{-1} + I_{1,2}(\theta_0)^{-1})$.

Let $G_{x,\cdot,n}(\cdot) = G_{0,x,\cdot,n}(\cdot) = (G_{x,h,n}(A); h \in \mathbf{R}^d, A \in \mathcal{B}(\mathbf{R}^d))$ denote the transition kernel, which is

$$G_{x,h,n} = N(I_{1,2}(\theta_0)^{-1}(I_{2|1}(\theta_0)h + Z_n), I_{1,2}(\theta_0)^{-1}I_{2|1}(\theta_0)I_{1,2}(\theta_0)^{-1} + I_{1,2}(\theta_0)^{-1}).$$

For any probability measures P, Q on $(\mathcal{Z}, \mathcal{C})$, the Hellinger distance $H(P, Q)$ is defined as

$$H(P,Q)^2 = \frac{1}{2}\int_{z\in\mathcal{Z}} (\sqrt{p(z)} - \sqrt{q(z)})^2 \mu(dz),$$

where μ is a σ-finite measure on $(\mathcal{Z}, \mathcal{C})$ such that $P \ll \mu, Q \ll \mu$ and $p = dP/d\mu$, $q = dQ/d\mu$. Note that the Hellinger distance is smaller than 1, and

$$\frac{1}{2}\|P - Q\|_{\mathrm{TV}} \le H(P,Q)(2 - H(P,Q)^2)^{1/2}.$$

Lemma 2.1. *Let a, c be d-dimensional vectors, and B, D be positive definite $d \times d$-matrices. Let $P = N(a, B)$ and $Q = N(c, D)$. Then*

$$\begin{aligned} H(P,Q)^2 =& 1 - \det((B+D)/2)^{-1/2}\det(B)^{1/4}\det(D)^{1/4} \\ & \times \exp(-\frac{1}{4}\langle a - c, (D+B)^{-1}(a-c)\rangle). \end{aligned}$$

Proof. By definition, we have

$$\begin{aligned} H(P,Q)^2 &= 1 - \int_{\mathbf{R}^d} \phi(x; a, B)^{1/2}\phi(x; c, D)^{1/2}dx \\ &= 1 - \int_{\mathbf{R}^d} \frac{\det(BD)^{-1/4}}{(2\pi)^{d/2}} e^{-\left(\langle x-a, B^{-1}(x-a)\rangle + \langle x-c, D^{-1}(x-c)\rangle\right)/4} dx \\ &= 1 - \det(BD)^{-1/4}\det((B^{-1} + D^{-1})/2)^{-1/2}e^{R/4}, \end{aligned}$$

where R is

$$\langle B^{-1}a + D^{-1}c, (B^{-1} + D^{-1})^{-1}(B^{-1}a + D^{-1}c)\rangle - \langle a, B^{-1}a\rangle - \langle c, D^{-1}c\rangle.$$

Taking $E = B^{-1} + D^{-1}$, the first term is

$$\begin{aligned}&\langle Ea + D^{-1}(c-a), E^{-1}(B^{-1}(a-c) + Ec)\rangle \\ &= -\langle a-c, D^{-1}E^{-1}B^{-1}(a-c)\rangle + \langle D^{-1}(c-a), c\rangle + \langle a, B^{-1}a + D^{-1}c\rangle.\end{aligned}$$

Since $D^{-1}E^{-1}B^{-1} = (D+B)^{-1}$, we have $R = -\langle a-c, (D+B)^{-1}(a-c)\rangle$ and hence the claim follows. □

Lemma 2.2. *Let X, Y, Z be $d \times d$-matrices. Assume X and Z are positive definite matrices. Let $\Pi = N(0, X)$ and let $V : \mathbf{R}^d \times \mathcal{B}(\mathbf{R}^d) \to [0,1]$ be a transition kernel such that $V_h = N(Yh, Z)$. Then*

(a) Π is the invariant probability measure for V if and only if $YXY^T + Z = X$. In particular, $X - YXY^T$ is positive definite.

(b) Assume Π is the invariant probability measure for V, and there exists a constant $r < 1$ such that there exist non-singular $d \times d$-matrix P and $d \times d$-matrix Λ satisfying $Y = P^{-1}\Lambda P$ and $\|\Lambda\|_2 = r$. Then there exist constants $c_1, c_2 > 0$ such that

$$\|\Pi - V_h^i\|_{\mathrm{TV}}^2 \le (c_1 + c_2|h|^2)r^{2i} \ (h \in \mathbf{R}^d),$$

where V_h^i is defined by $V_h^i(\cdot) = \int_g V_h^{i-1}(dg)V_g(\cdot)$ for $i \ge 1$ and $V_h^0 = \delta_h$.

Proof. The distribution $\int_{\mathbf{R}^d} \Pi(dh)V_h$ is $N(0, YXY^T + Z)$. Then the first claim is apparent.

We show the second claim. Let $a = 0$, $B = X$, $c = Y^i h$, and $D = X - Y^i X (Y^i)^T$. Then $\Pi = N(a, B)$ and $V_h^i = N(c, D)$. We have

$$\begin{aligned}&2^{-3}\|\Pi - V_h^i\|_{\mathrm{TV}}^2 \le H(\Pi, V_h^i)^2 \le 1 - \det(B^{-1/2}DB^{-1/2})^{1/4}\exp(-\frac{1}{4}c'(D+B)^{-1}c) \\ &\le 1 - \det(B^{-1/2}DB^{-1/2})^{1/4} + 1 - \exp(-\frac{1}{4}\langle c, (D+B)^{-1}c\rangle) \\ &\le 1 - \det(B^{-1/2}DB^{-1/2})^{1/4} + \frac{1}{4}\langle c, (D+B)^{-1}c\rangle,\end{aligned}$$

since $\det((B+D)/2) \le \det(B)$. We calculate the upper bounds of $1 - \det(B^{-1/2}DB^{-1/2})^{1/4}$ and $\langle c, (D+B)^{-1}c\rangle$.

The matrix $B^{-1/2}DB^{-1/2} = I - X^{-1/2}Y^iX(Y^i)^TX^{-1/2}$ is positive definite, and if λ is its eigenvalue and u is its eigenvector, then

$$\begin{aligned}|(1-\lambda)u| &= |X^{-1/2}Y^iX(Y^i)^TX^{-1/2}u| \\ &= |X^{-1/2}P^{-1}\Lambda^iPX(P^{-1}\Lambda^iP)^TX^{-1/2}u| \\ &\le \|X^{-1/2}\|_2^2\|P\|_2^2\|P^{-1}\|_2^2\|X\|_2\|\Lambda\|_2^{2i}|u|.\end{aligned}$$

Therefore, there exists a constant $c > 0$ such that $|1-\lambda| \leq cr^{2i}$. Let $n \in \mathbf{N}$ to be $cr^{2i} < 1$ and let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $B^{-1/2}DB^{-1/2}$. Then for any $i \geq n$, we have

$$0 \leq 1-\det(B^{-1/2}DB^{-1/2})^{1/4} = 1-\prod_{j=1}^{d}\lambda_j^{1/4} \leq 1-(1-cr^{2i})^{d/4} \leq c^* r^{2i}$$

for some $c^* > 0$. Then there exists $c_1 > 0$ such that $1-\det(B^{-1/2}DB^{-1/2})^{1/4} \leq 2^{-3}c_1 r^{2i}$ for any $i \in \mathbf{N}$.

Let $D_1 = X - YXY^T$. Then D_1 is also a positive definite matrix, and we have $(D+B)-(D_1+B) = Y(X - Y^{i-1}X(Y^{i-1})^T)Y^T$ is nonnegative definite. Therefore, $\|(D+B)^{-1}\|_2 \leq \|(D_1+B)^{-1}\|_2$, and we have

$$\begin{aligned}&\langle c,(D+B)^{-1}c\rangle = \langle Y^i h,(D+B)^{-1}Y^i h\rangle = \langle P^{-1}\Lambda^i Ph,(D+B)^{-1}P^{-1}\Lambda^i Ph\rangle\\ &\leq \|P^{-1}\|_2^2\|P\|_2^2\|(D_1+B)^{-1}\|_2|h|^2 r^{2i} \leq 2^{-3}c_2|h|^2 r^{2i}\end{aligned}$$

for $c_2 = 2^3\|P^{-1}\|_2^2\|P\|_2^2\|(2X-YXY^T)^{-1}\|_2 > 0$. Hence the claim follows. □

Let $G_{x,n} = N(I(\theta_0)^{-1}Z_n, I(\theta_0)^{-1})$.

Corollary 2.1. *The distribution $G_{x,n}$ is the invariant distribution of $G_{x,\cdot,n}$.*

Proof. Let we denote $A = I(\theta_0)$, $B = I_{2|1}(\theta_0)$ and $C = I_{1,2}(\theta_0)$. Since $C = A+B$, we have $C^{-1}BA^{-1}+C^{-1} = A^{-1}$, and therefore,

$$\begin{aligned}&C^{-1}BA^{-1}BC^{-1}+C^{-1}BC^{-1}+C^{-1} = (C^{-1}BA^{-1}+C^{-1})BC^{-1}+C^{-1}\\ &= A^{-1}BC^{-1}+C^{-1} = A^{-1}.\end{aligned}$$

Hence for $X = A^{-1}$, $Y = C^{-1}B$ and $Z = C^{-1}BC^{-1}+C^{-1}$, we have $YXY^T + Z = X$. □

We assume the following conditions.

Assumption 2.2.3. *For any $h \in \mathbf{R}^d$, $P^{1,2}_{0,h,n}$ and $P^{1,2}_{0,n}$ are mutually contiguous.*

Assumption 2.2.4. *For any bounded continuous function $f : \mathbf{R}^d \to \mathbf{R}$, a sequence of random variables $(Z_n^{2|1} : \mathcal{X}_n \times \mathcal{Y}_n \to \mathbf{R}^d; n \in \mathbf{N})$ satisfies the following property:*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\left|\int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy)f(Z_n^{2|1}(x,y)) - \int_{\mathbf{R}^d} G^{2|1}_{0,h}(dw)f(w)\right| \to 0. \quad (2.1)$$

Let $Z_n^{1,2} = Z_n + Z_n^{2|1}$.

Assumption 2.2.5.

$$\lim_{n\to\infty}\int_{\mathcal{X}_n\times\mathcal{Y}_n} P_{0,n}^{1,2}(dxdy)\|F_{x,y,n}^{1|2} - G_{Z_n^{1,2}}^{1|2}\|_{\mathrm{TV}} = 0.$$

Assumption 2.2.6. *For $P_{0,n}$-almost all x, the transition kernel $F_{x,\cdot,n}$ has an invariant probability distribution $F_{x,n}$ and that*

$$\lim_{n\to\infty}\int_{\mathcal{X}_n} P_{0,n}(dx)\|F_{x,n} - G_{x,n}\|_{\mathrm{TV}} = 0.$$

Assumption 2.2.7. *There exists $\delta > 0$ such that for any $\epsilon_n \to 0$,*

$$\lim_{n\to\infty}\int_{\mathcal{X}_n} P_{0,n}(dx) \sup_{|h-g|<\epsilon_n, |h|,|g|\le\delta n^{1/2}} \|P_{x,h,n}^{2|1} - P_{x,g,n}^{2|1}\|_{\mathrm{TV}} \to 0.$$

We note a few comments for the above assumptions. Assumption 2.2.4 is the asymptotic normality condition of $\mathcal{L}(Z_n^{2|1}|P_{x,h,n}^{2|1})$ and in particular, $(Z_n^{2|1}; n \in \mathbf{N})$ is $P_{0,n}^{1,2}$-tight. Assumptions 2.2.5 and 2.2.6 are related to the Bernstein von-Mises theorem for $\mathcal{E}_n$ and $\mathcal{E}_n^{1,2}$. In most applications of the Gibbs sampler, $\mathcal{E}_n$ and $\mathcal{E}_n^{1,2}$ are experiments of independent and identically distributed (i.i.d.) observations or experiments of time discrete Markov chains. For the case of i.i.d. observations, the sufficient conditions of Assumptions 2.2.5 and 2.2.6 are studied for example, in [1], [42]. For the case of Markov chains, those conditions are studied for example, in [3]. See also [16], [41], [6] and [13]. In a later section, we treat some sufficient conditions for the above assumptions for i.i.d. case.

Let $(T, \mathcal{T})$ be a measurable space, and Ψ be any index set. Let $(\|\cdot\|_{u_{1:k}}^*; k \in \mathbf{N}, u_i \in \Psi)$ denote a family of semi-norms satisfying the following conditions for any $k, l \in \mathbf{N}$ and $u_1, \ldots, u_{k+l} \in \Psi$.

1. A real valued function $\|\cdot\|_{u_{1:k}}^*$ is a semi-norm on the linear space

   $$\mathcal{S}_k = \{\nu; \nu \text{ is a signed measure on } (T^k, \mathcal{T}^k) \text{ with } \|\nu\|_{\mathrm{T.V}} < \infty\},$$

   that is, for any $\nu, \mu \in \mathcal{S}_k$ and $\alpha \in \mathbf{R}$, we have

   $$\|\nu + \mu\|_{u_{1:k}}^* \le \|\nu\|_{u_{1:k}}^* + \|\mu\|_{u_{1:k}}^*,$$
   $$\|\alpha\nu\|_{u_{1:k}}^* \le |\alpha|\|\nu\|_{u_{1:k}}^*.$$

2. $\|\nu\|^*_{u_{1:k}} \leq \|\nu\|_{\mathrm{TV}}$ for any $\nu \in \mathcal{S}_k$.

3. Let $(\mathcal{Z}, \mathcal{C})$ be a measurable space. If $K = (K_{z,s}(S); z \in \mathcal{Z}, s \in T^k, S \in \mathcal{T}^l)$ satisfies $K_{z,s} \in \mathcal{S}_l$ for any $z \in \mathcal{Z}, s \in T^k$ and $K_{\cdot,\cdot}(S) : \mathcal{Z} \times T^k \to \mathbf{R}$ is $\mathcal{C} \times \mathcal{T}^k$-measurable, then

    (a) $\|K_{\cdot,\cdot}\|^*_{u_{1:l}} : \mathcal{Z} \times T^k \to \mathbf{R}$ is $\mathcal{C} \times \mathcal{T}^k$-measurable.

    (b) for any $\mu \in \mathcal{S}_k$, the following property holds:

$$\|\mu(ds)K_{z,s}(dS)\|^*_{u_{1:k+l}} \leq \|(\mu(ds)\|K_{z,s}\|^*_{u_{k+1:k+l}})\|^*_{u_{1:k}}.$$

Example 2.1. *Let $\|\nu\|^*_{u_{1:k}} = \|\nu\|_{\mathrm{TV}}$ for any $u_1, \ldots, u_k$. Then it satisfies the above conditions if $\mathcal{T}$ is countably generated.*

Example 2.2. *If Ψ is a set of $\mathcal{T}$-measurable functions $u : T \to \mathbf{R}$ such that $|u| \leq 1$, then*

$$\|\nu\|^*_{u_{1:k}} = |\int_{z=(z_1,\ldots,z_k)\in\mathcal{Z}^k} \nu(dz)u_1(z_1)\cdots u_k(z_k)|$$

satisfies the above conditions.

In the following lemma, we also assume for any $s, t \in T$, then $s + t \in T$.

Lemma 2.3. *Let $(\mathcal{Z}_n, \mathcal{C}_n, P_n)$ be a sequence of probability spaces. Let $U_{\cdot,\cdot,n} = (U_{z,s,n}(S); z \in \mathcal{Z}_n, s \in T, S \in \mathcal{T})$ and $V_{\cdot} = (V_s(S); s \in T, S \in \mathcal{T})$ be transition kernels and $\nu_{z,n}$, $\mu_{z,n}$ be probability measures on $(T, \mathcal{T})$. Let $(W_n : \mathcal{Z}_n \to T; n \in \mathbf{N})$ be a sequence of random variables. Suppose for any $l \in \mathbf{N}$, there exist a test $\omega_{n,l} : \mathcal{Z}_n \to [0,1]$ and a finite measure λ_l on $(T, \mathcal{T})$ such that $(1-\omega_{n,l})V^l_{z,\mu,n}(A - W_n) \leq \lambda_l(A)$ $(A \in \mathcal{T})$ and $P_n(\omega_{n,l}) \to 0$. Assume for any $s \in T$, we have*

$$\int_{\mathcal{Z}_n} P_n(dz)\|U_{z,s+W_n,n} - V_{s+W_n}\|^*_u \to 0,$$

$$\int_{\mathcal{Z}_n} P_n(dz)\|\nu_{z,n} - \mu_{z,n}\|^*_u \to 0 \ (u \in \Psi).$$

Then

$$\lim_{n\to\infty} \int P_n(dz)\|U^{(0:k)}_{z,\nu,n} - V^{(0:k)}_{z,\mu,n}\|^*_{u_{0:k}} = 0.$$

Proof. The case $k = 0$ is clear by the assumption, since $U^{(0:0)}_{z,\nu,n} = \nu_{z,n}$ and $V^{(0:0)}_{z,\mu,n} = \mu_{z,n}$. We assume that the case $k = l$ is proved. Then

$$\begin{aligned}
&\|U^{(0:l+1)}_{z,\nu,n} - V^{(0:l+1)}_{z,\mu,n}\|^*_{u_{0:l+1}} \\
&= \|(U^{(0:l)}_{z,\nu,n} - V^{(0:l)}_{z,\mu,n})(ds_{1:l})U_{z,s_l,n}(ds_{l+1}) + V^{(0:l)}_{z,\mu,n}(ds_{1:l})(U_{z,s_l,n} - V_{s_l})(ds_{l+1})\|^*_{u_{0:l+1}} \\
&\leq \|U^{(0:l)}_{z,\nu,n} - V^{(0:l)}_{z,\mu,n}\|^*_{u_{0:l}} + \int V^l_{z,\mu,n}(ds)\|U_{z,s,n} - V_s\|^*_{u_{l+1}} \\
&\leq \|U^{(0:l)}_{z,\nu,n} - V^{(0:l)}_{z,\mu,n}\|^*_{u_{0:l}} + \int \lambda_l(ds)\|U_{z,s+W_n,n} - V_{s+W_n}\|^*_{u_{l+1}} + 2\omega_{n,l},
\end{aligned}$$

and hence the case $k = l + 1$ follows. □

Lemma 2.4. *Let Assumptions 2.2.1, 2.2.2 and 2.2.4 be satisfied. For any bounded continuous function $f : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{C}$ such that $\partial f(x,y)/\partial x$ exists and continuous, the following value tends in $P_{0,n}$-probability to* 0*:*

$$\int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy) f(Z_n(x), Z^{2|1}_n(x,y)) - \int_{\mathbf{R}^d} G^{2|1}_{0,h}(dg) f(Z_n(x), g).$$

Proof. For simplicity, suppose $|f| \leq 1$. Fix $C > 0$ and consider a test $\omega_{n,1} = 1_{\{|Z_n| > C\}}$. We will show that the supremum over $|z| \leq C$ of the following value tends in $P_{0,n}$-probability to 0:

$$|\int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy) f(z, Z^{2|1}_n(x,y)) - \int_{\mathbf{R}^d} G^{2|1}_{0,h}(dg) f(z,g)|. \tag{2.2}$$

By Assumption 2.2.4, the above value tends in $P_{0,n}$-probability to 0 for any fixed z. Choose a sequence of subsets of $B_C(0)$, $C_1 \subset C_2 \subset \ldots$ such that for each $n \in \mathbf{N}$ the number of elements of C_n is finite and $B_C(0) \subset \cup_{z \in C_n} B_{n^{-1}}(z)$. Then there exists $m_n \to \infty$ such that the supremum over C_{m_n} of (2.2) tends in $P_{0,n}$-probability to 0. Taking $\omega_{n,2}(x) = 1_{\{|x| > C\}}$, and let $M = \sup_{x,y \in C^2} |\partial_x f(x,y)|$, then the difference between the supremum over $|z| \leq C$ and the supremum over C_{m_n} of (2.2) is bounded above by

$$2\int P^{2|1}_{x,h,n}(dy)\omega_{n,2}(Z^{2|1}_n(x,y)) + 2\int G^{2|1}_{0,h}(dg)\omega_{n,2}(g) + 2m_n^{-1}M.$$

Then the first and the second term tends to 0 if we let $C = C_n$ tend to ∞. On the other hand, we can choose $C_n \to \infty$ such that the third term tends to 0. Hence, the claim follows. □

Theorem 2.1 (Bernstein-von Mises theorem for Gibbs sampler). *Let Assumptions 2.2.1-2.2.5 be satisfied. Then*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\|F^{(1:k)}_{x,h,n} - G^{(1:k)}_{x,h,n}\|_{\mathrm{TV}} \to 0,$$

for any $h \in \mathbf{R}^d$ and $k \in \mathbf{N}$. Moreover, under Assumptions 2.2.1-2.2.6, we have

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\|F^{(1:k)}_{x,n} - G^{(1:k)}_{x,n}\|_{\mathrm{TV}} \to 0.$$

Proof. We show $\int P_{0,n}(dx)\|F^{(1:k)}_{x,h,n} - G^{(1:k)}_{x,h,n}\|_{\mathrm{TV}}$ tends to 0 by induction. First, we show the case $k = 1$. By the triangular inequality, we obtain

$$\begin{aligned}\|F_{x,h,n} - G_{x,h,n}\|_{\mathrm{TV}} &= \|\int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy)F^{1|2}_{x,y,n}(\cdot) - G_{x,h,n}\|_{\mathrm{TV}} \\ &\leq P^{2|1}_{x,h,n}(\|F^{1|2}_{x,y,n} - G^{1|2}_{Z^{1,2}_n}\|_{\mathrm{TV}}) + \|P^{2|1}_{x,h,n}(G^{1|2}_{Z^{1,2}_n}(\cdot)) - G_{x,h,n}\|_{\mathrm{TV}}.\end{aligned}$$

The expectation of the first term in the right hand side tends to 0 by Assumptions 2.2.3 and 2.2.5. The expectation of the second term in the right hand side is

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\Big(\int_{\mathbf{R}^d} |\int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy)\frac{dG^{1|2}_{Z^{1,2}_n}}{d\mathrm{Leb}}(z) - \frac{dG_{x,h,n}}{d\mathrm{Leb}}(z)|dz\Big). \tag{2.3}$$

For any z, taking $f_z(x_1, x_2) = \phi(z; I_{1,2}(\theta_0)^{-1}(x_1 + x_2), I_{1,2}(\theta_0)^{-1})$, then

$$\frac{dG^{1|2}_{Z^{1,2}_n}}{d\mathrm{Leb}}(z) = f_z(Z_n, Z^{2|1}_n) \text{ and } \frac{dG_{x,h,n}}{d\mathrm{Leb}}(z) = \int_{\mathbf{R}^d} G^{2|1}_{0,h}(dg)f_z(Z_n, g).$$

For any sequence $C_n > 0$, (2.3) is

$$\begin{aligned}&2\int_{\mathcal{X}_n} P_{0,n}(dx)\int_{\mathbf{R}^d}\Big(\int_{\mathbf{R}^d} G^{2|1}_{0,h}(dg)f_z(Z_n, g) - \int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy)f_z(Z_n, Z^{2|1}_n)\Big)^+ dz \\ &\leq 2\int_{\mathcal{X}_n} P_{0,n}(dx)\int_{B_{C_n}(0)}\Big(\int_{\mathbf{R}^d} G^{2|1}_{0,h}(dg)f_z(Z_n, g) - \int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy)f_z(Z_n, Z^{2|1}_n)\Big)^+ dz \\ &+ 2\int_{\mathcal{X}_n} P_{0,n}(dx)\int_{B_{C_n}(0)^c} G_{x,h,n}(dz).\end{aligned} \tag{2.4}$$

Let $C_n \equiv C$. Since $f_z(x,y)$ is bounded above with respect to (z,x_1,x_2), using Fubini's theorem, (2.4) is

$$2\int_{B_C(0)}\int_{\mathcal{X}_n} P_{0,n}(dx)\Big(\int_{\mathbf{R}^d} G^{2|1}_{0,h}(dg)f_z(Z_n,g) - \int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy)f_z(Z_n,Z^{2|1}_n)\Big)^+ dz,$$

and the integrand with respect to dz tends to 0 by Lemma 2.4, and hence, (2.4) tends to 0 by the dominated convergence theorem. Therefore, we can choose $C_n \to \infty$ such that (2.4) tends to 0.

For any $\epsilon > 0$, take $C_1 > 0$ and $\omega_n = 1_{\{|Z_n(x)|>C_1\}}$ to be $\limsup_{n\to\infty} P_{0,n}(\omega_n) \leq \epsilon$. Since there exists a finite measure λ_h on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$ such that $(1-\omega_n)G_{x,h,n}(dz) \leq \lambda_h(dz)$. Then we have

$$\limsup_{n\to\infty}\int_{\mathcal{X}^n} P_{0,n}(dx)\int_{B_{C_n}(0)^c} G_{x,h,n}(dz) \leq \limsup_{n\to\infty}(P_{0,n}(\omega_n)+\lambda_h(B_{C_n}(0)^c)) \leq \epsilon.$$

Hence (2.3) tends to 0, and the case $k=1$ is proved.

Next, we consider the general case. Fix any $k \in \mathbf{N}$. Let S_s be a shift of a kernel K such that, $S_s(K)_h(A) = K_{h-s}(A-s)$. Set $W_n = I(\theta_0)^{-1}Z_n$, $U_{x,\cdot,n} = S_{W_n}(F_{x,\cdot,n})$, $V = S_{W_n}(G_{x,\cdot,n})$, and $\nu_{x,n} = \mu_{x,n} = \delta_h(\cdot - W_n)$. Let $F_{x,h,n} = \delta_{\{0\}}$ $(h \notin H_n)$. Then we can apply Lemma 2.3 and hence the first claim follows. For the second claim, set $\nu_{x,n} = F_{x,n}(\cdot - W_n)$ and $\mu = G_{x,n}(\cdot - W_n)$. □

We state some comments about the theorem. The local convergence rate are determined by the normal kernel $G_{x,\cdot,n}$. This fact was stated using simple models by some authors (for example, [33]). By Theorem 2.1 and Lemma 2.2, for some $n \in \mathbf{N}$, we have for any $i \geq n$,

$$\|F_{x,n} - F^i_{x,h,n}\|_{\mathrm{TV}} = \|G_{x,n} - G^i_{x,h,n}\|_{\mathrm{TV}} + o_{P_{0,n}}(1) \leq \sqrt{c_1 + c_2|h|^2 r^{2i}} + o_{P_{0,n}}(1)$$

for some $c_1, c_2 > 0$, where r is the largest eigenvalue of $J(\theta_0)$. Therefore, r is considered to be the rate of convergence of the marginal distribution of $F_{x,\cdot,n}$.

2.2.2 Multi-Stage Gibbs Sampler

We define the Multi-stage Gibbs sampler. Fix some $k \in \mathbf{N}$ and $d_1, \ldots, d_k \in \mathbf{N}$ such that $\sum_{i=1}^k d_i = d$. Let $(\Theta, \mathcal{F}) = \otimes_{i=1}^k(\Theta_i, \mathcal{F}_i)$ where $\Theta_i \subset \mathbf{R}^{d_i}$. Using this representation, we write $(H_n, \mathcal{H}_n) = \otimes_{i=1}^k(H_{n,i}, \mathcal{H}_{n,i})$, and

$$Z_n(x) = (Z_{n,1}(x), \ldots, Z_{n,k}(x)),$$
$$Z^{1,2}_n(x,y) = (Z^{1,2}_{n,1}(x,y), \ldots, Z^{1,2}_{n,k}(x,y)).$$

For each $i = 1, \dots, k$, let $(F^{2|1,i}_{x,y,h,n} : x \in \mathcal{X}_n, y \in \mathcal{Y}_n, h = (h_1, \dots, h_k) \in H_n)$ be a family of probability measure on $(H_{n,i}, \mathcal{H}_{n,i})$ which does not depend on h_i. We assume the following.

Assumption 2.2.8. *For any $i = 1, \dots, k$, we have for* Leb*-almost all $h \in \mathbf{R}^d$,*

$$\int P^{1,2}_{0,n}(dxdy) \| F^{1|2,i}_{x,y,h,n} - G^{1|2,i}_{Z^{1,2}_{n,i},h,n} \|_{\mathrm{TV}} \to 0,$$

where $G^{1|2,i}_{g,h}$ $(g \in \mathbf{R}^{d_i}, h \in \mathbf{R}^d)$ is

$$N(I_{1,2}(\theta_0)^{-1}_{i,i}(\sum_{j \neq i} I_{1,2}(\theta_0)_{i,j} h_j - g), I_{1,2}(\theta_0)^{-1}_{i,i}).$$

Note that $G^{1|2,i}_{g,h}$ does note depend on h_i.

Lemma 2.5. *Let Assumptions 2.2.3 and 2.2.5 be satisfied. Assume $F^{1|2}_{x,y,n}$ is absolutely continuous with respect to the Lebesgue measure for any n and $P^{1,2}_{0,n}$-almost all x, y. Let $f^{1|2}_{x,y,n}(h)$ denote $dF^{1|2}_{x,y,n}/d\mathrm{Leb}$ and if*

$$f^{1|2}_{x,y,n}(h_{1:i-1}, H_{n,i}, h_{i+1:k}) = \int_{H_{n,i}} f^{1|2}_{x,y,n}(h_{1:i-1}, g, h_{i+1:k}) dg \neq 0,$$

then let

$$F^{1|2,i}_{x,y,h,n}(dg) = \frac{f^{1|2}_{x,y,n}(h_{1:i-1}, g, h_{i+1:k})}{f^{1|2}_{x,y,n}(h_{1:i-1}, H_{n,i}, h_{i+1:k})} dg$$

and $F^{1|2,i}_{x,y,h,n} = \delta_{\{0\}}$ otherwise. Then Assumption 2.2.8 holds.

Proof. The following value tends in $P^{1,2}_{0,n}$-probability to 0 by Assumption 2.2.5:

$$\int |g^{1|2}_{Z^{1,2}_n}(h_{1:i-1}, H_{n,i}, h_{i+1:k}) - f^{1|2}_{x,y,n}(h_{1:i-1}, H_{n,i}, h_{i+1:k})| dh_{1:i-1} dh_{i+1:k}, \tag{2.5}$$

where $g^{1|2}_{Z^{1,2}_n}$ is $dG^{1|2}_{Z^{1,2}_n}/d\mathrm{Leb}$. We define

$$(G^{1|2,i}_{Z^{1,2}_{n,i},h})^*(dw) = 1_w(H_{n,i}) \frac{g^{1|2}_{Z^{1,2}_n}(h_{1:i-1}, w, h_{i+1:k})}{g^{1|2}_{Z^{1,2}_n}(h_{1:i-1}, H_{n,i}, h_{i+1:k})} dw.$$

For any bounded set $K \subset \mathbf{R}^d$, and for a large $n \in \mathbf{N}$ such that $K \subset H_{n,i}$, we have

$$\begin{aligned}
&\int_K G^{1|2}_{Z^{1,2}_n}(dh) \| G^{1|2,i}_{Z^{1,2}_{n,i},h} - F^{1|2,i}_{x,y,h,n} \|_{\mathrm{TV}} \\
&\leq \int_K G^{1|2}_{Z^{1,2}_n}(dh) \| (G^{1|2,i}_{Z^{1,2}_{n,i},h})^*(dw) - \frac{f^{1|2}_{x,y,n}(h_{1:i-1}, w, h_{i+1:k})}{g^{1|2}_{Z^{1,2}_n}(h_{1:i-1}, H_{n,i}, h_{i+1:k})} dw \|_{\mathrm{TV}} \\
&+ \int_K G^{1|2}_{Z^{1,2}_n}(dh) \| \frac{f^{1|2}_{x,y,n}(h_{1:i-1}, w, h_{i+1:k})}{g^{1|2}_{Z^{1,2}_n}(h_{1:i-1}, H_{n,i}, h_{i+1:k})} dw - F^{1|2,i}_{x,y,h,n} \|_{\mathrm{TV}} \\
&+ \int_K G^{1|2}_{Z^{1,2}_n}(dh) \| (G^{1|2,i}_{Z^{1,2}_{n,i},h})^* - G^{1|2,i}_{Z^{1,2}_{n,i},h} \|_{\mathrm{TV}}.
\end{aligned}$$

The first term in the right hand side is bounded above by $\|G^{1|2}_{Z^{1,2}_n} - F^{1|2}_{x,y,n}\|_{\mathrm{TV}}$, and the second term is bounded above by (2.5). Therefore, the first and the second term tends in $P^{1,2}_{0,n}$-probability to 0. The third term is

$$\begin{aligned}
&\int_K G^{1|2}_{Z^{1,2}_n}(dh) 2 (1 - \frac{g^{1|2}_{Z^{1,2}_n}(h_{1:i-1}, H_{n,i}, h_{i+1:k})}{g^{1|2}_{Z^{1,2}_n}(h_{1:i-1}, \mathbf{R}^{d_i}, h_{i+1:k})}) \\
&\leq 2 G^{1|2}_{Z^{1,2}_n}(\mathbf{R}^{d_1} \times \cdots \times H^c_{n,i} \times \cdots \times \mathbf{R}^{d_k}).
\end{aligned}$$

Fix $C > 0$ and consider a test $\omega = 1_{\{|Z^{1,2}_n| > C\}}$. Then there exists a finite measure λ such that $(1-\omega) G^{1|2}_{Z^{1,2}_n} \leq \lambda$. Then, the third term is bounded by $\omega + \lambda(\mathbf{R}^{d_1} \times \cdots \times H^c_{n,i} \times \cdots \times \mathbf{R}^{d_k})$. Then we can choose $C_n \to \infty$ instead of fixed C, such that the third term tends in $P^{1,2}_{0,n}$-probability to 0. Hence we have

$$\int P_{0,n}(dx) \int_K G^{1|2}_{Z^{1,2}_n}(dh) \| G^{1|2,i}_{Z^{1,2}_{n,i},h} - F^{1|2,i}_{x,y,h,n} \|_{\mathrm{TV}} \to 0.$$

Using the test ω, there exists a constant c such that Leb $\leq c(1-\omega) G^{1|2}_{Z^{1,2}_n}$ on K. Therefore, we have

$$\int_K \int P_{0,n}(dx) \| G^{1|2,i}_{Z^{1,2}_{n,i},h} - F^{1|2,i}_{x,y,h,n} \|_{\mathrm{TV}} dh \to 0.$$

Hence the claim follows. □

Assume that we have an observation x from $P_{0,n}$.

Step 0 Set $h^0 = (h^0_1, \ldots, h^0_k) \in H_n$, then, go to Step 1.

Step i Generate y^i from $P^{2|1}_{h^{(i-1)},x,n}$. Then, for $j = 1, \ldots, k$, generate h^i_j from $F^{1|2,j}_{x,y^i,(h^i_{1:j-1},h^{i-1}_{j:k}),n}$. Then set $h^i = (h^i_1, \ldots, h^i_k)$ and go to Step $i+1$.

This procedure defines a Markov chain $h^1, h^2, \ldots$, and let $F_{1,x,\cdot,n}(\cdot)$ denote its transition kernel. We are going to show that this transition kernel tends to a simple transition kernel of the following Markov chain $h^1, h^2, \ldots$.

Step 0 Set $h^0 \in \mathbf{R}^d$, then, go to Step 1.

Step i Generate g_i from $G^{2|1}_{Z_n,h^{i-1}} = N(I_{2|1}(\theta_0)h^{i-1} + Z_n, I_{2|1}(\theta_0))$. Then, for $j = 1, \ldots, k$, generate h^i_j from $G^{1|2,j}_{g_i,(h^i_{1:j-1},h^{i-1}_{j:k})}$. Then set $h^i = (h^i_1, \ldots, h^i_n)$ and go to Step $i+1$.

Let $G_{1,x,\cdot,n}(\cdot)$ denote the transition kernel, that is,

$$N((C^l)^{-1}(B + C^l - C)(g - I(\theta_0)^{-1}Z_n), (C^l)^{-1}(B + \mathrm{diag}(C))((C^l)^{-1})^T),$$

where $B = I_{2|1}(\theta_0)$ and $C = I_{1,2}(\theta_0)$. Then $G_{x,n} = N(I(\theta_0)^{-1}Z_n, I(\theta_0))$ is the invariant distribution of $G_{1,x,\cdot,n}$ by Lemma 2.2.

Proposition 2.1. *Let Assumptions 2.2.1-2.2.5, and 2.2.8 be satisfied. Then for* Leb-*almost all* $h = (h_1, \ldots, h_k)$*,*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\|F^{(1:k)}_{1,x,h,n} - G^{(1:k)}_{1,x,h,n}\|_{\mathrm{TV}} \to 0.$$

Moreover, under Assumptions 2.2.1-2.2.6, and 2.2.8, we have

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\|F^{(1:k)}_{1,x,n} - G^{(1:k)}_{1,x,n}\|_{\mathrm{TV}} \to 0.$$

Proof. We show the case $k = 1$.

$$\begin{aligned}
&\|F_{1,x,h,n} - G_{1,x,h,n}\|_{\mathrm{TV}} \\
&= \|P^{2|1}_{x,h,n}(\prod_{i=1}^{k} F^{1|2,i}_{x,y,(f_{1:i-1},h_{i:k}),n}(df_i)) - G_{1,x,h,n}(df_1 \cdots df_k)\|_{\mathrm{TV}} \\
&\leq \sum_{i=1}^{k} \|P^{2|1}_{x,h,n}(\prod_{j=1}^{i-1} G^{1|2,j}_{Z^{1,2}_{n,i},(f_{1:j-1},h_{j:k})}(df_j)(F^{1|2,i}_{x,,y,(f_{1:i-1},h_{i:k}),n} - G^{1|2,i}_{Z^{1,2}_{n,i},(f_{1:i-1},h_{i:k})}(df_i))\|_{\mathrm{TV}} \\
&+ \|P^{2|1}_{x,h,n}(\prod_{i=1}^{k} G^{1|2,i}_{Z^{1,2}_{n,i},(f_{1:i-1},h_{i:l})}(df_i)) - G_{1,x,h,n}(df_1 \cdots df_k)\|_{\mathrm{TV}}.
\end{aligned}$$

Fix $c > 0$ and let $\omega_n(x) = 1_{\{|Z_n(x)|>c\}}$. For any $\epsilon > 0$, take $c > 0$ to be $\limsup_{n\to\infty} P_{0,n}(\omega_n) \leq \epsilon$. Then there exists a finite measure λ_i on $\otimes_{j=1}^{i-1}\mathbf{R}^{d_i}$ such that $(1-\omega_n)\prod_{j=1}^{i-1} G^{1|2,i}_{Z^{1,2}_{n,i},(f_{1:j-1},h_{j:k})}(df_j) \leq \lambda_i$. Then, the expectation of the first term in the right hand side tends to 0 by Assumption 2.2.8. The expectation of the second term also tends to 0 by Lemma 2.4.

The case of $k > 1$ is almost identical to Theorem 2.1. The remaining convergence also follows from the same reason. □

The comparison with the two-stage Gibbs sampler is not clear. In this case,

$$J_1(\theta_0) = (I_{1,2}(\theta_0)^l)^{-1}(I_{1,2}(\theta_0)^l - I(\theta_0)),$$
$$\Sigma_1(\theta_0) = (I_{1,2}(\theta_0)^l)^{-1}(I_{2|1}(\theta_0) + \mathrm{diag}(I_{1,2}(\theta_0)))((I_{1,2}(\theta_0)^l)^{-1})^T.$$

As we mentioned in Section 1.3.2 that the order of the largest eigenvalue of $J_0(\theta_0)$ and $J_1(\theta_0)$ vary with a change in the matrices $I(\theta_0)$ and $I_{2|1}(\theta_0)$.

2.2.3 Convergence of Marginal Distribution

Lemma 2.6. *For any $h \in \mathbf{R}^d$, and $\omega_{n,h} : \mathcal{X}_n \to [0,1]$, we assume $\int P_{0,n}(dx)\omega_{n,h}(x) \to 0$. For any $\epsilon_n \to 0$, and for any compact set K of $\mathbf{R}^d$, we also assume*

$$\int P_{0,n}(dx) \sup_{h\in K, |\delta|<\epsilon_n} |\omega_{n,h} - \omega_{n,h+\delta}| \to 0.$$

Then, there exists a sequence $M_n \to \infty$ such that

$$\lim_{n\to\infty} \int P_{0,n}(dx) \sup_{|h|<M_n} \omega_{n,h} = 0.$$

Proof. For $M > 0$, define $C_0 \subset C_1 \subset \cdots \subset B_M(0)$ such that C_i has a finite number of elements, and $B_M(0) \subset \cup_{h\in C_i} B_{i^{-1}}(h)$ for any $i \in \mathbf{N}$. Then we obtain

$$\int P_{0,n}(dx) \sup_{|h|<M} \omega_{n,h} \leq \int P_{0,n}(dx) \sup_{h\in C_{m_n}} \omega_{n,h} + \sup_{|\delta|<m_n^{-1}, |h|<M} |\omega_{n,h} - \omega_{n,h+\delta}| \to 0$$

for some $m_n \to \infty$. Hence, there exists $M_n \to \infty$ such that $\int P_{0,n}(dx) \sup_{|h|<M_n} \omega_{n,h}$ tends to 0. □

We assume that a probability measure $\nu_{x,n}$ satisfies the following property: for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} P_{0,n}(\nu_{x,n}(B_\delta(0)^c)) \leq \epsilon. \tag{2.6}$$

Lemma 2.7. *Let $U_{x,\cdot,n}$ and $V_{\cdot}$ be transition kernels on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$ with invariant distribution $U_{x,n}$ and V. Let $\nu_{x,n}$ be a probability measure on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$ satisfying (2.6). Assume there exists a sequence of random variable $(W_n; n \in \mathbf{N})$ such that (W_n) is $P_{0,n}$-tight and*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\|U_{x,n} - V_{\cdot}\|_{\mathrm{TV}} \to 0, \quad \int_{\mathcal{X}_n} P_{0,n}(dx)\|U_{x,h+W_n,n} - V_{h+W_n}\|_{\mathrm{TV}} \to 0.$$

Moreover, for any $\epsilon_n \to 0$, and for any compact set K of $\mathbf{R}^d$, we have

$$\int_{\mathcal{X}_n} P_{0,n}(dx) \sup_{h,g\in K, |h-g|<\epsilon_n} (\|U_{x,h+W_n,n} - U_{x,g+W_n,n}\|_{\mathrm{TV}} + \|V_{h+W_n} - V_{g+W_n}\|_{\mathrm{TV}}) \to 0, \tag{2.7}$$

and $\lim_{l\to\infty} \sup_{h\in K} \|V - V_h^l\|_{\mathrm{TV}} = 0$ for any compact set $K \subset \mathbf{R}^d$. Then for any $l_n \to \infty$, we have

$$\lim_{n\to\infty} \int P_{0,n}(dx)\|U_{x,n} - U_{x,\nu,n}^{l_n}\|_{\mathrm{TV}} = 0.$$

Proof. Since $U_{x,n}$ is the invariant distribution of $U_{x,\cdot,n}$, $\|U_{x,n} - U_{x,\nu,n}^l\|_{\mathrm{TV}}$ is decreasing function with respect to l. Therefore, we fix $l_n \equiv l$ sufficiently large. By triangular inequality,

$$\|U_{x,n} - U_{x,\nu,n}^l\|_{\mathrm{TV}} \le \|U_{x,n} - V\|_{\mathrm{TV}} + \|V - V_{x,\nu,n}^l\|_{\mathrm{TV}} + \|V_{x,\nu,n}^l - U_{x,\nu,n}^l\|_{\mathrm{TV}}. \tag{2.8}$$

where $U_{x,\nu,n}^l, V_{x,\nu,n}^l$ denote the k-th marginal distribution of each Markov chain $U_{x,\cdot,n}, V_{\cdot}$ starting from $h_0 \sim \nu_{x,n}$. The first term integrated by $P_{0,n}$ tends to 0 by assumption. We show that the integral of the second term and the third term with respect to $P_{0,n}$ tend to 0 as $n \to \infty$. First, we show the convergence of the second term. Fix any $\epsilon > 0$. By consistency of $\nu_{x,n}$ there exists $M > 0$ and $N \in \mathbf{N}$ such that

$$\int P_{0,n}(dx)(\nu_{x,n}(B_M(0)^c) \le \epsilon/2 \ (N \le n).$$

Therefore we have

$$\int P_{0,n}(dx)\|V - V_{x,\nu,n}^l\|_{\mathrm{TV}} \le \int P_{0,n}(dx) \int_{\mathbf{R}^d} \|V - V_h^l\|_{\mathrm{TV}} \nu_{x,n}(dh)$$
$$\le \epsilon + \sup_{h\in B_M(0)} \|V - V_h^l\|_{\mathrm{TV}}$$

and the right hand side does not depend on n. When $l \to \infty$, by assumption, the right hand side tends to ϵ. Since $\epsilon > 0$ is arbitrary, the integral with respect to $P_{0,n}$ of the second term in (2.8) tends to 0.

Next, we show that the integral with respect to $P_{0,n}$ of the third term in (2.8) also tends to 0. We have

$$\|V^l_{x,\nu,n} - U^l_{x,\nu,n}\|_{\mathrm{TV}} \le 2\nu_{x,n}(B_M(W_n)^c) + \sup_{|h|<M} \|V^l_{h+W_n} - U^l_{x,h+W_n,n}\|_{\mathrm{TV}}.$$

Let $\omega_{h,n}(x) = \|V^l_{h+W_n} - U^l_{x,h+W_n,n}\|_{\mathrm{TV}}$, then by assumption and Lemma 2.3, $\int P_{0,n}(dx)\omega_{h,n}(x) \to 0$. By assumption, we have

$$|\omega_{h,n}(x) - \omega_{g,n}(x)| \le \|U_{x,h+W_n,n} - U_{x,g+W_n,n}\|_{\mathrm{TV}} + \|V_{h+W_n} - V_{g+W_n}\|_{\mathrm{TV}}.$$

Taking $|h-g| < \epsilon_n \to 0$, by Lemma 2.6, we have $\int P_{0,n}(dx) \sup_{|h|<M_n} \omega_{h,n}(x) \to 0$ for some $M_n \to \infty$. Then the claim follows. □

Corollary 2.2. *Let Assumptions 2.2.1-2.2.8 be satisfied. If $\mu = (\mu_{x,n}; x \in \mathcal{X}_n, n \in \mathbf{N})$ satisfies (2.6), then for any $l_n \to \infty$,*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\|F_{i,x,n} - F^{l_n}_{i,x,\mu,n}\|_{\mathrm{TV}} \to 0 \ (i = 0, 1).$$

Proof. Fix any $i = 0, 1$. Let S_s be a shift of a kernel K such that, $S_s(K)_h(A) = K_{h-s}(A - s)$. Set $W_n = I(\theta_0)^{-1}Z_n$, $U_{x,\cdot,n} = S_{W_n}(F_{i,x,\cdot,n})$, $V_\cdot = S_{W_n}(G_{i,x,\cdot,n})$, and $\nu_{x,n} = \mu_{x,n}(\cdot - W_n)$. Using Lemma 2.7, it is enough to show that the convergence (2.7) holds. By definition

$$\|F^l_{i,x,h,n} - F^l_{i,x,g,n}\|_{\mathrm{TV}} \le \|P^{2|1}_{x,h,n} - P^{2|1}_{x,g,n}\|_{\mathrm{TV}},$$

therefore, the convergence of the first term in the left hand side of (2.7) follows by Assumption 2.2.8. Since $V_h = N(Ah, B)$ for some $d \times d$-matrix A, B, by Lemma 2.1, the second term in the left hand side is

$$\|V_h - V_g\|^2_{\mathrm{TV}} \le H(V_h, V_g)^2 = 2(1 - \exp(-\langle A(h-g), B^{-1}A(h-g)\rangle/8)).$$

Therefore, by Lemma 2.6, then the claim follows by Lemma 2.7. □

2.2.4 Point Estimation

Let $d_g, d_h \in \mathbf{N}_0$ such that $d_g + d_h = d$. Let $K_\cdot(\cdot) : \mathbf{R}^{d_g} \times \mathcal{B}(\mathbf{R}^{d_g}) \to [0, 1]$ be a probability transition kernel. For the above transition kernel $K_\cdot(\cdot)$, we will define another probability transition kernel $K^{1,2}_\cdot(\cdot) : \mathbf{R}^d \times \mathcal{B}(\mathbf{R}^d) \to [0, 1]$ to be

$$K^{1,2}_{g,h_1} = K^{1,2}_{g,h_2} \ (g \in \mathbf{R}^{d_g}, h_1, h_2 \in \mathbf{R}^{d_h})$$
$$K^{1,2}_{g,h}(A \times \mathbf{R}^{d_h}) = K_g(A) \ (g \in \mathbf{R}^{d_g}, h \in \mathbf{R}^{d_h}, A \in \mathcal{B}(\mathbf{R}^{d_g})).$$

If the transition kernel $K_{\cdot}(\cdot)$ has an invariant probability measure K, then for any fixed $h \in \mathbf{R}^{d_h}$, $K^{1,2}(\cdot) = \int_{\mathbf{R}^{d_g}} K(dg^*)K^{1,2}_{g^*,h}(\cdot)$ is an invariant probability measure of $K^{1,2}_{\cdot}(\cdot)$. For any probability measure ν on $(\mathbf{R}^{d_g}, \mathcal{B}(\mathbf{R}^{d_g}))$, we define a probability measure $\nu K^{1,2}(\cdot) = \int_{\mathbf{R}^{d_g}} \nu(dg^*)K^{1,2}_{g^*,h}(\cdot)$.

Assume we have $((g^1, h^1), \dots, (g^m, h^m))$ from $(K^{1,2}_{\nu K^{1,2}})^{(1:m)}$, we define the Gibbs sampling estimator $\overline{h}^m$ as follows. Let $h^i = (h^i_1, \dots, h^i_d)$ and $\overline{h}^m = (\overline{h}^m_1, \dots, \overline{h}^m_d)$. Let $h_i^{(1)} \leq h_i^{(2)} \leq \dots \leq h_i^{(m)}$ denote the ordered sequence of $h_i^1, \dots, h_i^m$. Then for any $i = 1, \dots, d$, we define

$$\overline{h}^m_i(h) = \begin{cases} h_i^{(j)} & \text{when } m = 2j+1 \\ (h_i^{(j)} + h_i^{(j+1)})/2 & \text{when } m = 2j. \end{cases} \tag{2.9}$$

Lemma 2.8. *Let $U_{x,\cdot,n}$ and $V_{\cdot}$ be probability transition kernels on $(\mathbf{R}^{d_g}, \mathcal{B}(\mathbf{R}^{d_g}))$ with invariant measures $U_{x,n}$ and V. Assume for any $k \in \mathbf{N}$, we have*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)(U^{1,2}_{x,n})^{(1:k)} \Rightarrow (V^{1,2})^{(1:k)}.$$

We assume the existence of $\nu_{x,n}$, which is a probability measure, such that for any $l_n \to \infty$,

$$\lim_{n\to\infty} \int_{\mathcal{X}_n} P_{0,n}(dx) \| U^{l_n}_{x,\nu,n} - U_{x,n} \|_{\mathrm{TV}} = 0. \tag{2.10}$$

Further assume that the transition kernel $V^{1,2}_{\cdot}$ is positive Harris recurrent and for any $\epsilon > 0$, $c^i_\epsilon := \int V^{1,2}(d((g_1, \dots, g_d), (h_1, \dots, h_d)))1_{\{h_i \geq \epsilon\}} < 1/2$ for any $i = 1, \dots, d$. Then, for any $l_n \to \infty$ and $\nu^{1,2} = \nu U^{1,2}_{x,\cdot,n}$, we have

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\Big(\int (U^{1,2}_{x,\nu^{1,2},n})^{(1:l_n)}(d(g,h))(|\overline{h}_{l_n}(h)| \wedge 1) \Big) \to 0.$$

Proof. Fix any $i = 1, \dots, d$. For any $\epsilon > 0$, we have

$$\{(x,g,h) \in \mathcal{X}_n \times \mathbf{R}^{l(d_g+d_h)}; \overline{h}^l_i(h) \geq \epsilon\} = \{(x,g,h); \sum_{j=1}^{l} 1_{\{h_i^j \geq \epsilon\}} \geq \frac{l}{2}\}$$

$$= \{(x,g,h); \frac{1}{l}\sum_{j=1}^{l}(1_{\{h_i^j \geq \epsilon\}} - c^i_\epsilon) \geq \frac{1}{2} - c^i_\epsilon\}.$$

Therefore, by Chebyshev's inequality, we have

$$(U^{1,2}_{x,n})^{(1:l)}(\overline{h}^l_i(h) \geq \epsilon) \leq (\frac{1}{2} - c^i_\epsilon)^{-1}(U^{1,2}_{x,n})^{(1:l)}(\frac{1}{l}\sum_{j=1}^{l} 1_{\{h_i^j \geq \epsilon\}} - c^i_\epsilon)^{+}.$$

For any $m \leq l$ and any sequence $(a_j; j \in \mathbf{N})$, we have

$$\sum_{j=1}^{l} a_j = \sum_{j=0}^{[l/m]-1} (\sum_{k=1}^{m} a_{jm+k}) + \sum_{j=[l/m]m+1}^{l} a_k$$

where $[x]$ denotes the largest integer smaller than x.

For $h = (h_1, \ldots, h_d)$, let $a_i(h) = 1_{\{h_i \geq \epsilon\}} - c_\epsilon^i$ $(i = 1, \ldots, d)$. First, consider the case $\nu_{x,n} = U_{x,n}$. Since $U_{x,n}^{1,2}$ is the invariant measure of $U_{x,\cdot,n}^{1,2}$, we have

$$(U_{x,n}^{1,2})^{(1:l)}(\frac{1}{l}\sum_{j=1}^{l} a_i(h^j))^+ \leq (U_{x,n}^{1,2})^{(1:m)}(\frac{1}{m}\sum_{j=1}^{m} a_i(h^j))^+ + \frac{l - [l/m]m}{l}.$$

Therefore, for any $l_n \to \infty$ and $m \in \mathbf{N}$, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \int_{\mathcal{X}_n} P_{0,n}(dx)\Big(\int (U_{x,n}^{1,2})^{(1:l_n)}(d(g,h))(\overline{h}_i^{l_n}(h) > \epsilon)\Big) \\
&\leq (\frac{1}{2} - c_\epsilon)^{-1} \limsup_{n\to\infty} \int_{\mathcal{X}_n} P_{0,n}(dx)\Big(\int (U_{x,n}^{1,2})^{(1:m)}(d(g,h))(\frac{1}{m}\sum_{j=1}^{m} a_i(h^j))^+\Big) \\
&= (\frac{1}{2} - c_\epsilon)^{-1} \int (V^{1,2})^{(1:m)}(d(g,h))(\frac{1}{m}\sum_{j=1}^{m} a_i(h^j))^+.
\end{aligned}$$

Then by ergodicity of the transition kernel, the right hand side tends to 0 when $m \to \infty$. Therefore the claim follows when $\nu_{x,n} = U_{x,n}$.

Next, we show that for any $\nu_{x,n}$ satisfying (2.10), the same conclusion holds. For any $m_n \to \infty$ such that $m_n/l_n \to 0$, we have

$$\frac{1}{l_n}\sum_{j=1}^{l_n} a_j = \frac{1}{l_n}\sum_{j=1}^{m_n-1} a_j + (\frac{l_n - m_n}{l_n})\frac{1}{l_n - m_n}\sum_{j=m_n}^{l_n} a_j.$$

Then, the first term is negligible, and the integral of the second term is

$$\begin{aligned}
&\int (U_{x,\nu^{1,2},n}^{1,2})^{(1:l_n)}(d(g,h))((\frac{l_n - m_n}{l_n})\frac{1}{l_n - m_n}\sum_{j=m_n}^{l_n} a_i(h^j))^+ \\
&\leq \|U_{x,\nu,n}^{m_n} - U_{x,n}^{m_n}\|_{\mathrm{TV}} + \int (U_{x,n}^{1,2})^{(1:l_n - m_n)}(d(g,h))(\frac{1}{l_n - m_n}\sum_{j=0}^{l_n - m_n} a_i(h^j))^+.
\end{aligned}$$

The first term tends to 0 by assumption. The second term also tends to 0 by the previous arguments. □

In the following corollary, we set $F^{1,2}_{x,(g^*,h^*),n}(dgdh) = F_{x,g^*,n}(dg)\delta_g(h)$.

Corollary 2.3. *Let Assumptions 2.2.1-2.2.8 be satisfied. For any $l_n \to \infty$, and for $\nu_{x,n}$ such that (2.6), we have*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\Big(\int_{H_n^{l_n}} F^{(1:l_n)}_{i,x,\nu,n}(dh)(|\overline{h}^{l_n}(h) - m(F_{x,n})| \wedge 1)\Big) \to 0 \ (i = 0, 1).$$

where $m(\mu)$ is the one of its median of the measure μ.

Proof. Fix any $i = 0, 1$. Let S_s be a shift of a kernel K such that, $S_s(K)_h(A) = K_{h-s}(A - s)$. Set $W_n = I(\theta_0)^{-1}Z_n$, $U_{x,\cdot,n} = S_{W_n}(F_{i,x,\cdot,n})$, $V = S_{W_n}(G_{i,x,\cdot,n})$, and $\mu_{x,n} = \nu_{x,n}(\cdot - W_n)$. Then, this is easy corollary of Lemma 2.8. Note that $\int P_{0,n}(dx)|m(G_{Z_n}) - m(F_{x,n})| \wedge 1$ tends to 0 by the second condition of Assumption 2.2.5. □

Corollary 2.4. *Let Assumptions 2.2.1-2.2.8 be satisfied. For any $l_n \to \infty$, and for $\nu_{x,n}$ such that (2.6), and taking $\overline{\theta}_n^{l_n} = \theta_0 + \overline{h}^{l_n} n^{-1/2}$, we have*

$$\int_{\mathcal{X}_n} P_{0,n}(dx) \int_{H_n^{l_n}} F^{(1:l_n)}_{i,x,\nu,n}(dh)(n^{1/2}|\overline{\theta}_n^{l_n} - \overline{\theta}_n| \wedge 1) \to 0 \ (i = 0, 1),$$

where $\overline{\theta}_n = \theta_0 + m(F_{x,n})n^{-1/2}$.

Proof. Easy. □

2.2.5 Some Speed Up Methods

We consider some speed up methods. The *Rao-Blackwellization* is one of them which was introduced by [11]. Consider we have an observation x from $P_{0,n}$.

Step 0 Set $g_0 \in H_n$, then, go to Step 1.

Step i Generate y_i from $P^{2|1}_{g_{i-1},x,n}$. Then, compute $h_i = \int g F^{1|2}_{x,y_i,n}(dg)$, generate g_i from $F^{1|2}_{x,y_i,n}$ and go to Step $i+1$.

This iteration defines a Markov chain $(g_1, h_1), (g_2, h_2), \ldots$. Let $F^{1,2}_{x,\cdot,n}$ denote the transition kernel.

Step 0 Set $g_0 = g \in \mathbf{R}^d$, then, go to Step 1.

Step i Generate f_i from $G^{2|1}_{Z_n,g_{i-1}}$ and compute $h_i = I_{1,2}(\theta_0)^{-1}f_i$. Then, generate g_i from $G^{1|2}_{f_i}$.

This iteration defines a Markov chain $(g_1, h_1), (g_2, h_2), \ldots$. Let $G^{1,2}_{x,\cdot,n}$ denote the transition kernel.

For any signed measure ν on $(\mathbf{R}^{kd}, \mathcal{B}(\mathbf{R}^{kd}))$ and for any $u_1, \ldots, u_k \in \mathbf{R}^d$, let

$$\|\nu\|_{u_{1:k}} = |\int_{\mathbf{R}^d \times \cdots \times \mathbf{R}^d} \exp(i \sum_{j=1}^{k} \langle u_j, z_j \rangle) \nu(dz_1, \ldots, dz_k)|.$$

Assumption 2.2.9. *We have*

$$\lim_{n\to\infty} \int_{\mathcal{X}_n \times \mathcal{Y}_n} P^{1,2}_{0,n}(dxdy)| \int_{H_n} t F^{1|2}_{x,y,n}(dt) - \int_{\mathbf{R}^d} t G^{1|2}_{Z^{1|2}_n}(dt)| = 0.$$

Proposition 2.2. *Under Assumptions 2.2.1 to 2.2.5 and 2.2.9, we have*

$$\lim_{n\to\infty} \int_{\mathcal{X}_n} P_{0,n}(dx) \| (F^{1,2}_{x,(g,h),n})^{(1:k)} - (G^{1,2}_{x,(g,h),n})^{(1:k)} \|_{u_{1:2k}} = 0$$

and if Assumption 2.2.6 is also satisfied, then

$$\lim_{n\to\infty} \int_{\mathcal{X}_n} P_{0,n}(dx) \| (F^{1,2}_{x,n})^{(1:k)} - (G^{1,2}_{x,,n})^{(1:k)} \|_{u_{1:2k}} = 0.$$

Proof. We show the case $k = 1$.

$$\begin{aligned}
&\|F^{1,2}_{x,(g,h),n} - G^{1,2}_{x,(g,h),n}\|_{u_{1:2}} \\
&= \int_{\mathcal{Y}_n} P^{2|1}_{x,g,n}(dy) \|\delta_{\{\int t F^{1|2}_{x,y,n}(dt)\}} - \delta_{\{I_{1,2}(\theta_0)^{-1} Z^{1,2}_n\}}\|_{u_1} + \|F^{1|2}_{x,y,n} - G^{1|2}_{Z^{1,2}_n}\|_{u_2} \\
&\quad + \| \int_{\mathcal{Y}_n} P^{2|1}_{x,g,n}(dy) \delta_{\{I_{1,2}(\theta_0)^{-1} Z^{1,2}_n\}}(dh_1) G^{1|2}_{Z^{1,2}_n}(dg_1) - G^{1,2}_{x,g,n}(dg_1 dh_1)\|_{u_{1:2}}.
\end{aligned}$$

The second term integrated by $P_{0,n}$ tends to 0 by Assumptions 2.2.3 and 2.2.5, and the last term tends to 0 by Lemma 2.4. By a Taylor expansion, the first term is

$$\|\delta_{\{\int t F^{1,2}_{x,y,n}(dt)\}} - \delta_{\{I_{1,2}(\theta_0)^{-1} Z^{1,2}_n\}}\|_{u_1} \leq |u_1| \cdot | \int t F^{1|2}_{x,y,n}(dt) - \int t G^{1|2}_{Z^{1,2}_n}(dt)|.$$

Therefore by Assumption 2.2.9 the above value tends to 0, and hence the case $k = 1$ is proved.

In the general case, we use a shift of a kernel and Lemma 2.3. For any kernel $K_{x,\cdot,n} = (K_{x,(g,h),n}(A); g, h \in \mathbf{R}^d, A \in \mathbf{B}(\mathbf{R}^{2d}))$, we define a shift $S_s(K_{x,\cdot,n})_{(g,h)}(A) = K_{x,(g-s,h-s)}(A - (s,s))$. Set $W_n = I(\theta_0)^{-1} Z_n$, $U_{x,\cdot,n} = S_{W_n}(F^{1,2}_{x,\cdot,n})$, $V. = S_{W_n}(G^{1,2}_{x,\cdot,n})$, $\nu_{x,n} = F_{x,n}(\cdot - W_n)$ and $\mu_{x,n} = G_{x,n}(\cdot - W_n)$. Then we can apply Lemma 2.3 and hence the claim follows. □

We note a few comments about this proposition. First, in generally, $(F^{1,2}_{x,(g,h),n})^{(1:k)}$ does not tend to $(G^{1,2}_{x,(g,h),n})^{(1:k)}$ in $\|\cdot\|_{\mathrm{TV}}$ sense, since $P^{2|1}_{x,g,n}$ may be a probability measure on a discrete space and therefore $h_i = h_i(x, y_i)$ may be a random variable taking discrete values.

The Rao-Blackwellization use only $(h_1, h_2, \ldots, h_k)$ for estimation and define $\overline{h}^m_i$ as (2.9). This marginal distribution tends to a Markov chain of kernel $G_{2,x,\cdot,n}$, which is

$$J_2(\theta_0) = J_0(\theta_0)$$
$$\Sigma_2(\theta_0) = I_{1,2}(\theta_0)^{-1}(I_{2|1}(\theta_0) + I_{2|1}(\theta_0)I_{1,2}(\theta_0)^{-1}I_{2|1}(\theta_0))I_{1,2}(\theta_0)^{-1}.$$

Then $\det(\Sigma_2(\theta_0)) < \det(\Sigma_0(\theta_0))$, since $\det(\Sigma_2(\theta_0))/\det(\Sigma_0(\theta_0))$ is equal to $\det(I_{2|1}(\theta_0))/\det(I_{1,2}(\theta_0))$ and $\det(I_{2|1}(\theta_0)) < \det(I_{1,2}(\theta_0))$. Therefore, in this sense, the Rao-Blackwellization method is more efficient than the Gibbs sampler defined in Subsection 2.2.1. Note that if $\det(I_{2|1}(\theta_0))$ is small relative to $\det(I(\theta_0))$, then the Rao-Blackwellization is efficient, but $\det(I_{2|1}(\theta_0))$ is large relative to $\det(I(\theta_0))$, then the difference between the Gibbs sampler and its Rao-Blackwellization method is small.

Corollary 2.5. *Let assumptions 2.2.1-2.2.9 be satisfied. Let $\nu_{x,n}$ be a measure such that (2.6), We define $\nu^{1,2} = \nu_{x,n}F^{1,2}_{x,\cdot,n}$. For any $l_n \to \infty$ we have*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\Big(\int_{H^{l_n}_n}(F^{1,2}_{x,\nu^{1,2},n})^{(1:l_n)}(d(g,h))(|\overline{h}^{l_n}(h) - m(F_{x,n})| \wedge 1)\Big) \to 0.$$

Corollary 2.6. *Let assumptions 2.2.1-2.2.9 be satisfied. Let $\nu_{x,n}$ be a measure such that (2.6), We define $\nu^{1,2} = \nu_{x,n}F^{1,2}_{x,\cdot,n}$. For any $l_n \to \infty$, taking $\overline{\theta}^{l_n}_n = \theta_0 + \overline{h}^{l_n}n^{-1/2}$, we have*

$$\int_{\mathcal{X}_n} P_{0,n}(dx)\int_{H^{l_n}_n}(F^{1,2}_{x,\nu^{1,2},n})^{(1:l_n)}(d(g,h))(n^{1/2}|\overline{\theta}^{l_n}_n - \overline{\theta}_n| \wedge 1) \to 0,$$

where $\overline{\theta}_n = \theta_0 + m(F_{x,n})n^{1/2}$.

Note that improvement by the Rao-Blackwellization has already been addressed in [23] in another sense.

2.3 Independent and Identically Distributed Observations

We consider independent and identically distributed observations. As in Section 1.4, we take two parametric families $(\mathcal{X}, \mathcal{A}, P_\theta; \theta \in \Theta)$, $(\mathcal{Y}, \mathcal{B}, P^{2|1}_{x,\theta}; \theta \in$

$\Theta, x \in \mathcal{X}$). Let $\theta_0 \in \Theta$. We set $\mathcal{X}_n = \mathcal{X}^n, \mathcal{A}_n = \mathcal{A}^n$, $P_{h,n} = P^n_{\theta_0+hn^{-1/2}}$, $\mathcal{Y}_n = \mathcal{Y}^n, \mathcal{B}_n = \mathcal{B}^n$ and $P^{2|1}_{x,h,n}(dy) = \prod_{i=1}^n P^{2|1}_{x_i,\theta_0+hn^{-1/2}}(dy_i)$.

We also define $P^{1,2}_{\theta,\vartheta}(dxdy) = P_\theta(dx)P^{2|1}_{x,\vartheta}(dy)$ and $P^{1,2}_{g,h,n} = P^{1,2}_{\theta_0+gn^{-1/2},\theta_0+hn^{-1/2}}$, and we write $P^{1,2}_{\theta,\theta} = P^{1,2}_\theta$ and $P^{1,2}_{h,h,n} = P^{1,2}_{h,n}$

Let Assumption 1.4.3 be satisfied. Let $\tilde{\eta}^{2|1}_{\theta_0}(x,y) = 2\eta^{2|1}_{\theta_0}(x,y)/p^{1,2}_{\theta_0}(x,y)^{1/2}$ if $p^{1,2}_{\theta_0}(x,y)^{1/2} > 0$ and $\tilde{\eta}^{2|1}_{\theta_0}(x,y) = 0$ otherwise. Let $Z^{2|1}_n = n^{-1/2}\sum_{i=1}^n \tilde{\eta}^{2|1}_{\theta_0}(x_i,y_i)$ if we have observations $(x_1,\ldots,x_n)$ and $y_1,\ldots,y_n)$.

Proposition 2.3. *Under Assumption 1.4.3, Assumption 2.2.4 is satisfied.*

Proof. First, we show that for $(P_{\theta_0})^\infty$-almost all $(x_1,x_2,\ldots) \in \mathcal{X}^\infty$,

$$\mathcal{L}(Z^{2|1}_n|P^{2|1}_{(x_1,\ldots,x_n),0,n}) \Rightarrow N(0, I_{2|1}(\theta_0)).$$

This claim will follow from the following three properties by the Lindeberg Central Limit Theorem (for example, Theorem 27.2 of [2]). We show

$$\int_{\mathcal{Y}} P^{2|1}_{x,\theta_0}(dy)\langle\tilde{\eta}^{2|1}_{\theta_0}(x,y),h\rangle = 0 \ (P_{\theta_0} \text{ a.s. } x),$$

$$n^{-1}\sum_{i=1}^n \int_{\mathcal{Y}} P^{2|1}_{x_i,\theta_0}(dy)\langle\tilde{\eta}^{2|1}_{\theta_0}(x_i,y),h\rangle^2 \to \langle h, I_{2|1}(\theta_0)h\rangle,$$

$$n^{-1}\sum_{i=1}^n \int_{\mathcal{Y}} P^{2|1}_{x_i,\theta_0}(dy)\langle\tilde{\eta}^{2|1}_{\theta_0}(x_i,y),h\rangle^2 1_{\{|\langle\tilde{\eta}^{2|1}_{\theta_0}(x_i,y),h\rangle|>n^{1/2}\epsilon\}} \to 0,$$

in $P_{0,n}$-almost surely. The first equation follows by Assumption 1.4.3 and the second convergence follows by the law of large numbers. The last convergence also follows by the law of large numbers, and hence, the claim follows.

Let $\epsilon > 0$. For simplicity, let $|f| \le 1$. By the expansion of the likelihood ratio, we have

$$\sum_{i=1}^n \log\frac{p^{2|1}_{x_i,\theta_0+n^{-1/2}h}(y_i)}{p^{2|1}_{x_i,\theta_0}(y_i)} = \langle h, Z^{2|1}_n\rangle - \frac{1}{2}\langle h, I_{2|1}(\theta_0)h\rangle + R_n(h)$$

where $R_n(h)$ tends in $P^{1,2}_{0,n}$-probability to 0. Let $\omega_n = 1_{\{|R_n(h)|>\epsilon\}}$ and for some $M > 0$, let $\psi : \mathbf{R}^d \to [0,1]$ be a continuous function such that $\psi(w) = 0$ if $|w| > M$ and $\psi(w) = 1$ if $|w| < M/2$. Then the integrand of (2.1) with respect to $P_{0,n}(dx)$ is

$$\int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy) f(Z^{2|1}_n(x,y)) - \int_{\mathbf{R}^d} G^{2|1}_{0,h}(dw) f(w) = \sum_{i=1}^{6} r^{(i)}_n$$

where $r^{(i)}_n, i = 1, \ldots, 6$ are

$$\begin{aligned}
r^{(1)}_n &= \int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy) f(Z^{2|1}_n(x,y))\omega_n(x,y) \\
r^{(2)}_n &= \int_{\mathcal{Y}_n} P^{2|1}_{x,h,n}(dy) f(Z^{2|1}_n(x,y))\psi(Z^{2|1}_n(x,y))(1-\omega_n(x,y)) \\
r^{(3)}_n &= \int_{\mathcal{Y}_n} P^{2|1}_{x,0,n}(dy) f(Z^{2|1}_n(x,y)) \exp(\langle h, Z^{2|1}_n\rangle - \frac{1}{2}\langle h, I_{2|1}(\theta_0)h\rangle) \\
&\quad \times (\exp(R_n(h)) - 1)(1 - \psi(Z^{2|1}_n))(1-\omega_n) \\
r^{(4)}_n &= -\int_{\mathcal{Y}_n} P^{2|1}_{x,0,n}(dy) f(Z^{2|1}_n(x,y)) \exp(\langle h, Z^{2|1}_n\rangle - \frac{1}{2}\langle h, I_{2|1}(\theta_0)h\rangle)(1-\psi(Z^{2|1}_n))\omega_n \\
r^{(5)}_n &= \int_{\mathcal{Y}_n} P^{2|1}_{x,0,n}(dy) f(Z^{2|1}_n(x,y)) \exp(\langle h, Z^{2|1}_n\rangle - \frac{1}{2}\langle h, I_{2|1}(\theta_0)h\rangle)(1-\psi(Z^{2|1}_n)) \\
&\quad - \int_{\mathbf{R}^d} G^{2|1}_{0,h}(dw) f(w)(1-\psi(w)) \\
r^{(6)}_n &= -\int_{\mathbf{R}^d} G^{2|1}_{0,h}(dw) f(w)\psi(w).
\end{aligned}$$

Since $Z^{2|1}_n$ is tight with respect to $P^{1,2}_{0,n}$, we can choose $C, M > 0$ such that $\limsup_{n\to\infty} \int P_{0,n}(dx)|r^{(2)}_n| \le \epsilon$ and $|r^{(6)}_n| \le \epsilon$. By contiguity, $\int P_{0,n}(dx)(|r^{(1)}_n| + |r^{(4)}_n|) \le C \int P^{2|1}_{0,h,n}(dxdy)\omega_n$ tends to 0, and $|r^{(3)}_n| \le 1 - e^{-\epsilon}$. The convergence of $\int P_{0,n}(dx)|r^{(5)}_n| \to 0$ is from Lindeberg Central Limit Theorem. Then, choose appropriate $\epsilon_n \to 0$ instead of ϵ, the claim follows. □

Assumption 2.3.1. *If $s \neq t$, then $P_s \neq P_t$.*

Assumption 2.3.2. *There exists an integer n and a test $\omega_n = \omega_n(x_1, \ldots, x_n)$ on $(\mathcal{X}^n, \mathcal{A}^n)$, such that there exists a constant $\epsilon_0 \in (0, 1/2)$ and a compact subset K of Θ such that*

$$P^{(n)}_{\theta_0}(\omega_n) \le \epsilon_0, P^{(n)}_{\theta}(1-\omega_n) \le \epsilon_0 \ (\forall \theta \in K^c).$$

Assumption 2.3.3. *The prior distribution Q is absolutely continuous with respect to the Lebesgue measure, and its derivative is continuous, positive and bounded around θ_0.*

Assumption 2.3.4. *For some $\epsilon > 0$, there exists a constant $C > 0$ and for any $s, t \in B_\epsilon(\theta_0)$, we have*

$$H(P^{2|1}_{x,s}, P^{2|1}_{x,t}) \leq M(x)|s-t|,$$

where $M(x) \in L^2(P_{\theta_0})$.

We define the families of probability measure $(F_{x,n}; x \in \mathcal{X}_n)$ and $(F^{1,2}_{x,y,n}; x \in \mathcal{X}_n, y \in \mathcal{Y}_n)$ as follows:

$$F_{x,n}(A) = \frac{\int_{\theta_0+An^{-1/2}} P_s(x)Q(ds)}{\int_\Theta P_t(x)Q(dt)}, \; F^{1,2}_{x,y,n}(A) = \frac{\int_{\theta_0+An^{-1/2}} P^{1,2}_s(x,y)Q(ds)}{\int_\Theta P^{1,2}_t(x,y)Q(dt)}.$$

The following is the Bernstein von-Mises Theorem for independent and identically distributed observations. This version of the Bernstein von-Mises Theorem is proved in [6]. Note that the existence of uniformly consistent test for $(P_\theta; \theta \in \Theta)$ imply the existence of the test for $(P^{1,2}_\theta; \theta \in \Theta)$.

Proposition 2.4 (Bernstein von-Miese's Theorem for i.i.d.). *Under Assumptions 1.4.3 and 2.3.1-2.3.3, Assumptions 2.2.5 and 2.2.6 are satisfied.*

Proposition 2.5. *Under Assumption 2.3.4, Assumption 2.2.7 holds.*

Proof. Using the Hellinger distance, we obtain the following:

$$\begin{aligned}
&\|P^{2|1}_{x,h,n} - P^{2|1}_{x,g,n}\|_{\mathrm{TV}} \\
&= \int |\prod_{i=1}^n p^{2|1}_{x_i,\theta_0+hn^{-1/2}}(y_i) - \prod_{i=1}^n p^{2|1}_{x_i,\theta_0+gn^{-1/2}}(y_i)| \prod_{i=1}^n \mu^{2|1}_{x_i}(dy_i) \\
&\leq 2\Big(\int \Big(\prod_{i=1}^n \sqrt{p^{2|1}_{x_i,\theta_0+hn^{-1/2}}(y_i)} - \prod_{i=1}^n \sqrt{p^{2|1}_{x_i,\theta_0+gn^{-1/2}}(y_i)}\Big)^2 \prod_{i=1}^n \mu^{2|1}_{x_i}(dy_i)\Big)^{1/2} \\
&= 2\Big(2(1 - \int \prod_{i=1}^n \sqrt{p^{2|1}_{x_i,\theta_0+hn^{-1/2}}(y_i)}\sqrt{p^{2|1}_{x_i,\theta_0+gn^{-1/2}}(y_i)} \prod_{i=1}^n \mu^{2|1}_{x_i}(dy_i))\Big)^{1/2} \\
&= 2^{3/2}\Big(1 - \prod_{i=1}^n (1 - \frac{1}{2}H(P^{2|1}_{x_i,\theta_0+hn^{-1/2}}, P^{2|1}_{x_i,\theta_0+gn^{-1/2}})^2)\Big)^{1/2}.
\end{aligned}$$

By Assumption 2.3.4, when $|h-g| < \epsilon$ for some $\epsilon > 0$, the last term is bounded above by the following value:

$$2^{3/2}\Big(1 - \prod_{i=1}^n (1 - \frac{M(x_i)^2}{2n}\epsilon^2)\Big)^{1/2}.$$

Therefore, by the Schwarz inequality, we have

$$\begin{aligned}
&\int_{\mathcal{X}_n} P_{0,n}(dx)2^{3/2}\Big(1-\prod_{i=1}^{n}(1-\frac{M(x_i)^2}{2n}\epsilon^2)\Big)^{1/2}\\
&\leq 2^{3/2}\Big(\int P_{0,n}(dx)\big(1-\prod_{i=1}^{n}(1-\frac{M(x_i)^2}{2n}\epsilon^2)\big)\Big)^{1/2}\\
&\to 2^{3/2}(1-\exp(-P_{\theta_0}(M^2/2)\epsilon^2))^{1/2}.
\end{aligned}$$

Therefore, taking $\epsilon = \epsilon_n \to 0$, Assumption 2.2.7 holds. □

Chapter 3

The Metropolis-Hastings Algorithm for a Fat-tail Target Distribution

3.1 Introduction

Various forms of Markov chain Monte Carlo methods are widely used for simulation of a probability density $p(x)dx$ on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$, and the Metropolis-Hastings algorithms form a popular sub-class of those.

In order to describe the Metropolis-Hastings algorithms for the *target distribution* p, we first consider a *candidate transition kernel* Q which generates potential transition for a discrete time Markov chain. In this paper we will assume that there exists a measurable (in both variables) function $q(x,y)$ such that $Q(x,dy) = q(x,y)dy$.

In the Metropolis-Hastings algorithm, a candidate transition is accepted with probability $\alpha(x,y) = \min\{1, p(y)q(y,x)/(p(x)q(x,y))\}$, otherwise, the jump is rejected and the chain remains its original state. Thus the actual Metropolis-Hastings chain $(M_n^x; n \in \mathbf{N}_0)$ starting from $M_0^x = x$ is defined as follows:

$$\left\{ \begin{array}{lcl} Y_n^x & \sim & q(M_{n-1}^x, y)dy \ (n \in \mathbf{N}) \\ M_n^x & = & \left\{ \begin{array}{ll} Y_n^x & \text{with probability } \alpha(M_{n-1}^x, Y_n^x) \\ M_{n-1}^x & \text{with probability } 1 - \alpha(M_{n-1}^x, Y_n^x). \end{array} \right. \end{array} \right. \tag{3.1}$$

In this paper, we mainly consider two classes of the Metropolis-Hastings algorithms. One is called "random-walk based", in which

$$q(x,y) = q^*(x-y), \tag{3.2}$$

where the q^* is a probability density on $\mathbf{R}^d$. The other is called *Metropolis adjusted Langevin algorithm* or simply, *Langevin algorithm* whose candidate transition kernel is

$$Q(x, dy) \sim \mathrm{N}(x + \frac{1}{2}\langle h\nabla \log p(x)\rangle, h), \tag{3.3}$$

where h is a positive constant, and ∇ denotes the gradient operator. This class is motivated by the Langevin diffusion satisfied by

$$dX_t = dB_t + \frac{1}{2}\nabla \log p(X_t)dt;\ X_0 = x, \tag{3.4}$$

for a Brownian motion $(B_t; t \in \mathbf{R}^+)$. The Langevin algorithm and other Langevin diffusion based algorithms are studied in, for example, [14], [36], [37], [38] and [34].

We are concerned with the rate of convergence of these algorithms for a probability density $p(x)dx$. It is known that the rate of convergence depends on the tail of the distribution $p(x)dx$ (cf. [29], [35]). For example, the tail of p needs to be uniformly exponential for geometric ergodicity for the Metropolis-Hastings algorithms based on random-walk candidate distributions (Theorem 3.3 of [29]). The similar statement was proved in [36] for the Langevin algorithm.

In this paper, we assume that p has heavy tails. The Metropolis-Hastings algorithms when p has heavy tails were studied in, for example, [8] and [10]. A significant step in this direction was made by [17], which served as a basis of the present study. They showed that the random-walk with Gaussian increment based algorithm and the Langevin algorithm converge at the same polynomial rate to p with heavy tails. Moreover, they showed that the convergence rate of a random-walk based algorithm is improved by using a distribution with heavier tails. Their results can be validated for a certain class of probability distribution p. The class of functions they considered consists of p that satisfying

$$p(x) = \frac{l(|x|)}{|x|^\eta}\ (|x| \to \infty), \tag{3.5}$$

with $\eta > d$ where $|\cdot|$ denotes the Euclidean norm and l is a normalized slowly varying function such that $l(x) \to a > 0$ $(x \to \infty)$. Therefore, p should be a symmetric function in the limit. It is not easy to relax the condition. The difficulty comes from the fact that if the target is not symmetric, then the acceptance ratio is difficult to treat.

We show that when p has heavy tails, the behavior of the Langevin algorithm in the tail area in $\mathbf{R}^d$ is similar to that of the Langevin diffusion itself. For this fact, almost all proposal is accepted in the tail area and polynomial rate of convergence of the Langevin algorithm follows from ergodicity of the Langevin diffusion. We do not have to assume technical conditions for p. We only assume that the probability density p is C^2 and

$$\lim_{|x|\to\infty} |\nabla \log p(x)| = 0, \ \lim_{|x|\to\infty} \|\nabla^T \nabla \log p(x)\| = 0,$$

where $\|(a_{i,j})_{i,j=1,\ldots,d}\| = (\sum_{i,j=1}^d a_{i,j}^2)^{1/2}$ and $\nabla^T f(x)$ denotes the Jacobi matrix for the vector $f(x)$. Then we propose an algorithm with transformation, which transform heavy tails of p into lighter tails, and by using that we can improve the convergence rate. The convergence rate is the same for the random-walk based algorithm with heavier increment distribution, which is proposed in [17], though this convergence for the new algorithm is validated for a wider class of target distributions.

In Section 3.2, we formulate central limit theorems for Markov chains with polynomial ergodicity. Those results are used for concrete examples in Section 3.4. In Section 3.3, which is the main part of this chapter, we prove generalized version of a polynomial rate of convergence for the Langevin algorithm. Then we propose an improved algorithm and prove its convergence. In Section 3.4 we demonstrate the efficiency of our methods by numerical calculations.

3.2 Markov Chain and its Polynomial Ergodicity

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(E, \mathcal{E})$ a measurable space where $\mathcal{E}$ is a countably generated σ-algebra. Let $(X_n; n \in \mathbf{N}_0)$ be a discrete time Markov chain having state space $(E, \mathcal{E})$. The transition kernel of $(X_n; n \in \mathbf{N}_0)$ is denoted by P:

$$\mathbf{P}(X_n \in A | X_{n-1}) = P(X_{n-1}, A) \text{ a.s.}$$

This transition kernel P can be interpreted as a linear operator on a function space by defining $Pf(x) = \int P(x, dy)f(y)$. If P_1, P_2 are two kernels, their product P_1P_2 is defined by $(P_1P_2)(x, A) = \int P_1(x, dy)P_2(y, A)$. The iterates P^n is defined by $P^1 = P$ and $P^n = P^{n-1}P$.

Markov chain will be assumed to be irreducible, aperiodic and positive Harris recurrent; for definitions, see [30]. Note that for the Metropolis-Hastings algorithms (3.1), if $p(x)$ and $q(x, y) > 0$ are continuous in both variables, then the Markov chain is $p(x)dx$-irreducible, aperiodic and any compact set of positive Lebesgue measure is a small set (Lemma 1.2 of [29]). Hitting time τ_A of a set $A \in \mathcal{E}$ is defined by $\tau_A = \inf\{n \geq 1; X_n \in A\}$. Hitting times of a petite set play an important role in the ergodicity of Markov chain. A subset $\mathcal{E}^+$ of $\mathcal{E}$ is defined by

$$\mathcal{E}^+ = \{A \in \mathcal{E}; A \text{ has a positive measure by an irreducibility measure}\}.$$

Let $V : E \to \mathbf{R}^+$ be an $\mathcal{E}$-measurable function. Let $\|\cdot\|_V$ be a norm over the space of signed measures on $(E, \mathcal{E})$ to $\mathbf{R}$ defined by

$$\|\nu\|_V := \sup_{|f| \leq V} |\nu(f)| \ (\nu : \text{ signed measure}).$$

When $V \equiv 1$, the norm corresponds to the total variation.

Sub-geometric rate of convergence is studied in, for example, by [40], [9], [17], [18] and [8]. In [18], they proved the following theorem.

Theorem 3.1 (Jarner and Roberts). *Suppose a Markov chain $(X_n; n \in \mathbf{N}_0)$ with transition kernel P is irreducible and aperiodic. Suppose that there exist an $\mathcal{E}$-measurable function $V : E \to [1, \infty)$, constants $c, b > 0$, $0 \leq \gamma < 1$, and a small set C, such that*

$$PV(x) \leq V(x) - cV(x)^\gamma + b1_C(x). \tag{3.6}$$

Then there exists a probability measure Π and the following polynomial convergence property holds for any $x \in E$ where $1 \leq \beta \leq 1/(1-\gamma)$ and $V_\beta(x) = V(x)^{1-\beta(1-\gamma)}$:

$$(n+1)^{\beta-1}\|P^n(x,\cdot) - \Pi\|_{V_\beta} \to 0. \tag{3.7}$$

In particular, $\gamma/(1-\gamma)$ is the polynomial order of convergence in total variation norm.

A central limit theorem is said to hold for f if $\Pi(|f|) < \infty$ and there exists $0 < \sigma^2 < \infty$ such that

$$\frac{S_n(\bar{f})}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2) \ (n \to \infty),$$

where $\bar{f} = f - \Pi(f)$ and $S_n(f) = \sum_{i=1}^n f(X_i)$. We need some lemmas to prove central limit theorems. These lemmas are closely related to Theorem 11.3.9 of [30], Proposition 3.1 of [40] and Theorem 3.2 of [18]. First lemma is merely a modification of Theorem 3.2 of [18].

Lemma 3.1. *Let $A, C \in \mathcal{E}$, $W_i : E \to [1, \infty)$ $(i = 0, 1, \dots, k)$ and $PW_i - W_i \le W_{i+1} + \beta_i 1_C$, $i = 0, 1, \dots, k$. Then for any $l = 0, 1, 2, \dots, k$, we have*

$$\mathbf{E}_x[\sum_{n=0}^{\tau_A-1} \frac{(n+l)!}{n!} W_{l+1}(X_n)] \le l! W_0(x) + \sum_{m=0}^{l} \beta_m \mathbf{E}_x[\sum_{n=0}^{\tau_A-1} \frac{(n+m)!}{n!} 1_C(X_n)]. \tag{3.8}$$

In particular, if $A, C \in \mathcal{E}^+$ and C is a petite set, then there exists a constant $c < \infty$ such that for any $l = 0, 1, 2, \dots, k$, we have

$$\mathbf{E}_x[\sum_{n=0}^{\tau_A-1} \frac{(n+l)!}{n!} W_{l+1}(X_n)] \le l! W_0(x) + c. \tag{3.9}$$

Proof. At the first step, from the assumption $PW_0 - W_0 \le W_1 + \beta_0 1_C$ and using Theorem 11.3.2 of [30], we obtain

$$\mathbf{E}_x[\sum_{n=0}^{\tau_A-1} W_1(X_n)] \le W_0(x) + \beta_0 \mathbf{E}_x[\sum_{n=0}^{\tau_A-1} 1_C(X_n)].$$

At the l-th step, we have

$$\frac{(n+l)!}{n!} PW_l - \frac{(n+l-1)!}{(n-1)!} W_l \le -\frac{(n+l)!}{n!} W_{l+1} + l \frac{(n+l-1)!}{(n-1)!} W_l + \frac{(n+l)!}{n!} \beta_l 1_C.$$

Then using Theorem 11.3.2 of [30], we obtain

$$\mathbf{E}_x[\sum_{n=0}^{\tau_A-1} \frac{(n+l)!}{n!} W_{l+1}(X_n)] \le l \mathbf{E}_x[\sum_{n=0}^{\tau_A-1} \frac{(n+l-1)!}{(n-1)!} W_l(X_n)] + \beta_l \mathbf{E}_x[\sum_{n=0}^{\tau_A-1} \frac{(n+l)!}{n!} 1_C(X_n)]. \tag{3.10}$$

From this fact, the first claim of the lemma can be obtained easily by using induction. Second claim is $\sup_x \mathbf{E}_x[\sum_{n=0}^{\tau_A-1} 1_C(X_n)] < \infty$, which is stated in Theorem 11.3.11 of [30]. □

Lemma 3.2. *Let (P, V, γ, C, b, c) satisfy the drift condition (3.6), $A, C \in \mathcal{E}^+$ and C be a petite set. Then for any $\eta \in (0, 1]$, there exist constants c_1, c_2 such that for any (not necessarily integer) $l \in [0, \eta/(1-\gamma) - 1]$, we have*

$$\mathbf{E}_x[\sum_{n=0}^{\tau_A-1} (n+1)^l V^{\eta-(l+1)(1-\gamma)}(X_n)] \le c_1 V^\eta(x) + c_2. \tag{3.11}$$

In particular, if we take $l = \eta/(1-\gamma) - 1$, then we have

$$\mathbf{E}_x[\tau_A^{l+1}] \leq (l+1)(c_1 V^\eta(x) + c_2). \tag{3.12}$$

Proof. From Lemma 3.5 of [18], for any integer $k \in [0, \eta/(1-\gamma))$, there exist constants c_k, b_k such that $PV^{\eta-k(1-\gamma)} - V^{\eta-k(1-\gamma)} \leq -c_k V^{\eta-(k+1)(1-\gamma)} + b_k 1_C$. Then from the previous lemma, for any integer $l \in [0, \eta/(1-\gamma))$ and for some $c > 0$, we have

$$\mathbf{E}_x[\sum_{n=0}^{\tau_A - 1} \frac{(n+l)!}{n!} V^{\eta-(l+1)(1-\gamma)}(X_n) \prod_{k=0}^{l} c_k] \leq l! V^\eta(x)$$
$$+c.$$

Since $(n+1)^l \leq (n+l)!/n!$ we obtain (3.11) for any integer $l \in [0, \eta/(1-\gamma))$. Next we consider the equation for any real number $l \in [0, \eta/(1-\gamma) - 1]$. For any $t \in [l-1, l)$, we know

$$(\frac{n+1}{V(x)^{(1-\gamma)}})^{l-1} + (\frac{n+1}{V(x)^{(1-\gamma)}})^{l} \geq (\frac{n+1}{V(x)^{(1-\gamma)}})^{t}, \tag{3.13}$$

hence the claim follows. □

In the following theorem, $L^p = L^p(E, \mathcal{E}, \Pi)$ denotes the space of p-power integrable functions f, $\int |f(x)|^p \Pi(dx) < \infty$.

Theorem 3.2. *Let (P, V, γ, C, b, c) satisfy the drift condition (3.6), and $C \in \mathcal{E}^+$ and C is a petite set. Then for any $\eta \geq 1/2$ such that $\Pi(V^{\gamma+2\eta-1})$, for any $\epsilon > (1-\gamma)/(\eta - (1-\gamma))$, a central limit theorem for the Markov chain holds for any f which is in $L^{2+\epsilon}$ or $|f| \leq d\ V^{\gamma+\eta-1}$ where d is a positive constant.*

Proof. First, we show the measure $\lambda(dx) = (\Pi I_{|f|})(dx) = |f(x)|\Pi(dx)$ is $|f|$-regular for any $f \in L^{2+\epsilon}$ where ϵ is in the above range. If the claim holds, then using Theorem 7.6 of [31], this Markov chain has a central limit theorem.

Consider $f \in L^{2+\epsilon}$. For any $A \in \mathcal{E}^+$, using Hölder's inequality, and for any $p, q > 1$ such that $p^{-1} + q^{-1} = 1$, we have

$$\begin{aligned}\mathbf{E}_\lambda[\sum_{n=0}^{\tau_A - 1} |f|(X_n)] &= \sum_{n=0}^{\infty} \mathbf{E}_\Pi[|f|(X_0)|f|(X_n)1_{\{n<\tau_A\}}] \\ &\leq \sum_{n=0}^{\infty} \|f\|_{L^p} (\mathbf{E}_\Pi[|f|(X_0)^q 1_{\{n<\tau_A\}}])^{\frac{1}{q}}.\end{aligned}$$

Since $1_{\{n<\tau_A\}} \leq (\tau_A/n)^r$ for any $r > 1$ and $\beta \in (0,1]$, we have

$$\begin{aligned}\mathbf{E}_\Pi[|f|(X_0)^q 1_{\{n<\tau_A\}}] &\leq \mathbf{E}_\Pi[|f|(X_0)^q \mathbf{E}_\Pi[1_{\{n<\tau_A\}}|\mathcal{F}_0]] \\ &\leq \mathbf{E}_\Pi[|f|(X_0)^q \mathbf{E}_\Pi[(\frac{\tau_A}{n})^{\frac{\beta}{1-\gamma}}|\mathcal{F}_0]] \\ &\leq \beta n^{-\frac{\beta}{1-\gamma}} \mathbf{E}_\Pi[|f|(X_0)^q(c_1 V^\beta(X_0) + c_2)].\end{aligned}$$

Then for any $p', q' > 1$ such that $p'^{-1} + q'^{-1} = 1$, we have

$$\mathbf{E}_\Pi[|f|(X_0)^q V^\beta(X_0)] \leq \|f\|_{L^{p'q}}(\mathbf{E}_\Pi[V^{\beta q'}(X_0))])^{\frac{1}{qq'}}.$$

Sufficient conditions for $\mathbf{E}_\lambda[\sum_{n=0}^{\tau_A-1}|f|(X_n)] < \infty$ are $f \in L^p = L^{p'q}$, $\beta q' = \gamma + 2\eta - 1$ and $q(1-\gamma) < \beta$. We can find β and p, q, p', q' which satisfy the above sufficient conditions for any $f \in L^{2+\epsilon}$. Hence the claim follows.

The case of $|f| \leq d\ V^{\gamma+\eta-1}$ is quite similar. The only difference is the last inequality. In this case we do not have to use Hölder's inequality but the inequality $|f| \leq d\ V^{\gamma+\eta-1}$. □

If a Markov chain is geometrically ergodic, then integrability condition of the drift function like the above is not necessary. However, in sub-geometric case, we need it. Since we know $\Pi(V^\gamma) < \infty$ from the drift condition, $\eta = 1/2$ requires no assumption for the integrability of V. In the case $\eta = 1/2$, central limit theorems for the Markov chain are already showed in Theorem 9 of [19], which uses a mixing theory.

Markov chain is said to be reversible when $\Pi(dx)P(x,dy) = \Pi(dy)P(y,dx)$. Metropolis-Hastings chain is reversible. We can show a slight extension of the above result when the Markov chain is reversible.

Theorem 3.3. *Let (P, V, γ, C, b, c) satisfy the drift condition (3.6), and $C \in \mathcal{E}^+$ and C be a petite set. Further, we assume that the Markov chain is reversible. Then for any $\eta \geq 1/2$ such that $\Pi(V^{\gamma+2\eta-1}) < \infty$, and for $\epsilon = (1-\gamma)/(\eta-(1-\gamma))$, the Markov chain has a central limit theorem for any $f \in L^{2+\epsilon}$.*

Proof. The proof of the theorem uses the same argument as above. Since the Markov chain is reversible, we have

$$\begin{aligned}\mathbf{E}_\lambda[\sum_{n=0}^{\tau_A-1}|f|(X_n)] &\leq \sum_{n=0}^{\infty}\mathbf{E}_\Pi[1_{\{n<\tau_A\}}|f|(X_0)^2]^{\frac{1}{2}}\mathbf{E}_\Pi[1_{\{n<\tau_A\}}|f|(X_n)^2]^{\frac{1}{2}} \\ &= \sum_{n=0}^{\infty}\mathbf{E}_\Pi[1_{\{n<\tau_A\}}|f|(X_0)^2] = \mathbf{E}_\Pi[\tau_A|f|(X_0)^2].\end{aligned}$$

Using Lemma 3.2, and Schwarz's inequality, the claim follows. □

3.3 Algorithm and Main Theorems

3.3.1 Langevin Algorithms

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ be a filtered probability space. Let $p : \mathbf{R}^d \to \mathbf{R}$ be a strictly positive C^1 function and consider the stochastic differential equation (3.4). Under certain conditions, there exists a unique solution to the stochastic differential equation and the solution has an invariant measure $p(x)dx$. Let $(Y^x_n; n \in \mathbf{N}_0)$ be an Euler-Maruyama discretization of $(X^x_t; t \in \mathbf{R}^+)$, that is,

$$Y^x_n = \sqrt{h}W_n + hb(Y^x_{n-1});\ Y^x_0 = x, \tag{3.14}$$

where $W_n := h^{-1/2}(B_{hn} - B_{h(n-1)})$. In [36], they proved that if $|\nabla \log p(x)| \to 0$ $(|x| \to \infty)$ then the Langevin algorithm does not converge at geometric rate (Theorem 4.2 of [36]). We are going to prove its polynomial rate of convergence. First, we show polynomial ergodicity for this Markov chain, the candidate chain of the Langevin algorithm.

Theorem 3.4. *Let $p : \mathbf{R}^d \to \mathbf{R}$ be a C^1 function. Suppose there exists $\eta > d$, such that*

$$\limsup_{|x|\to\infty} \langle x, \nabla \log p(x)\rangle \leq -\eta, \quad \lim_{|x|\to\infty} |\nabla \log p(x)| = 0. \tag{3.15}$$

Then the Euler-Maruyama discretization $(Y^x_n; n \in \mathbf{N}_0)$ satisfies the drift condition (3.6) for any $h > 0$, $2 < s < 2 + \eta - d$, $V(x) = (|x|^2 + 1)^{s/2}$, $\gamma = (s-2)/s$ and a compact set C of positive Lebesgue measure. In particular, the upper bound of the polynomial convergence rate of the total variation norm is $(\eta - d)/2$.

Proof. It is enough to show

$$\limsup_{|x|\to\infty} \frac{PV(x) - V(x)}{V(x)^\gamma} < 0, \tag{3.16}$$

since $C = \{|x| \leq N\}$ is a small set for any $N > 0$. Let $(X^x_t, t \in [0,1])$ be a stochastic process satisfying $dX^x_t = dB_t + b(x)dt$, where B_t is a standard

Brownian motion. Then $\mathcal{L}(X_h) = \mathcal{L}(Y_1^x)$ and

$$\begin{aligned} PV(x) - V(x) &= \mathbf{E}[V(X_h^x) - V(x)] \\ &= \mathbf{E}[\int_0^h \sum_{i=1}^d \frac{\partial V}{\partial x_i}(X_t^x)dX_t^{x,i} + \frac{1}{2}\frac{\partial^2 V}{\partial x_i \partial x_j}(X_t^x)d\langle X^{x,i}, X^{x,j}\rangle_t] \\ &= \frac{sh}{2}\mathbf{E}[\int_0^h (|X_t^x|^2+1)^{\frac{s}{2}-1}(2\sum_{i=1}^d X_t^{x,i}b^i(x) + s - 2 + d) \\ &\quad -(|X_t^x|^2+1)^{\frac{s}{2}-2}dt]. \end{aligned}$$

Since $X_t^x = x + B_t + tb(x)$, after some calculations such as $\limsup \mathbf{E}[(|X_t|^2 + 1)^n] \cdot |x|^{-2n} \leq 1$, we have

$$\begin{aligned} \limsup_{|x|\to\infty} \frac{PV(x) - V(x)}{V(x)^\gamma} &= \limsup_{|x|\to\infty} \frac{sh}{2}(2\sum_{i=1}^d x_i b^i(x) + s - 2 + d) \\ &\leq \frac{sh}{2}(-\eta + s - 2 + d). \end{aligned}$$

When $2 < s < 2 + \eta - d$, $\limsup_{|x|\to\infty}(PV(x) - V(x))/V(x)^\gamma < 0$ by the above inequality. □

In [10], they have already addressed polynomial ergodicity of a tempered Langevin diffusion. Their results are more general than our results though they do not consider Markov chains but continuous stochastic processes. Roughly speaking, our theorem corresponds to the discretization of Theorem 16 of [10] when a parameter $d = 0$ in a sense of the rate of convergence in $\|\|_f$-norm.

Next we show the convergence of the Langevin algorithm. Let $(M_n^x; n \in \mathbf{N}_0)$ be the Metropolis-Hastings chain of the Langevin algorithm starting from $M_0^x = x$, that is,

$$\left\{ \begin{array}{lcl} Y_n^x &=& M_{n-1}^x + \sqrt{h}W_n + hb(M_{n-1}^x) \\ M_n^x &=& \left\{ \begin{array}{ll} Y_n^x & \text{with probability } \alpha(M_{n-1}^x, Y_n^x) \\ M_{n-1}^x & \text{with probability } 1 - \alpha(M_{n-1}^x, Y_n^x). \end{array} \right. \end{array} \right. \tag{3.17}$$

where $b = \nabla \log p(x)/2$ and $q(x,y)$ is the density of the transition kernel (3.3), that is,

$$q(x,y) = \frac{1}{(2h\pi)^{\frac{d}{2}}} \exp(-\frac{|y - x - hb(x)|^2}{2h}). \tag{3.18}$$

This Langevin algorithm does not have geometrical ergodicity but polynomial ergodicity.

Theorem 3.5. *Let $p : \mathbf{R}^d \to \mathbf{R}$ be a C^2 function satisfying (3.15) and*

$$\lim_{|x|\to\infty} \|\nabla^T \nabla \log p(x)\| = 0. \tag{3.19}$$

Then the Metropolis-Hastings chain of the Langevin algorithm $(M_n^x; n \in \mathbf{N}_0)$ satisfies the drift condition (3.6) for $2 < s < 2+\eta-d$, $V(x) = (|x|^2+1)^{s/2}$, $\gamma = (s-2)/s$ and a compact set C. In particular, the upper bound of the polynomial convergence rate for the total variation norm is $(\eta-d)/2$.

Proof. We know by Theorem 3.4, there exist constants $c < 1, b > 0$ and a compact set C of positive Lebesgue measure satisfying

$$\begin{aligned} PV(x) &= \mathbf{E}[V(X_h^x)\alpha(x,X_h^x)] + \mathbf{E}[V(x)(1-\alpha(x,X_h^x))] \\ &= \mathbf{E}[V(X_h^x)] - \mathbf{E}[(V(X_h^x)-V(x))(1-\alpha(x,X_h^x))] \\ &\le V(x) - cV(x)^\gamma + b1_C(x) - \mathbf{E}[(V(X_h^x)-V(x))(1-\alpha(x,X_h^x))], \end{aligned}$$

where $dX_t^x = dB_t + b(x)dt$. Hence it is enough to show $\lim_{|x|\to\infty} |\mathbf{E}[(V(Y_1^x) - V(x))(1-\alpha(x,X_h^x))]|/V(x)^\gamma = 0$ when $\gamma = (s-2)/s$. By Schwarz's inequality,

$$\mathbf{E}[(V(X_h^x) - V(x))(1-\alpha(x,X_h^x))]| \le \mathbf{E}[(V(X_h^x)-V(x))^2]^{\frac{1}{2}} \mathbf{E}[(1-\alpha(x,X_h^x))^2]^{\frac{1}{2}}$$

Since the first term, $\limsup \mathbf{E}[(V(X_h^x) - V(x))^2]^{\frac{1}{2}} V(x)^{-\gamma} \le 1$, we will check $\lim \mathbf{E}[(1-\alpha(x,X_h^x))^2]^{\frac{1}{2}} = 0$. Let $\beta(x,y) = p(y)q(y,x)/(p(x)q(x,y))$, then

$$\begin{aligned} \mathbf{E}[(1-\alpha(x,X_h^x))^2] &\le \mathbf{E}[(1-\beta(x,X_h^x))^2] \le \mathbf{E}[\log \beta(x,X_h^x)^2] \\ &= \mathbf{E}[(\log p(X_h^x) - \log p(x) + \log q(X_h^x, x) - \log q(x, X_h^x))^2] \\ &= \mathbf{E}[(\log p(X_h^x) - \log p(x) - (b(x) + b(X_h^x))^T (X_h^x - x) \\ &\quad -\frac{h}{2}(|b(X_h^x)|^2 - |b(x)|^2))^2]. \end{aligned}$$

It is easy to check $\lim \mathbf{E}[((b(x) - b(X_h^x))^T (X_h^x - x))^2] = 0$ and $\lim \mathbf{E}[(|b(X_h^x)|^2 - |b(x)|^2)^2] = 0$. The remainder of the above is

$$\begin{aligned} &\mathbf{E}[(\log p(X_h^x) - \log p(x) - 2b(x)^T (X_h^x - x))^2] = \\ &\mathbf{E}[(\int_0^h \sum_{i=1}^d (\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x))dX_t^{x,i} + \frac{1}{2}\frac{\partial^2 \log p}{\partial x_i^2}(X_t^x)dt)^2], \end{aligned}$$

and the main part of the above equation is

$$\mathbf{E}[(\int_0^h \sum_{i=1}^d (\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x))dW_t^i)^2] =$$
$$\mathbf{E}[\int_0^h \sum_{i=1}^d (\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x))^2 dt]$$
$$\leq \mathbf{E}[\sup_{0\leq t\leq h} \sum_{i=1}^d (\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x))^2].$$

We want to show the last term in the above goes to 0 if $|x| \to \infty$. For any $\epsilon > 0$, there exist $\delta_1, \delta_2, \delta_3 > 0$ such that

$$\|\nabla^T \nabla \log p(x)\|^2 < \frac{\epsilon}{2d^3 C_h} \ (|x| \geq \delta_1)$$
$$h|b(x)| \leq \delta_1 \wedge 1 \ (|x| \geq \delta_2)$$
$$\sup_{\xi \in \mathbf{R}^d} |\nabla \log p(\xi)|^2 4d\mathbf{P}(\sup_{0\leq t\leq h} |B_t| > \delta_3) < \frac{\epsilon}{2},$$

where $C_h = \mathbf{E}[\sup_{0\leq t\leq h}(|W_t|+1)^2]$ which is a bounded constant by Doob's inequality. Let $|x| > \delta_1 + \delta_2 + \delta_3$, then we divide the term into two parts,

$$\mathbf{E}[\sup_{0\leq t\leq h} \sum_{i=1}^d (\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x))^2 (1_{\{\sup_{0\leq t\leq h}|B_t|>\delta_3\}} + 1_{\{\sup_{0\leq t\leq h}|B_t|\leq\delta_3\}})].$$

The first term is bounded above by $4d \sup_{\xi\in\mathbf{R}^d} \|\nabla \log p(\xi)\|^2 \mathbf{P}(\sup_{0\leq t\leq h}|B_t| > \delta_3) \leq \epsilon/2$. By a Taylor expansion, the second term is bounded above by

$$\begin{aligned}\mathbf{E}[\sup_{0\leq t\leq h} \sum_{i=1}^d (\sum_{j=1}^d \sup_{|\xi|>\delta_1} |\frac{\partial^2 \log p}{\partial x_i \partial x_j}(\xi)||X_t^x - x|)^2] &\leq \frac{\epsilon}{2C_h}\mathbf{E}[\sup_{0\leq t\leq h}|X_t^x - x|^2] \\ &\leq \frac{\epsilon}{2C_h}\mathbf{E}[\sup_{0\leq t\leq h}(|W_t| + t|b(x)|)^2] \leq \frac{\epsilon}{2}.\end{aligned}$$

Hence $\mathbf{E}[\sup_{0\leq t\leq h}\sum_{i=1}^d(\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x))^2]$ goes to 0. □

When $d = 1$ and the target distribution can be written in the form $p(x) = C|x|^{-\eta}$ when $|x|$ is large enough, [17] have already proved the same result. Moreover the proof of [17] is the basis of the proof of Theorem 3.5, though the assumptions of Theorem 3.5 is more general.

When $d > 1$, [17] have also proved that the random-walk based Metropolis-Hastings algorithms have the same order of convergence as the Langevin algorithm when the increment distributions of the random-walks have light tails (Proposition 3.5 in [17]). The random-walk based algorithms are simpler than the Langevin algorithm in the sense of computer calculation, it is better to use the former algorithms if the convergence theorem can be validated for a wide class of target distributions p. However their results for random walk based algorithms are validated for a smaller class of target distributions. They assumed our assumptions and a roundness property about $A(x) = \{y; p(x) \leq p(y)\}$ and $A(x)$ should be a convex set when $|x|$ is large enough in their paper. For example, the two-dimensional probability distribution function $p(x, y) \propto (x^4+y^2+1)^{-1}$ does not satisfy the extra properties. This distribution function satisfies (3.15), (3.19) and $\eta = 2$, but $A(x)$ is not a convex set. In fact, the distribution satisfies $\limsup |x| \cdot |\nabla \log \pi(x)| < \infty$ and $\limsup |x|^2 \cdot \|\nabla^T \nabla \log \pi(x)\| < \infty$.

Many probability distributions which have heavy tails satisfy property (3.15), (3.19). For example, Student's t distribution satisfies the properties.

Example 3.1 (Multivariate Student's t distribution). *Consider following d-dimensional Student's t distribution with $m > 0$ degrees of freedom,*

$$p(x) = \frac{\Gamma(\frac{m+d}{2})}{\Gamma(\frac{m}{2})(m\pi)^{\frac{d}{2}}}(\det \Sigma)^{-\frac{1}{2}}(1 + \frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{m})^{-\frac{m+d}{2}}. \tag{3.20}$$

It satisfies $\limsup_{|x|\to\infty} |x| \cdot |\nabla \log p(x)| < \infty$, $\limsup_{|x|\to\infty} x^T \cdot \nabla \log p(x) \leq -(m+d)$ and $\limsup_{|x|\to\infty} |x|^2 \|\nabla^T \nabla \log p(x)\| < \infty$. The proof uses the fact that for the positive definite matrix Σ, there exists $\lambda > 0$ such that $\lambda |x|^2 \leq x^T \Sigma^{-1} x$. By Theorem 3.3, the Langevin algorithm with proposal p has a central limit theorem for $L^{2+\epsilon}$ with $\epsilon > 4/(m-2)$.

Example 3.2 (An example which does not satisfy (3.15)). *Consider the following probability distribution function:*

$$p(x) \propto \prod_{i=1}^{d} \frac{1}{1+x_i^2} \quad (x = (x_1, \ldots, x_d) \in \mathbf{R}^d).$$

This function satisfy the left hand side of (3.15) but right hand side of it. Since

$$|\nabla \log p(x)| = (\sum_{i=1}^{d} (\frac{2x_i}{1+x_i^2})^2)^{\frac{1}{2}},$$

if we take $x = (0, t, \ldots, t)$ and $t \to \infty$, then $|\nabla \log p(x)| \to 2$.

3.3.2 Transformed Langevin Algorithm

We introduce a transformation of a Markov chain $(M_n^x; n \in \mathbf{N}_0)$ by a function $F : \mathbf{R}^d \to \mathbf{R}^d$. Suppose there is a C^2 function $f : \mathbf{R} \to \mathbf{R}$ which holds $f'(x) > 0$, $f(0) = 0$ and $\lim_{x \to 0} f(x)/x \neq 0$ such that

$$F(x) = \begin{cases} f(|x|)\frac{x}{|x|} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

Then F is a C^2 function with $\det \nabla^T F(x) > 0$. Under certain conditions, if a Markov chain $(M_n^x; n \in \mathbf{N}_0)$ with an invariant measure $p^*(x)dx := p(F(x)) \det \nabla^T F(x) dx$ satisfies (3.6), then $(F(M_n^x); n \in \mathbf{N}_0)$ has an invariant measure $p(x)dx$ and satisfies (3.6) (Proposition 3.1).

Let p be a d-dimensional probability distribution function and and $V : \mathbf{R}^d \to [0, \infty)$ be a norm-like function, that is, for any $r > 0$, $\{x; V(x) \leq r\}$ is a relatively compact set.

Proposition 3.1. *Let $|F^{-1}(x)|$ be a norm-like function. Let $p(x) > 0$ be a C^1 function and $Q^*(x, dy) = q^*(x, y)dy$ be a transition kernel where $q^*(x, y) > 0$ is continuous in both variables. Let $(M_n^*; n \in \mathbf{N}_0)$ be a Metropolis-Hastings chain with a candidate kernel Q^* and an invariant measure $p^*(x)dx$. Suppose there exist a compact set C^* with positive Lebesgue measure, a function $V^* : \mathbf{R}^d \to [1, \infty)$ and constants $0 \leq \gamma \leq 1$, $b, c > 0$ such that the drift condition (3.6) holds. Then for $(M_n = F(M_n^*); n \in \mathbf{N}_0)$, there exist constants γ, b, c, a compact set $C \supset C^*$ with positive Lebesgue measure such that the drift condition (3.6) for C, $V = V^* \circ F^{-1}$.*

Proof. Denote the transition kernel of $(M_n^*; n \in \mathbf{N}_0)$ by P^* and that of $(M_n; n \in \mathbf{N}_0)$ by P. First, we show that $(M_n; n \in \mathbf{N}_0)$ is a Metropolis-Hastings chain with the candidate kernel

$$Q(x, dy) = q(x, y)dy := q^*(F^{-1}(x), F^{-1}(y)) \det \nabla^T F^{-1}(y) dy,$$

and the invariant probability measure $p(x)dx$. Let $(Y_n^*; n \in \mathbf{N}_0)$ be a candidate chain of $(M_n^*; n \in \mathbf{N}_0)$ and denote the acceptance ratio for the Metropolis-Hastings chain by α^*. Let $Y_n := F(Y_n^*)$, then

$$\mathbf{P}(Y_n \in A | Y_{n-1}) = \int_{F^{-1}(A)} q^*(Y_{n-1}^*, y)dy = \int_A q(Y_{n-1}, y)dy,$$

hence Q is its transition kernel. Let $\alpha(x, y) = 1 \wedge p(y)q(y, x)/(p(x)q(x, y))$ then

$$\alpha^*(x, y) = 1 \wedge \frac{p^*(y)q^*(y, x)}{p^*(x)q^*(x, y)} = 1 \wedge \frac{p(F(y))q(F(y), F(x))}{p(F(x))q(F(x), F(y))} = \alpha(F(x), F(y)),$$

and it proves the first claim. Because q is strictly positive and continuous in both variables by its definition, $(M_n; n \in \mathbf{N}_0)$ is irreducible and any compact set with positive Lebesgue measure is a small set by Lemma 1.2 of [29]. By the conditions

$$\begin{aligned} P^*V^* - V^* \leq cV^{*\gamma} + b1_{C^*} \quad &\Rightarrow \quad P(V \circ F) - V \circ F \leq c(V \circ F)^\gamma + b1_{C^*} \\ &\Rightarrow \quad PV - V \leq cV^\gamma + b1_{C^*}(F^{-1}(x)) \ (x \in \mathbf{R}^d). \end{aligned}$$

Since C^* is a compact set, there is $r > 0$ such that $C^* \subset \{|x| \leq r\}$, then $\{F^{-1}(x) \in C^*\} \subset \{|F^{-1}(x)| \leq r\}$ and if we take $C = \overline{\{|F^{-1}(x)| \leq r\}}$, then C is a compact set since $|F^{-1}|$ is a norm-like function. We can take C large enough to have positive Lebesgue measure, hence C is a small set. Then for C, V, γ, b, c, the drift condition (3.6) holds. □

We take $f(x) = x^{2/(2-r)}$ $(x > 1)$ and set properly to satisfy above conditions when $x \leq 1$. When $|x| > 1$, $\nabla^T F(x) = (I_d + r/(2-r)x \cdot x^T/|x|^2)|x|^{r/(2-r)}$ and $\det \nabla^T F(x) = (2/(2-r))|x|^{dr/(2-r)}$. When $(M_n^x; n \in \mathbf{N}_0)$ is from the Langevin algorithm we call this transform algorithm, the transformed Langevin algorithm.

For practical purpose, it is convenient to take $f(x) \equiv x (x \leq 1)$ and it is enough to establish the following conclusion, though it does not a C^2 function. We restrict f to be a C^2 function in our proof since it simplifies our proof.

Theorem 3.6. *Let p be a C^2 function that satisfies*

$$\limsup_{|x| \to \infty} \langle x, \nabla \log p(x) \rangle \leq -\eta, \tag{3.21}$$

$$\lim_{|x| \to \infty} |x|^{\frac{r}{2}} \cdot |\nabla \log p(x)| = 0, \tag{3.22}$$

$$\lim_{|x| \to \infty} |x|^r \cdot \|\nabla^T \nabla \log p(x)\| = 0. \tag{3.23}$$

Consider the Transformed Langevin algorithm by F when $0 \leq r < 2$. Then the drift condition (3.6) holds for $2 < s < 2 + (\eta - d)(2/(2-r))$, $V(x) = (|F^{-1}(x)|^2 + 1)^{s/2}$ and $\gamma = (s-2)/s$. In particular, the upper bound of the polynomial order of convergence in total variation norm is $(\eta - d)/(2-r)$.

Proof. If p^* satisfies the properties (3.15) and (3.19), by using Theorem 3.5 for $(M_n^x; n \in \mathbf{N}_0)$, the claim follows by Proposition 3.1. Through the proof, we assume $|x| > 1$. By the definition of p^*, $\nabla \log p^*(x) = (\nabla^T F(x))^T \cdot (\nabla \log p)(F(x)) + \nabla \log \det \nabla^T F(x)$. Because $x^T \cdot \nabla^T F(x) =$

$(2/(2-r))F(x)^T$, $\nabla \log \det \nabla^T F(x) = (dr/(2-r))x/|x|^2$ and $\|\nabla^T F(x)\| \cdot |x| \leq (d^{1/2} + r/(2-r))|F(x)|$, we obtain

$$\begin{aligned} x^T \cdot \nabla \log p^*(x) &= \frac{2}{2-r} F(x)^T \cdot \nabla \log p(F(x)) + \frac{dr}{2-r}, \\ |x| \cdot |\nabla \log p^*(x)| &\leq (d^{\frac{1}{2}} + \frac{r}{2-r})|F(x)| \cdot |\nabla \log p(F(x))| + \frac{dr}{2-r}. \end{aligned}$$

Let $\eta^* = (1 - r/2)^{-1}(\eta - rd/2)$, then the following properties hold since $|F(x)|^{r/2} = |F(x)|/|x|$:

$$\limsup_{|x|\to\infty} x^T \cdot \nabla \log p^*(x) \leq -\eta^*, \quad \lim_{|x|\to\infty} |x|^{\frac{r}{2}} |\nabla \log p^*(x)| = 0.$$

Next we show that $\|\nabla^T \nabla \log p^*(x)\|$ goes to 0 in the limit. We take some steps to calculate it. In the following calculations, we sometimes drop the operator "$\cdot$" and the state x to simplify the inequalities and equations. First, divide $\|\nabla^T \nabla \log p^*(x)\|$ into two parts, $\|\nabla^T (\nabla^T F(x))^T \nabla \log p(F(x))\|$ and $\|\nabla^T \nabla \log \det \nabla^T F(x)\|$. About the second term, it is easy to see

$$|x|^2 \|\nabla^T \nabla \log \det \nabla^T F(x)\| \leq (dr(2-r))(d^{1/2} + 2).$$

Now consider the first term. We have

$$\begin{aligned} \nabla^T((\nabla^T F(x)^T) \cdot \nabla \log p(F(x))) &= \nabla^T(|x|^{\frac{r}{2-r}} \nabla \log p(F(x))) \\ &+ \frac{r}{2-r} \nabla^T(|x|^{\frac{r}{2-r}-2} x \cdot x^T \cdot \nabla \log p(F(x))). \qquad (3.24) \end{aligned}$$

Then the first term in the above is $\nabla \log p(F) \nabla^T |x|^{r/(2-r)} + |x|^{r/(2-r)} \nabla^T (\nabla \log p(F))$. The norm of the first term in it is smaller than $(r/(2-r))|F(x)| \cdot |x|^{-2} \cdot |\nabla \log p(F(x))|$ and the second term is

$$\begin{aligned} \||x|^{\frac{r}{2-r}} \nabla^T(\nabla \log p(F(x)))\| &\leq |x|^{\frac{r}{2-r}} \|\nabla^T \nabla \log p(F(x)) \nabla^T F(x)\| \\ &\leq (d^{\frac{1}{2}} + \frac{r}{2-r})|F(x)|^{\frac{r}{2}} |\nabla^T \nabla \log p(F(x))|, \end{aligned}$$

hence both of them converge to 0. Finally, we show that the norm of the second term in (3.24) goes to 0 in the limit. We write

$$\begin{aligned} A := \nabla^T(|x|^{\frac{r}{2-r}-2} x x^T \nabla \log p(F(x))) &= x x^T \nabla \log p(F(x)) \nabla^T |x|^{\frac{r}{2-r}-2} \\ &+ |x|^{\frac{r}{2-r}-2} \nabla^T (x x^T \nabla \log p(F(x))). \end{aligned}$$

Since

$$\begin{aligned} \nabla^T(x x^T \nabla \log p(F(x))) &= x x^T (\nabla^T (\nabla \log p(F(x))) \\ &+ x^T \nabla \log p(F(x)) I_d + \nabla \log p(F(x))^T x, \end{aligned}$$

we obtain

$$\begin{array}{rcl} |x|^2\|A\| & \leq & |x|^4|\nabla \log p(F)| \cdot |\nabla^T |x|^{\frac{r}{2-r}-2}| \\ & & +|x|^{\frac{r}{2-r}-2}(|x|^4\|\nabla^T\nabla \log p(F))\| \cdot \|\nabla^T F\| + |x|^3|\nabla \log p(F)|(d^{1/2}+1)) \\ & \leq & (\frac{r}{2-r} + d^{\frac{1}{2}} - 1)|F| \cdot |\nabla \log p(F)| + (d^{\frac{1}{2}} + \frac{r}{2-r})|F|^2\|\nabla^T\nabla \log p(F)\|. \end{array}$$

Then by (3.22) and (3.23), $\|A\|$ converges to 0. □

As we showed in Example 3.1, the Langevin algorithm with m degree of freedom Student's t proposal distribution has a central limit theorem for $L^{2+\epsilon}$ with $\epsilon > 4/(m-2)$. On the other hand, transformed chain has a central limit theorem for $L^{2+\epsilon}$ with $\epsilon > 2(2-r)/(m-(2-r))$.

In [17], they proved the same kind of improvements of the rate of convergence in another way. We transformed the chain to gain the heaviness of the tail. On the other hand, they weighted q^* of the transition kernel $Q(x, dy) = q^*(|x-y|)dy$. They took q^* as a probability distribution function of Student's t distributions instead of normal distributions. However they supposed stronger conditions, which is described in (3.5).

We can transform the random-walks based Metropolis-Hastings algorithm instead of the Langevin algorithm as Theorem 3.6. However we cannot prove the improvements like this theorem without some extra conditions, for example, $A(x) = \{p(y) \geq p(x)\}$ should be a convex set. I cannot make out whether these difficulties are avoidable or are essential problems for the schemes.

3.4 Calculation

We now check the performance of the Metropolis-Hastings algorithms. In practice, we should choose good parameters. As stated in the previous subsection, we use f as $f(x) \equiv x$ $(x \leq 1)$.

Example 3.3 (Multivariate t distribution). *Consider the multivariate t distribution (3.20) with the degree of freedom* $m = 3$, *mean* $\mu = (2, 2)^T$ *and*

$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}.$$

Start point $X_0 = (2, 3)$. *We produced* $M = 100,000$ *parallel Markov chains* $(X_n^m = (X_{1,n}^m, X_{2,n}^m)^T; n \in \mathbf{N}_0)$ $(m = 1, \ldots, M)$ *by four algorithms below for* *each and calculated mean squared error*

$$\mathrm{MSE}_{N,M} = \sum_{m=1}^{M} \frac{(\sum_{n=1}^{N} g(X_n^m) - \Pi(g))^2}{N * M}. \tag{3.25}$$

We took $N = 500, 1000, 2500$ *for each and* $g(x, y) = x$. *We consider the following algorithms:*

- *Random-walk with Gaussian increment distribution based algorithm. (Table 3.1)*
- *Langevin algorithm. (Table 3.2)*
- *Random-walk with Student's t increment distribution (degree of freedom is* 1*) based algorithm. (Table 3.3)*
- *Transformed Langevin algorithm (*$r = 1$*). (Table 3.4)*
- *Transformed Langevin algorithm (*$r = 1.2$*). (Table 3.5)*

These algorithms have central limit theorems for $L^{2+\epsilon}$ *by Theorem 3.3, where the value of* ϵ *differs as follows:* $\epsilon > 4$ *for the first and second algorithms,* $\epsilon > 1$ *for the third and fourth, and* $\epsilon > 8/11$ *for the last one. Therefore in this case, Markov chain produced by the last algorithm have a central limit theorem, but we can not say anything about the others using Theorem 3.3.*

In Tables 3.1 through 3.5, transformed algorithm $r = 1.2$ *works well in this case. However you should choose good parameters to obtain such an improvement. When* $r = 1.2$, *the algorithm behaves badly for* $h = 10$.

Example 3.4. *The following example is anti-convex probability distribution:*

$$p(x, y) \propto \frac{1}{(x^4 + y^2 + 1)^3}. \tag{3.26}$$

In this example, $\eta = 6$.

We consider the following algorithms:

- *Random-walk with Gaussian increment distribution based algorithm. (Table 3.6)*
- *Langevin algorithm. (Table 3.7)*
- *Random-walk with Student's t increment distribution (degree of freedom is* 1*) based algorithm. (Table 3.8)*

- *Transformed Langevin algorithm ($r = 1$). (Table 3.9)*

Some of these algorithms have central limit theorems for $L^{2+\epsilon}$ where the value of ϵ differs as follows: $\epsilon > 8/5$ for the second algorithm, $\epsilon > 4/7$ for the last one. Since this probability distribution is not symmetric, we do not know whether other algorithms have a central limit theorem.

In Tables 3.6 through 3.9, we used the same starting point X_0 and the same number of parallel Markov chains M as the previous example. The first algorithm is not so bad and the second algorithm is better than the last one. Transformation does not always show improvements.

3.5 Conclusion

The purpose of this paper is to introduce the Metropolis-Hastings algorithms that can deal with a wide class of heavy-tailed target distributions. We proved the convergence rate and sufficient conditions for convergence for these algorithms. The transformed algorithm is of the same rate of convergence as the heavy-tailed proposal random-walk algorithm, though the latter algorithm needs strong assumptions for the target.

Next, we want to prove the differences between the random-walk with Gaussian increment distribution based algorithm and the Langevin algorithm. Numerical calculation suggests that the asymptotic variance of the estimator $\widehat{\Pi(f)}_N = N^{-1}\sum_{n=1}^{N} f(M_n^x)$ of the Langevin algorithm is smaller than that of the random-walk based algorithm when the target distribution is not symmetric. Therefore symmetricity seems to be an important condition for the latter algorithm.

Table 3.1: Example 3.3: Random-walk with Gaussian increment distribution based algorithm.

	h=30	h=40	h=50
N=500	111.70	95.69	90.83
N=1000	100.23	101.83	105.98
N=2500	159.91	267.78	127.22

Table 3.2: Example 3.3: Langevin algorithm.

	h=30	h=40	h=50
N=500	113.62	121.6	123.18
N=1000	157.27	120.72	137.58
N=2500	165.95	143.11	172.00

Table 3.3: Example 3.3: Random-walk with Student's t increment distribution (degree of freedom is 1) based algorithm.

	h=8	h=10	h=12
N=500	138.55	134.16	129.60
N=1000	139.53	143.13	144.25
N=2500	147.98	144.21	145.02

Table 3.4: Example 3.3: Transformed Langevin algorithm by $r = 1$.

	h=1	h=2	h=3	h=10
N=500	45.65	33.59	45.93	457.5
N=1000	48.24	35.48	48.67	903.19
N=2500	49.14	40.15	45.83	2149.07

Table 3.5: Example 3.3: Transformed Langevin algorithm by $r = 1.2$.

	h=1	h=1.2	h=1.4
N=500	41.71	41.02	43.61
N=1000	42.59	41.87	43.28
N=2500	42.91	44.08	42.54

Table 3.6: Example 3.4: Random-walk with Gaussian increment distribution based algorithm.

	h=0.5	h=1	h=1.5
N=500	2.121	2.024	2.246
N=1000	2.109	2.038	2.256
N=2500	2.124	2.025	2.258

Table 3.7: Example 3.4: Langevin algorithm.

	h=0.25	h=0.50	h=0.75
N=500	1.064	0.569	0.688
N=1000	1.070	0.566	0.689
N=2500	1.074	0.570	0.688

Table 3.8: Example 3.4: Random-walk with Student's t increment distribution (degree of freedom is 1) based algorithm.

	h=0.05	h=0.1	h=0.2
N=500	4.906	4.782	4.817
N=1000	4.943	4.821	4.875
N=2500	5.018	4.851	4.931

Table 3.9: Example 3.4: Transformed Langevin algorithm by $r = 1$.

	h=0.06	h=0.08	h=0.10
N=500	1.309	1.222	1.324
N=1000	1.312	1.222	1.325
N=2500	1.328	1.226	1.328

Bibliography

[1] P. J. Bickel and J. A. Yahav. Some contributions to the asymptotic theory of Bayes solutions. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 11:257–276, 1969.

[2] Patrick Billingsley. *Probability and measure.* A Wiley-Interscience Publication., 3 edition, 1995.

[3] J. Borwanker, G. Kallianpur, and B. L. S. Prakasa Rao. The Bernstein-von Mises theorem for Markov processes. *Annals of Mathematical Statistics*, 42:1241–1253, 1971.

[4] Russell A. Boyles. On the Convergence of the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):47–50, 1983.

[5] Lucien Le Cam. On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 129–156. University of California Press, Berkeley and Los Angeles, 1956.

[6] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics : some basic concepts.* New York ; Tokyo : Springer-Verlag, 2nd edition, 2000.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[8] Fort G. Douc, R and E. Moulines. Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability*, 14:1353–1377, 2004.

[9] G. Fort and E. Moulines. V-subgeometric ergodicity for a Hastings-Metropolis algorithm. *Statistics Probability Letters*, 49:401–410, 2000.

[10] G. Fort and G. O. Roberts. Subgeometric ergodicity of strong Markov processes. *The Annals of Applied Probability*, 15:1565–1589, 2005.

[11] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398, 1990.

[12] James E. Gentle. *Matrix Algebra: Theory, Computations, and Applications in Statistics.* Springer-Verlag, 2007.

[13] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian nonparametrics.* Springer-Verlag, New York, 2003.

[14] U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 56:549–603, 1994.

[15] Jaroslav Hajek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 175–194, 1972.

[16] I.A. Ibragimov and R.Z. Has'minskii. *Statistical estimation, asymptotic theory.* New York : Springer-Verlag, 1981.

[17] S. F. Jarner and G. O. Roberts. Convergence of heavy tailed MCMC algorithms.

[18] S. F. Jarner and G. O. Roberts. Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, 12:224–247, 2002.

[19] G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

[20] Kenneth Lange. A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):425–437, 1995.

[21] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses.* Springer, 3rd edition, 2005.

[22] C. Liu, D. Rubin, and Y. Wu. Parameter epansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85(4):755–770, 1998.

[23] Jun S. Liu, Wing Hung Wong, and Augustine Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

[24] David G. Luenberger. *Linear and nonlinear programming.* Kluwer Academic Publishers, 2 edition, 2003.

[25] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* Wiley, 1997.

[26] Xiao-Li Meng. On the rate of convergence of the ECM algorithm. *Ann. Statist.*, 22(1):326–339, 1994.

[27] Xiao-Li Meng and Donald B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80(2):267–278, 1993.

[28] Xiao-Li Meng and David van Dyk. The EM Algorithm –An Old Folk-song Sung to a Fast New Tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):511–567, 1997.

[29] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24:101–121, 1996.

[30] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer, 1993.

[31] Esa Nummelin. *General irreducible Markov chains and nonnegative operators.* Number 83 in Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1984.

[32] David Pollard. Asymptotics via empirical processes. *Statistical Science.*, 4(4):341–366, 1989.

[33] G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B*, 59(2):291–317, 1997.

[34] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. International Workshop in Applied Probability. *Methodology and Computing in Applied Probability*, 4:337–357.

[35] G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110.

[36] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, 2:341–363, 1996.

[37] O. Stramer and R. L. Tweedie. Langevin-type models. I. Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, 1:283–306, 1999.

[38] O. Stramer and R. L. Tweedie. Langevin-type models. II. Self-targeting candidates for MCMC algorithms. *Methodology and Computing in Applied Probability*, 1:307–328, 1999.

[39] L. Tierney. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–17028, 1994.

[40] P. Tuominen and R. L. Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Advances in Applied Probability*, 26:775–798, 1994.

[41] A.W. van der Vaart. *Asymptotic statistics.* Cambridge ; New York : Cambridge University Press, 1998.

[42] A. M. Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B. Methodological*, 31:80–88, 1969.

[43] C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.