To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–26

Convergence of Griddy Gibbs Sampling and other perturbed Markov chains

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

The Griddy Gibbs sampling was proposed by Ritter and Tanner (1992) as a computationally efficient approximation of the well-known Gibbs sampling method. The algorithm is simple and effective and has been used successfully to address problems in various fields of applied science. However, the approximate nature of the algorithm has prevented it from being widely used: the Markov chains generated by the Griddy Gibbs sampling method are not reversible in general, so the existence and uniqueness of its invariant measure is not guaranteed. Even when such an invariant measure uniquely exists, there was no estimate of the distance between it and the probability distribution of interest, hence no means to ensure the validity of the algorithm as a means to sample from the true distribution.

In this paper, we show, subject to some fairly natural conditions, that the Griddy Gibbs method has a unique, invariant measure. Moreover, we provide L^p estimates on the distance between this invariant measure and the corresponding measure obtained from Gibbs sampling. These results provide a theoretical foundation for the use of the Griddy Gibbs sampling method. We also address a more general result about the sensitivity of invariant measures under small perturbations on the transition probability. That is, if we replace the transition probability P of any Monte Carlo Markov chain by another transition probability Q where Q is close to P, we can still estimate the distance between the two invariant measures. The distinguishing feature between our approach and previous work on convergence of perturbed Markov chain is that by considering the invariant measures as fixed points of linear operators on function spaces, we don't need to impose any further conditions on the rate of convergence of the Markov chain. For example, the results we derived in this paper can address the case when the considered Monte Carlo Markov chains are not uniformly ergodic.

Keywords: Griddy Gibbs, nonreversible Markov chain, perturbed Markov kernel, non-uniformly ergodic Markov chain

AMS Subject Classification: 60J20; 65C40; 92B05; 32A70.

1. Introduction

The need to generate samples from a probability function or estimate moments of such a distribution arises in many fields of applied science, including Bayesian statistics, computational physics, computational biology and computer science. A common difficulty in generating such samples is that the distribution (hereafter denoted by π) may be high-dimensional and computationally intractable. To resolve this problem, many sampling-based approaches have been proposed: the basic idea is to construct a Markov chain with a tractable transition mechanism that has π as its invariant distribution.

One of the most widely applicable methods to construct such a Markov chain is the method of Gibbs sampling. This algorithm generates an instance from the distribution of each variable in turn, conditional on the current values of the other variables. This reduces the sampling problem to a series of one-dimensional problems. The method of Gibbs sampling is very computationally effective, especially in the case when π is high-dimensional. Gibbs sampling applies even in the case that the distribution is known only up to a normalizing constant, which occurs commonly in fitting models to data.

However, the use of the Gibbs sampling method is hindered by several factors. First, the method requires the one-dimensional conditional densities to be known, or at least to be easy to sample directly. In most contexts, such knowledge about the conditional densities is usually not available. Second, in many fields of applied sciences, sampling from the conditional distributions is computationally expensive, despite the fact that they are one-dimensional. For instance, in systems biology, evaluating (up to a normalizing factor) the value of the distribution function π at one point might be equivalent to solving a high-dimensional system of differential equations. These high computational costs pose a serious problem in applying the algorithm in practice, which motivates the development of sampling methods which uses approximation of the full distribution to alleviate this difficulty.

To address these issues, Ritter and Tanner (1992) proposed in [1] an approximate method – the Griddy Gibbs method – as an alternative. The Griddy Gibbs sampling method evaluates the conditional density on a grid and uses piecewise linear or piecewise constant functions to approximate the cumulative distribution function of the conditional distributions based on these grid values. The resulting distribution is used to generate random variables with approximately the right distribution.

This method has been used successfully to address problems in various fields of applied science: statistical modeling and inference ([2–7]), machine learning ([8]), chemical analysis ([9]), systems biology ([10–12]), medical science([13]), statistical computing and data analysis([14, 15]), economics([16–18]), ecological modelling([19]), acoustics ([20]), and time series analysis ([21–23]). However, the approximate nature of the algorithm still prevents it from being widely used. The approximation by linear or constant functions leads to theoretical questions about the ergodic properties of the constructed Markov chains and about the validity of the algorithm as a means to sample from the true distribution.

Many adjustments to overcome the approximate nature of the algorithm have been proposed. In [24], a Metropolis chain is embedded in the algorithm to ensure that the equilibrium distribution is exactly π even on a coarse grid. In [25], a similar strategy is proposed, in which the Multiple-try Metropolis algorithm is embedded in the sampling process. In both approaches, the convergence of the algorithms are guaranteed, but the computational costs increase considerably and the algorithms are more difficult to set up. Another popular grid-based sampling algorithm is the adaptive rejection sampler (ARS), which is capable of sampling the distribution efficiently with a low number of grid points

[38]. However, ARS-type algorithms are designed for log-concave (or near log-concave) distributions and in general does not work well for general distribution with multiple modes. In all cases, the approximations are restricted to piecewise linear and piecewise constant functions.

In this paper, we show, assuming that the approximations to the distribution are bounded from above and bounded away from zero, that the Griddy Gibbs method has a unique, invariant measure. Moreover, we provide L^p estimates on the distance between this invariant measure and the corresponding measure obtained from Gibbs sampling. Subject to appropriate hypotheses, our main results about Griddy Gibbs are the following.

(1) Although the Markov chains generated by the Griddy Gibbs sampler are not reversible in general, they admit unique invariant measures.
(2) For $2 \leq p \leq \infty$, there is an L^p-estimate of the distance between the limit invariant measure and the correct distribution π, which guarantees the L^p-convergence of the algorithm.

The first result is obtained using tools from the theory of Markov processes. We then extend the Markov chain transition operator to L^p-spaces and use techniques from the theory of functional analysis on Hilbert vector spaces to prove the second result. These results provide a theoretical foundation for the use of the Griddy Gibbs sampling method with some guarantees of convergence.

In this paper, we go beyond these results for Griddy Gibbs sampling in two ways. First, the approximation scheme does not need to be piecewise linear or piecewise constant: any reasonable approximation scheme can be employed to obtain Griddy Gibbs sampling. In fact, in the case that the distribution is smooth but the computational cost of determining the value of the conditional distribution is much greater than the cost for approximation, high order polynomial interpolations are preferred since they increase the accuracy of the sampling process and reduce the number of function evaluations.

Second, we generalize our method to give results about the sensitivity of invariant measures under small perturbations on the transition probability. That is, if we replace the transition probability P of any Monte Carlo Markov chain by another transition probability Q where Q is close to P, can we still estimate the distance between the two invariant measures? Our paper provides a positive answer to this question, given some mild conditions imposed on Q. The distinguishing feature between our approach and other work [26–28] on convergence of perturbed Markov chain is that by considering the invariant measures as fixed points of linear operators on function spaces, we don't need to impose any further conditions on the rate of convergence of the Markov chain. For example, the results we derive in this paper can address the case when the considered Monte Carlo Markov chains are not uniformly ergodic. The result about the sensitivity of invariant measures under small perturbations thus guarantees the convergence and provides a way to assess the efficiency of Griddy Gibbs sampler and other perturbed Markov chains methods.

The paper is organized as follows. Section 2 provides the mathematical framework used in the paper, as well as descriptions of the Gibbs and Griddy Gibbs sampling methods. Section 3 discusses the existence, uniqueness and regularity results for the invariant measure. We develop in Section 4 results about the sensitivity of invariant measures under small perturbations on the transition probability for general non-uniformly ergodic Monte Carlo Markov chains. The estimates are then extended to the case when the distribution of interest has non-compact support in Section 5. Finally, we provide in

Section 6 numerical examples to illustrate our theoretical findings and demonstrate the utility of the Griddy Gibbs sampling method.

2. Mathematical framework

The problem addressed by the Gibbs algorithm is the following (see [29] for reference). We are given a density function $\hat{\pi}$, on a state space with bounded Lebesgue measure $D \subset \mathbb{R}^d$. This density gives rise to an absolutely continuous probability measure π on D, by

$$\pi(A) = \int_A \hat{\pi}(x)dx, \quad \forall A \in \mathcal{B}$$

where $\mathcal{B}$ denotes the σ-algebra of Borel sets on D. Without loss of generality, we assume throughout this paper that the the distribution π has finite variance. In other words, the density function $\hat{\pi} \in L^2(D)$.

In many applications, we want to estimate the expectations of functions $\phi : D \to \mathbb{R}$ with respect to π, i.e. we want to estimate

$$\pi(\phi) = E_\pi[\phi(X)] = \int_D \phi(x)\hat{\pi}(x)dx.$$

If D is high-dimensional, and $\hat{\pi}$ is a complicated function, then direct integration (either analytic or numerical) of these integrals is infeasible.

The classical Monte Carlo solution to this problem is to simulate independent and identically distributed random variables $X_1, X_2, ..., X_N$ with distribution π and then estimate $\pi(\phi)$ by

$$\pi_N(\phi) = \frac{1}{N}\sum_{i=1}^{N} \phi(X_i). \tag{1}$$

This gives an unbiased estimate with standard deviation of order $O(1/\sqrt{N})$. However, if π_u is complicated, it is difficult to directly simulate i.i.d. random variables from π. The Markov chain Monte Carlo (MCMC) approach is introduced instead to construct on D a Markov chain that is computationally efficient and that has π as a stationary distribution. That is, we want to define easily-simulated Markov chain transition probabilities $P(x, y)$ for $x, y \in D$ such that

$$\int_D \hat{\pi}(y)P(x, y)dy = \hat{\pi}(x), \quad for\ a.e.\ x \in D.$$

In principle, if we run the Markov chain (started from anywhere) to obtain samples X_n, then for large n the distribution of X_n will be approximately stationary, and the sequence $\{X_n\}$ can be used to estimate $\pi(\phi)$ as in equation (1).

2.1. *Gibbs transition*

The Gibbs transition is a transition probability on D defined as follows. First, the i^{th} component Gibbs transition P_i leaves all components except the i^{th} component unchanged and replaces the i^{th} component by a draw from the full distribution π conditional on all other components:

$$P_i(x_1,\ldots,x_i,\ldots,x_d) = \frac{\hat{\pi}(x_1,\ldots,x_i,\ldots,x_d)}{\int_{-\infty}^{\infty}\hat{\pi}(x_1,\ldots,t,\ldots,x_d)dt},$$

where t appears in the i^{th} position. The transition probability of the Gibbs sampler is defined as

$$P(x,y) = P_1(y_1,x_2,\ldots,x_d)P_2(y_1,y_2,x_3,\ldots,x_d)\cdots P_d(y_1,y_2,\ldots,y_d)$$

where $x = (x_1,x_2,\ldots,x_d)$ and $y = (y_1,y_2,\ldots,y_d)$.

Now let $\{X_n\}, n \geq 0$ be a time-homogeneous Markov process generated by the Gibbs sampling algorithm with transition probability P. We have

$$P(X_n \in A|X_0 = a) = P^n(a,A), \quad \forall A \in \mathcal{B}$$

where P^n is defined recursively by

$$P^1 = P, \qquad P^n(a,y) = \int_D P(x,y)P^{n-1}(a,x)dx.$$

We also define the transition operator T on $\mathcal{P}$(D), the space of probability measures on D, by

$$T\mu(A) = \int_D P(x,A)\mu(dx). \tag{2}$$

This transition operator can also be considered as a linear operator on $L^p(D)$, $1 \leq p \leq \infty$, by defining

$$Tf(y) = \int_D P(x,y)f(x)dx$$

Moreover, the operator T^n obtained by replacing P by P^n in (2) is equal to the operator obtained by applying T n times, $T^n = T \circ T \circ \ldots \circ T$.

2.2. *Ergodic properties of the Markov chains generated by the Gibbs sampling*

By standard results about ergodicity of Gibbs sampling method, we know that under rather general conditions, T admits a unique invariant measure, which is the distribution π that we want to sample, i.e. $T\pi = \pi$. Moreover, the distribution of X_n converges in total variation norm to π. We state here Theorem 6 from [30] that justifies the convergence of the Gibbs sampling method:

Theorem 2.1 ([30]) *Assume that for each $1 \leq i \leq d$, the conditional distributions $\pi(X_i|X_j, j \neq i)$ have densities, say p_i, with respect to some dominating measure ρ_i. Suppose further that for each $1 \leq i \leq d$, there is a set A_i with $\rho_i(A_i) > 0$, and a $\delta > 0$ such that for each $1 \leq i \leq d$*

(1) $\pi(X_i = x_i | X_j = x_j, j \neq i) > 0$ *whenever* $x_k \in A_k$ *for all* $k \leq i$ *and* $x_{i+1}, ..., x_d$ *arbitrary.*

(2) $\pi(X_i = x_i | X_j = x_j, j \neq i) > \delta$ *whenever* $x_k \in A_k$ *for all* $k \leq d$.

Then for π*-a.e.* $x \in D$*, we have* $\sup_{C \in \mathcal{B}} |P^n(x, C) - \pi(C)| \to 0$.

In the rest of the paper, we will assume that the distribution of interest π satisfies conditions of Theorem 2.1 and has finite variance.

2.3. *Griddy Gibbs transition and Griddy Gibbs sampler*

In the Griddy Gibbs sampling method, at each point in the sampling space and on each dimension, we use some approximation scheme to approximate P_i. The i^{th} component Griddy Gibbs transition leaves all components except the i^{th} component unchanged and replaces the i^{th} component by a draw from Q_i that approximates the conditional expectation on all other components, i.e.,

$$Q_i(x_1, ..., y_i, ..., x_d) \approx \frac{\hat{\pi}(x_1, ..., y_i, ..., x_d)}{\int_{-\infty}^{\infty} \hat{\pi}(x_1, ..., t, ..., x_d) dt}. \tag{3}$$

The Griddy Gibbs sampler is defined in a similar manner as in the definition of Gibbs sampler: starting with an initial value, we sequentially and randomly update the i^{th} component while fixing other components of the variables. However, for each i^{th} component, instead of the conditional distribution P_i (which is unknown or difficult to compute), we use the approximation Q_i as the guideline for the sampling process. The surrogate function Q_i is obtained by evaluating the conditional density P_i on a grid and uses some interpolation to approximate P_i based on these grid values. A simple one-dimensional simple, such as the inverse transform sampler, is then used to sample the i^{th} component corresponding to Q_i.

As with Gibbs sampler, the transition probability and transition operator of the Griddy Gibbs are defined as

$$Q(x, y) = Q_1(y_1, x_2, ..., x_d) Q_2(y_1, y_2, ..., x_d) ... Q_d(y_1, y_2, ..., y_d) \tag{4}$$

and

$$Q^n(a, y) = \int_D Q(x, y) Q^{n-1}(a, x) dx \qquad \text{and} \qquad S\mu(A) = \int_D Q(x, A) \mu(dx).$$

We note that since the approximations on each dimension are different, the Markov chain $\{Y_n\}$ generated by Griddy Gibbs algorithm is not reversible in general.

Throughout this paper, we will use the notation $\{X_n\}$, T, P to describe a Markov chain generated by Gibbs sampling, its transition operator and its transition probability, respectively. The corresponding notations for Griddy Gibbs are $\{Y_n\}$, S, Q. A comparison between the notations used for the Gibbs sampling and Griddy Gibbs sampling is provided in Figure 1.

$$\text{Gibbs:} \qquad \{X_n\} \xrightarrow{T,P} \pi$$

$$\text{Griddy Gibbs:} \qquad \{Y_n\} \xrightarrow{S,Q} \eta$$

Figure 1. Comparison between Gibbs sampling and Griddy Gibbs sampling: Although the two transition operators P and Q are close, the Markov chain $\{Y_n\}$ is not reversible in general, so the existence and uniqueness of the invariant measure η is not guaranteed. Even when η uniquely exists, an estimate of the distance between π and η is needed to guarantee the validity of the Griddy Gibbs sampling.

3. Existence, uniqueness, and regularity of the invariant measure of a Monte Carlo Markov chain generated by the Griddy Gibbs sampling

In this section, we will prove that the transition operator S (obtained from the Griddy Gibbs algorithm as in (3), (4)) admits a unique invariant measure η, assuming that the approximations Q_i are uniformly bounded above and away from zero:

$$\exists M, \epsilon > 0, \text{ such that } \epsilon \leq Q_i(x) \leq M, \ \forall 1 \leq i \leq d, \ \forall x \in D. \tag{5}$$

We also prove that under this condition, η is absolutely continuous with respect to Lebesgue measure and admits a bounded density function. We note that condition (5) is general and does not hinder the application of the Griddy Gibbs sampling method, since we can always use additional cutoff functions on the approximation scheme to guarantee the boundedness from above and below of f_i, without significantly affecting the accuracy of the approximation scheme.

The outline of the proof is as follow. By verifying Doeblin's condition (see Theorem 3.1), we prove the existence and uniqueness of the invariant measure η; moreover, the distribution of $\{Y_n\}$ (obtained by the Griddy Gibbs algorithm) converges to η in total variation norm. Using this and Lemma 3.1, we deduce that η is absolutely continuous with respect to Lebesgue measure. Finally, using Lemma 3.2, we prove that the density function of η is bounded.

3.1. *Existence and uniqueness*

To verify the existence and uniqueness of the invariant measure, we use the following result from [31] on the convergence of transition probabilities. As before, we will denote by $\mathcal{B}$ the σ-algebra of Borel sets on D.

Theorem 3.1 ([31]) *Suppose that the Markov chain Z_n with transition probability $K(x, \cdot)$ satisfies the Doeblin condition:*

$\exists k \in N, \epsilon > 0$, and a probability measure ϕ on $(D, \mathcal{B})$ such that

$$K^k(x, C) \geq \epsilon\phi(C), \forall x \in D, \forall C \in \mathcal{B}.$$

Then there exists a unique invariant probability measure ξ such that for all $n \in N$ and all $x \in D$,

$$\sup_{C \in \mathcal{B}} |K^n(x, C) - \xi(C)| \leq (1 - \epsilon)^{((n/k)-1)}.$$

Using this result, we can prove that under condition (5), the distribution of the Markov chain $\{Y_n\}$ generated by the Griddy Gibbs sampling method converges to a stationary distribution η in total variation norm. This is a direct analog of the convergence given in Theorem 2.1 above (although we still have to show that η is near π).

THEOREM 3.2 *(Existence and uniqueness of the invariant measure for S) Assume that the approximation scheme* $\{f_i\}_{i=1}^d$ *satisfies condition* (5). *Then there exists a unique probability measure* η *that is invariant under* S, *and this* η *satisfies*

$$\sup_{C \in B} |Q^n(x, C) - \eta(C)| \to 0$$

for all $x \in D$. *In other words,* $\forall x \in D$, $Q^n(x, \cdot) \to \eta(\cdot)$ *in total variation norm.*

Proof. We will prove that the transition probability Q constructed in the Griddy Gibbs sampling algorithm satisfies Doeblin's condition of Theorem 3.1. Recall that the transition probability in the Griddy Gibbs algorithm is given by

$$Q(x, C) = \int_C f_1(y_1, x_2, \dots, x_d) f_2(y_1, y_2, \dots, x_d) \cdots f_d(y_1, y_2, \dots, y_d)\ dy_1 dy_2 \dots dy_d.$$

Recall that $f_i \geq \epsilon$ on D from (5). Hence with Vol(C) denoting the Lebesgue measure of C, we have

$$Q(x, C) \geq \epsilon^d \text{Vol}(C), \forall x \in D, \forall C \in B.$$

This is Doeblin's condition with $k = 1$, ϕ is the Lebesgue measure on D, so applying Theorem 3.1, we have $\sup_{C \in \mathcal{B}} |Q^n(x, C) - \eta(C)| \to 0$. ■

3.2. *Some supporting lemmas*

To establish results about regularity of the invariant measure η of Markov chains generated by Griddy Gibbs sampling, we need the following two lemmas. The first result is about the absolute continuity of η, while the second result provides a basic inequality for the transition operator as a linear operator on L^p space. As we will see in the next section, the two lemmas allow us to prove that the invariant measure η has bounded density with respect to the Lebesgue measure.

The proofs are standard, but we sketch them for completeness.

LEMMA 3.1 *Let* μ_n *be a sequence of probability measures on* $(D, \mathcal{B})$ *that converges in total variation norm to a measure* μ. *Assume further that each* μ_n *is absolutely continuous with respect to Lebesgue measure. Then* μ *is also absolutely continuous w.r.t Lebesgue measure and admits a non-negative density function.*

Proof. Consider any Borel measurable set A with $|A| = 0$. By the assumption of absolute continuity, $\mu_n(A) = 0$, hence $\mu(A) = \lim \mu_n(A) = 0$. Since A was arbitrary, μ is absolutely continuous w.r.t. Lebesgue measure. ■

Throughout the rest of this section, for any two linear normed space U, V, we will denote by $\mathcal{L}(U, V)$ the space of al linear operator with the source domain and the target domain being U and V, respectively, equipped with the standard operator norm. We then have the following result.

Lemma 3.2 *For $1 \leq p \leq \infty$, let $K(x,y)$ be a bounded function on $D \times D$, and let*

$$\mathcal{V}g(y) = \int K(x,y)g(x)dx$$

for $g \in L^p(D)$. Then

(a) $\mathcal{V}$: $L^2(D) \to L^2(D)$ is a compact linear operator. Moreover

$$\|\mathcal{V}\|_{\mathcal{L}(L^2(D),L^2(D))} = \|K\|_{L^2(D\times D)}.$$

(b) $\mathcal{V}$: $L^1(D) \to L^1(D)$ is a bounded linear operator. Moreover, if $K(x,y)$ is a transition probability function, then

$$\|\mathcal{V}\|_{\mathcal{L}(L^1(D),L^1(D))} \leq 1.$$

(c) $\mathcal{V}$ maps $L^1(D)$ to $L^\infty(D)$, and

$$\|\mathcal{V}\|_{\mathcal{L}(L^1(D),L^\infty(D))} \leq \|K\|_\infty.$$

(d) If $g \in L^2(D)$ and $2 \leq p \leq \infty$ then

$$\|\mathcal{V}g\|_p \leq \|K\|_p \max\{\|g\|_1, \|g\|_2\}.$$

Before proceeding to provide the proof, we note that Lemma 3.2 plays a central role in the rest of the paper. To be more precise, parts (b) and (c) of the Lemma helps establish that the invariant measure η has bounded density, while parts (a) and (d) helps provide the L^p estimates of the sensitivity of the invariant measures of Markov chains under kernel perturbation (Section 4).

Proof. Part a) of the Lemma is a well-known result about Hilbert-Schmidt integral operators. For reference, cf. [33].

For b) and c), let $M = \sup_{D\times D} |K(x,y)| = \|K\|_\infty$. Then for all $y \in D$ we have

$$|\mathcal{V}g(y)| = \left|\int K(x,y)g(x)dx\right| \leq M \int |g(x)|dx.$$

In other words, $\|\mathcal{V}g\|_\infty \leq \|K\|_\infty \|g\|_1$.

Integrating over D gives

$$\|\mathcal{V}g\|_1 \leq \mathrm{Vol}(D)\|\mathcal{V}g\|_\infty \leq \mathrm{Vol}(D)\|K\|_\infty\|g\|_1,$$

which proves b) and c).

Finally, if K is a transition probability, we have

$$\int |\mathcal{V}g(y)|dy = \int \left|\int K(x,y)g(x)dx\right| dy \leq \int \int K(x,y)dy|g(x)|dx = \int |g(x)|dx$$

which implies $\|\mathcal{V}\|_{\mathcal{L}(L^1(D),L^1(D))} \leq 1$.

For d), consider the linear operator W defined on $L^2(D \times D)$ and on $L^\infty(D \times D)$ as

$$W\phi = \int_D \phi(x,y)g(x)dx$$

Then from part a) and c), we have W is a bounded linear operator that maps $L^2(D \times D)$ to $L^2(D)$, and maps $L^\infty(D \times D)$ to $L^\infty(D)$. Moreover, the following inequalities are satisfied:

$$\|W\phi\|_{L^2(D)} \leq \|g\|_{L^2(D)}\|\phi\|_{L^2(D\times D)}$$

$$\|W(\phi)\|_{L^\infty(D)} \leq \|g\|_{L^1(D)}\|\phi\|_{L^\infty(D\times D)}$$

Using Riesz-Thorin interpolation theorem (see [32]), we deduce that W also maps $L^p(D\times D)$ to $L^p(D)$, and

$$\|W\phi\|_{L^p(D)} \leq \max\{\|g\|_{L^2(D)}, \|g\|_{L^1(D)}\}\|\phi\|_{L^p(D\times D)}.$$

Replace ϕ by K, noticing that $\mathcal{V}g = W(K)$, we deduce

$$\|Lg\|_p \leq \|K\|_p \max\{\|g\|_1, \|g\|_2\}.$$

■

3.3. *Regularity*

These two previous lemmas allow us to prove the following result.

THEOREM 3.3 *(Regularity of invariant measure) The invariant measure η of S is absolutely continuous w.r.t Lebesgue measure on D. Moreover, there exists $\hat{\eta} \in L^\infty(D)$ so that for each $C \in \mathcal{B}$,*

$$\eta(C) = \int_C \hat{\eta}(x)dx.$$

Also, $\hat{\eta}$ is invariant under S: $S\hat{\eta} = \hat{\eta}$.

Proof. The proof of this theorem is straightforward from the previous theorems and lemma. From Theorem 3.2 and Lemma 3.1, we know that η is absolutely continuous and admits a density function:

$$\eta(dx) = \hat{\eta}(x)dx$$

with $\hat{\eta} \in L^1(D)$.

Now considering S as a bounded linear operator on $L^1(D)$, we have

$$\begin{aligned}\int_A \hat{\eta}(x)dx = \eta(A) = S\eta(A) = \int_D Q(x,A)\eta(dx)\\ = \int_D \int_A Q(x,y)dy\ \hat{\eta}(x)dx = \int_A \left(\int_D Q(x,y)\hat{\eta}(x)dx\right)\ dy.\end{aligned}$$

Since A was arbitrary, we deduce that

$$\hat{\eta}(x) = \int_D Q(x,y)\hat{\eta}(x)dx$$

or $\hat{\eta} = S\hat{\eta}$. From Lemma 3.2, S maps $L^1(D)$ to $L^\infty(D)$. Hence $\hat{\eta} = S\hat{\eta} \in L^\infty(D)$, so $\hat{\eta}$ is a bounded function.

■

Remark 3.1 *Since D is a subset with bounded measure of R^d, $\hat{\eta}$ also belongs to $L^p(D)$, for all $1 \leq p \leq \infty$.*

4. Sensitivity and convergence of non-uniformly ergodic Markov chains

Before proceeding to give result about the sensitivity of the invariant measures under perturbation, we want to make a remark that the assumption of uniformly boundedness away from 0 of the approximations Q_i was introduced only to guarantee the existence and uniqueness of an absolutely continuous invariant measure η.

As we mentioned before, we can always use additional cutoff functions on the approximation scheme to guarantee the boundedness from below of Q_i, without significantly affecting the accuracy of the approximation scheme. However, as an analysis of convergence of perturbed Monte Carlo Markov chains, condition(5) is replaced by any condition that guarantees the existence and uniqueness of the invariant measure η and the ergodicity of the Markov chain $\{Y_n\}$. In a similar manner, the assumptions of Theorem 2.1 can be replaced by the existence and uniqueness of the invariant measure π and the ergodicity of the Markov chain $\{X_n\}$.

In short, we will assume the following conditions in the subsequent analyses

(1) the invariant measures π, η of the Markov chain exists and are unique.
(2) the Markov chains $\{X_n\}$, $\{Y_n\}$ are ergodic (not necessarily uniformly ergodic).
(3) the distributions π, η have finite second moments.

The distinguishing feature between our approach and other work [26–28] on convergence of perturbed Markov chain is that by considering the invariant measures as fixed points of linear operators on function spaces, we don't need to impose any further conditions on the rate of convergence of the Markov chain. For that reason, the results we derived in this paper can address the case when the considered Monte Carlo Markov chains are not uniformly ergodic.

4.1. *Continuity of eigenspaces for eigenvalue 1*

We recall from the previous part of the paper that the two transition operators T and S admit unique absolutely continuous invariant measures π and η, respectively. Before proceeding to derive estimates of the distance between $\hat{\pi}$ and $\hat{\eta}$, we provide in this section two key lemmas to further investigate properties of the transition operators T and S as operators on $L^2(D)$.

In Lemma 4.1, we will prove that the eigenspaces correspond to eigenvalue $\lambda = 1$ of T and S are one-dimensional subspaces spanned by π and η, respectively. Lemma 4.2 investigates a special case when it is possible to estimate the distance between the positive invariant eigenvectors of two close operators.

Lemma 4.1 *Using the same notation as in Section 2.3 and consider T, S as operators on $H = L^2(D)$, we have*

- $\{v \in H : Tv = v\} = \langle \hat{\pi} \rangle$
- $\{v \in H : Sv = v\} = \langle \hat{\eta} \rangle$,

where $\langle \hat{\pi} \rangle$ denotes the span of $\hat{\pi}$.

Proof. Consider any $w \in H - \{0\}$ such that $Tw = w$. Then

$$\int |Tw(y)|dy = \int \left| \int P(x,y)w(x)dx \right| dy \leq \int \int P(x,y)dy|w(x)|dx = \int |w(x)|dx.$$

Equality happens only when

$$\left| \int P(x,y)w(x)dx \right| = \int |P(x,y)w(x)|dx$$

for a.e. $y \in D$.

Since $P(x,y) > 0$, this happens only if w does not change sign on D. Therefore, if we define

$$w^* = \frac{w}{\|w\|_{L^1(D)}}$$

then w^* is the density function of a probability measure on D. Moreover, we also have $Tw^* = w^*$. Since π is the unique invariant measure that is also a fixed point of T, we deduce that $w^* = \hat{\pi}$. Hence, $w \in \langle \hat{\pi} \rangle$. ■

Lemma 4.2 *Let M and N be Hilbert-Schmidt integral operators on $H = L^2(D)$. Assume further that $u, v \in H$ such that*

(i) $\|u\|_H = \|v\|_H = 1$
(ii) $\{w \in H : Mw = w\} = \langle u \rangle$
(iii) $\{w \in H : Nw = w\} = \langle v \rangle$
(iv) *u, v are positive functions.*

Then there exists $\alpha > 0$ that depends only on M such that

$$\|v - u\|_H \leq C(\alpha)\|M - N\|_{L(H,H)}.$$

Proof. Since H is a Hilbert space, we can write

$$H = \langle u \rangle \oplus K$$

where K is the orthogonal complement of the linear space spanned by u. For the sake of convenience, in the rest of the proof, we will denote $\|\cdot\|_H$ simply by $\|\cdot\|$.

First we show that there exists $\alpha > 0$ such that

$$\|(M-I)k\| \geq \alpha\|k\| \ \forall k \in K.$$

By way of contradiction, suppose that $\exists \alpha_n \to 0, \|k_n\| = 1, k_n \in K$ such that $\|Mk_n - k_n\| = \alpha_n$. Since M is a compact operator on H, by extracting a subsequence, we can assume that $Mk_n \to k_\infty \in H$. On the other hand, we have $\|Mk_n - k_n\| = \alpha_n \to 0$. By the triangle inequality, we have

$$\|k_n - k_\infty\| \leq \|k_n - Mk_n\| + \|Mk_n - k_\infty\| \to 0.$$

We deduce that $k_n \to k_\infty$, and hence that $\|k_\infty\| = 1$. Since K is closed we have $k_\infty \in K$, and since M is continuous we have $Mk_\infty = k_\infty$. By (ii), $Mu = u$ has no nontrivial solution in K, so we deduce that $k_\infty = 0$, which contradicts $\|k_\infty\| = 1$.

On the other hand, we can uniquely decompose

$$v = \lambda u + k \tag{6}$$

for some $\lambda \in \mathbb{R}, k \in K$. Since u and v are fixed points of M and N, respectively, we deduce that

$$Mv = M(\lambda u + k) = \lambda Mu + Mk = \lambda u + Mk$$

and

$$Nv = v = \lambda u + k.$$

Therefore

$$\|M - N\|_{L(H,H)} \geq \|Mv - Nv\| = \|(\lambda u + Mk) - (\lambda u + k)\| = \|Mk - k\| \geq \alpha\|k\|.$$

The orthogonal decomposition in (6) gives

$$1 = \|v\|^2 = \lambda^2\|u\|^2 + \|k\|^2 = \lambda^2 + \|k\|^2,$$

so

$$\lambda^2 = 1 - \|k\|^2 \geq 1 - \left(\frac{\|M-N\|_{L(H,H)}}{\alpha}\right)^2.$$

This plus the same decomposition also gives

$$\begin{aligned}\|v-u\|^2 &= (\lambda-1)^2\|u\|^2 + \|k\|^2 = \lambda^2 - 2\lambda + 1 + \|k\|^2 \\ &= 2(1-\lambda) = 2\frac{1-\lambda^2}{1+\lambda}.\end{aligned}$$

On the other hand, from the facts that u, v are positive functions (by (iv)) with $\|u\| = 1$ (by (i)) and the orthogonal decomposition of v, we have $\lambda = \langle u, v\rangle = \int_D uv\, dx \geq 0$.

Hence

$$\|v-u\|^2 \le 2(1-\lambda^2) = 2\|k\|^2 \le 2\frac{\|M-N\|^2_{L(H,H)}}{\alpha^2}$$

or

$$\|v-u\| \le \sqrt{2}\frac{\|M-N\|_{L(H,H)}}{\alpha}.$$

■

4.2. *Convergence results*

In this section, we answer the question about the sensitivity of the invariant measure of a Monte Carlo Markov chain under kernel perturbations: given that $\|P-Q\| < \epsilon$ (or equivalently, given a small perturbation on the transition operator), can we estimate the distance $\|\pi-\eta\|$ between the two invariant measures?

The outline of this section is as follows. Using Lemma 4.1 and 4.2, we derive the L^2-estimate of the distance between $\hat{\eta}$ and $\hat{\pi}$.

Then, knowing that S maps $L^1(D)$ to $L^\infty(D)$, we bound the L^∞-norm by L^2-norm to produce an L^∞-estimate, and then apply Lemma 3.2 to derive the L^p estimate for $2 \le p \le \infty$.

Since the proofs require us to switch back and forth between norms, let us recall that if $f \in L^\infty(D)$ then

$$\|f\|_1 \le C\|f\|_2 \text{ and } \|f\|_2 \le C\|f\|_\infty$$

where $C = \sqrt{\text{Vol}(D)}$.

Theorem 4.1 *(L^2-estimate)*
There exists $\delta(\pi), C(\pi) > 0$ such that for $\|P-Q\|_2 < \delta(\pi)$, we have

$$\|\hat{\pi}-\hat{\eta}\|_2 \le C(\pi)\,\|P-Q\|_2$$

Proof. For clarity, we replace $\hat{\pi}$ and $\hat{\eta}$ with π and η respectively. Applying Lemma 4.2 with $u = \frac{\pi}{\|\pi\|_2}$, $v = \frac{\eta}{\|\eta\|_2}$, we have

$$\left\|\frac{\pi}{\|\pi\|_2} - \frac{\eta}{\|\eta\|_2}\right\|_2 \le \sqrt{2}\frac{\|T-S\|}{\alpha}. \tag{7}$$

Then

$$\left\|\frac{\pi}{\|\pi\|_2} - \frac{\eta}{\|\eta\|_2}\right\|_1 \le C\left\|\frac{\pi}{\|\pi\|} - \frac{\eta}{\|\eta\|}\right\|_2 \le C\sqrt{2}\frac{\|T-S\|}{\alpha}$$

with $C = \sqrt{\mathrm{Vol}(D)}$. By the triangle inequality

$$\left| \left\| \frac{\pi}{\|\pi\|_2} \right\|_1 - \left\| \frac{\eta}{\|\eta\|_2} \right\|_1 \right| \leq C\sqrt{2}\frac{\|T-S\|}{\alpha}.$$

Since π and η are probability measures, we have $\|\pi\|_1 = \|\eta\|_1 = 1$, and hence

$$\left| \frac{1}{\|\pi\|_2} - \frac{1}{\|\eta\|_2} \right| \leq C\sqrt{2}\frac{\|T-S\|}{\alpha}.$$

This leads to

$$1 - \frac{\|\pi\|_2}{\|\eta\|_2} \leq C\sqrt{2}\frac{\|T-S\|}{\alpha}\|\pi\|_2. \tag{8}$$

If we assume further that the right hand side is less than 1, then

$$\|\eta\|_2 < \frac{\|\pi\|_2}{1 - C\sqrt{2}\frac{\|T-S\|}{\alpha}\|\pi\|_2}. \tag{9}$$

The triangle inequality plus (7) and (8) give

$$\begin{aligned}\|\pi-\eta\|_2 &\leq \left\| \pi - \frac{\|\pi\|_2\eta}{\|\eta\|_2} \right\|_2 + \left\| \eta - \frac{\|\pi\|_2\eta}{\|\eta\|_2} \right\|_2 \\ &\leq \sqrt{2}\frac{\|T-S\|}{\alpha}\|\pi\|_2 + C\sqrt{2}\frac{\|T-S\|}{\alpha}\|\pi\|_2\|\eta\|_2\end{aligned}$$

and then (9) gives

$$\|\pi-\eta\|_2 \leq \sqrt{2}\frac{\|T-S\|}{\alpha}\|\pi\|_2 \left(1 + C\frac{\|\pi\|_2}{1 - C\sqrt{2}\frac{\|T-S\|}{\alpha}\|\pi\|_2}\right). \tag{10}$$

Since T is defined by π, we can consider α as a function of π only. Moreover, with $\delta(\pi) = \alpha/(2C\sqrt{2}\,\|\pi\|_2)$ and $\|T-S\| \leq \delta(\pi)$, the right hand side of (8) is at most 1/2, and the constant in parentheses in (10) is at most $1 + 2C\|\pi\|_2$. Hence we define

$$C(\pi) = \frac{\sqrt{2}\,\|\pi\|_2\,(1 + 2C\,\|\pi\|_2)}{\alpha}$$

and note that from Lemma 3.2,

$$\|T-S\|_{\mathcal{L}(L^2,L^2)} = \|P-Q\|_{L^2(D\times D)}.$$

Hence for $\|P-Q\|_2 < \delta(\pi)$, changing back to the original notations, we have the desired estimate

$$\|\hat{\pi} - \hat{\eta}\|_2 \leq C(\pi)\,\|P-Q\|_2\,.$$

■

Remark 4.1 *From* (9) *and the choice of* $\delta(\pi)$*, we see that if* $\|P - Q\|_2 < \delta(\pi)$*, then* $\|\hat{\eta}\|_2 < 2\|\hat{\pi}\|_2$.

Theorem 4.2 *(*L^∞*-estimate)*
There exists $\delta'(\pi), C'(\pi) > 0$ *such that if* $P, Q \in L^\infty(D \times D)$ *and* $\|P - Q\|_\infty < \delta'(\pi)$ *then*

$$\|\hat{\pi} - \hat{\eta}\|_\infty \le C'(\pi)\, \|P - Q\|_\infty .$$

Proof. As in the proof of the previous theorem, we replace $\hat{\pi}$ and $\hat{\eta}$ with π and η respectively. Using part the fact that π and η are fixed by T and S, respectively, plus the triangle inequality and c) of Lemma 3.2, we have

$$\begin{aligned} \|\eta - \pi\|_\infty = \|S\eta - T\pi\|_\infty &\le \|S\eta - T\eta\|_\infty + \|T\eta - T\pi\|_\infty \\ &\le \|P - Q\|_\infty \|\eta\|_1 + \|P\|_\infty \|\eta - \pi\|_1 \\ &\le C\|P - Q\|_\infty \|\eta\|_2 + C\|P\|_\infty \|\eta - \pi\|_2, \end{aligned} \tag{11}$$

with $C = \sqrt{\mathrm{Vol}(D)}$. With $\delta(\pi)$ and $C(\pi)$ as in the previous theorem, define

$$\delta'(\pi) = \frac{\delta(\pi)}{C} \quad \text{and} \quad C'(\pi) = 2C\|\pi\|_2 + C^2\|P\|_\infty C(\pi).$$

If $\|P - Q\|_\infty < \delta'(\pi)$, then as mentioned previously,

$$\|P - Q\|_2 \le C\|P - Q\|_\infty < \delta(\pi).$$

We start with 11 and then use $\|\eta\|_2 < 2\|\pi\|_2$ and $\|\pi - \eta\|_2 \le C(\pi)\, \|P - Q\|_2$ from remark 4.1 and theorem 4.1 to get

$$\begin{aligned} \|\eta - \pi\|_\infty &\le C\|P - Q\|_\infty \|\eta\|_2 + C\|P\|_\infty \|\eta - \pi\|_2 \\ &\le 2C\|P - Q\|_\infty \|\pi\|_2 + C\|P\|_\infty C(\pi)\|P - Q\|_2 \end{aligned}$$

By collecting terms and noticing that $\|P - Q\|_2 \le C\|P - Q\|_\infty$, we deduce that

$$\begin{aligned} \|\eta - \pi\|_\infty &\le \left(2C\|\pi\|_2 + C^2\|P\|_\infty C(\pi)\right) \|P - Q\|_\infty \\ &= C'(\pi)\|P - Q\|_\infty \end{aligned}$$

■

Theorem 4.3 *(*L^p*-estimate,* $2 \le p \le \infty$*)*
Let $2 \le p \le \infty$*, there exists* $\delta'(\pi), C'(\pi) > 0$ *such that if* $P, Q \in L^p(D \times D)$ *and* $\|P - Q\|_p < \delta'(\pi)$ *then*

$$\|\hat{\pi} - \hat{\eta}\|_p \le C'(\pi)\, \|P - Q\|_p .$$

Proof. As before, we replace $\hat{\pi}$ and $\hat{\eta}$ with π and η respectively. Applying Lemma 3.2 (part d), noticing that η and π belong to $L^2(D)$, we have:

$$\|S\eta - T\eta\|_p \leq \|P - Q\|_p \max\{\|\eta\|_1, \|\eta\|_2\}$$

and

$$\|T\eta - T\pi\|_p \leq \|P\|_p \max\{\|\eta - \pi\|_1, \|\eta - \pi\|_2\}.$$

Using the fact that π and η are fixed by T and S, respectively, plus the triangle inequality and c) of Lemma 3.2, we have

$$\begin{aligned} \|\eta - \pi\|_p = \|S\eta - T\pi\|_p &\leq \|S\eta - T\eta\|_p + \|T\eta - T\pi\|_p \\ &\leq \|P - Q\|_p \max\{\|\eta\|_1, \|\eta\|_2\} + \|P\|_p \max\{\|\eta - \pi\|_1, \|\eta - \pi\|_2\} \\ &\leq C\|P - Q\|_p\|\eta\|_2 + C\|P\|_p\|\eta - \pi\|_2, \qquad (12) \end{aligned}$$

with $C = \sqrt{\text{Vol}(D)}$. The rest of the proof concludes as in the proof of the previous theorem. ■

5. Extension to non-compact support distributions

While most of the assumption of the method on the ergodicity of the Markov chains are quite general, one restriction of the method comes from the assumption of bounded parameter space D. Since the key ideas of our analysis of sensitivity of the invariant measures rely on moving back and forth between the L^p-norms, this condition could not be easily removed from the framework.

However, it is worth noting that for distributions with non-compact support, a variation of the Griddy Gibbs sampling method can be developed as followed: first, a rectangular domain D is chosen by prior knowledge about π, then the Griddy Gibbs sampling with $\pi_{new} = \pi|_D$ (normalized by a constant) is proceeded as usual. By our previous analyses, the Monte Carlo Markov chains generated by this process will have a unique invariant measure η whose distance to π can be estimated by the following theorem

Theorem 5.1 *Let $2 \leq p \leq \infty$. Assume that π has non-compact support on $\mathcal{R}^d$ and that there exist $C_1, C_2 > 0$ such that:*

$$\int_{\mathcal{R}^d} \|x\|_1 \hat{\pi}(x)^p \, dx \ \leq C_1 \qquad \textit{and} \qquad \int_{\mathcal{R}^d} \|x\|_1 \hat{\pi}(x) \, dx \ \leq C_2$$

where $\|x\|_1 = |x_1| + ... + |x_d|$. Let $D_t = \{x \in \mathcal{R}^d : \|x\|_\infty > t\}$ where $\|x\|_\infty = \max_i |x_i|$.

There exists $\delta'(\pi), C'(\pi) > 0$ such that if $P, Q \in L^p(D \times D)$, $\|P - Q\|_p < \delta'(\pi)$ and $t \geq C_2/2$ then

$$\|\hat{\pi} - \hat{\eta}\|_p \leq C'(\pi, D_t) \|P - Q\|_p + \frac{C_2}{2t}\|\hat{\pi}\|_p + \frac{C_1}{t\|\hat{\pi}\|_p} \qquad (13)$$

Proof. We denote $f = \hat{\pi}^p / \|\hat{\pi}\|_p$ and let X be a random variable with density function f. By Markov's inequality, we have for $i = 1, 2, ...d$

$$\begin{aligned}
\mathbb{P}[|X_i| > t] &\leq \frac{1}{t}\int_{\mathcal{R}^d} |x_i| f(x)\ dx \\
&= \frac{1}{t\|\hat{\pi}\|_p}\int_{\mathcal{R}^d} |x_i| \hat{\pi}^p(x)\ dx
\end{aligned}$$

Hence

$$\|\hat{\pi}\|_{L^p(\mathcal{R}^d \setminus D_t)} = \mathbb{P}[\|X\|_\infty > t] \leq \frac{\int_{\mathcal{R}^d} \|x\|_1 \hat{\pi}^p(x)\ dx}{t\|\hat{\pi}\|_p}$$

By a similar argument, we have

$$\|\hat{\pi}\|_{L^1(\mathcal{R}^d \setminus D_t)} \leq \frac{1}{t}\int_{\mathcal{R}^d} \|x\|_1 \hat{\pi}(x)\ dx \leq \frac{C_2}{t}$$

On the other hand, results for distribution with compact support in D implies

$$\left\| \hat{\eta} - \frac{\hat{\pi}}{\int_{D_t} \hat{\pi}} \right\|_{L^p(D)} \leq C'(\pi, D_t)\|P - Q\|_p,$$

We deduce that, for $t \geq 2C_2$, we have

$$\begin{aligned}
\|\hat{\pi} - \hat{\eta}\|_{L^p(\mathcal{R}^d)} &\leq \|\hat{\pi} - \hat{\eta}\|_{L^p(D)} + \|\hat{\pi}\|_{L^p(\mathcal{R}^d \setminus D_t)} \\
&\leq \left\| \hat{\eta} - \frac{\hat{\pi}}{\int_{D_t} \hat{\pi}(x)\ dx} \right\|_p + \frac{\int_{\mathcal{R}^d \setminus D_t} \hat{\pi}(x)\ dx}{\int_{D_t} \hat{\pi}} \|\hat{\pi}\|_p + \|\hat{\pi}\|_{L^p(\mathcal{R}^d \setminus D_t)} \\
&\leq C'(\pi, D_t) \left\| P - Q \right\|_p + \frac{C_2}{2t}\|\hat{\pi}\|_p + \|\hat{\pi}\|_{L^p(\mathcal{R}^d \setminus D_t)} \\
&\leq C'(\pi, D_t) \left\| P - Q \right\|_p + \frac{C_2}{2t}\|\hat{\pi}\|_p + \frac{C_1}{t\|\hat{\pi}\|_p}
\end{aligned}$$

■

Corollary 5.1 *Let $2 \leq p \leq \infty$ and assume that π has non-compact support on $\mathcal{R}^d$ and that there exists $C_3, C_4 > 0$ so that:*

(1) $\int_{\mathcal{R}^d} |x|^2 \hat{\pi}(x)\ dx \leq C_3 < \infty$ and
(2) $\|\hat{\pi}\|_{L^{2p-1}} \leq C_4 < \infty$

Then result (13) is true with $C_1 = \sqrt{C_3 C_4^p}$ and $C_2 = 1 + C_4$.

That is, if prior estimates on the second moment of π and the L^{2p-1}-norm of $\hat{\pi}$ are available, the Griddy Gibbs algorithm can be adjusted accordingly to produce a good estimate on the distance between the two invariant measures.

Proof. By Holder's inequality with

$$u(x) = \|x\|_1 \hat{\pi}(x)^{1/2}, \quad v(x) = \hat{\pi}^{p-1/2}$$

we have

$$\begin{aligned}\int_{\mathcal{R}^d} \|x\|_1 \hat{\pi}^p(x)\, dx &\leq \left(\int_{\mathcal{R}^d} \|x\|_1^2 \hat{\pi}(x)\, dx\right)^{1/2} \left(\int_{\mathcal{R}^d} \hat{\pi}^{2p-1}(x)\, dx\right)^{1/2} \\ &\leq \sqrt{C_1 C_2^p}.\end{aligned}$$

and

$$\int_{\mathcal{R}^d} \|x\|_1 \hat{\pi}(x)\, dx \leq \int_{\|x\|_1 \leq 1} \hat{\pi}(x)\, dx + \int_{\|x\|_1 > 1} \|x\|_1^2 \hat{\pi}(x)\, dx \leq 1 + C_4$$

■

6. Numerical examples

In this section, we provide numerical examples to illustrate our theoretical findings and demonstrate the utility of the Griddy Gibbs sampling method. First, we validate the estimates derived in previous sections in a simple 2D example. We then proceed to investigate the performance of the Griddy Gibbs sampling in a practical example arising from systems biology, in which it is necessary to employ the Griddy Gibbs sampling method, and demonstrate the use of the method in making inferences about the system.

6.1. *A 2D example*

In this example, we investigate the performance of the Griddy Gibbs sampling algorithm on grids of various resolutions in a simple 2D example. The chosen distribution for has the following density function

$$\pi(x,y) = \frac{1}{2} Beta\left(\frac{x+1}{2}, 2, 5\right) * Beta\left(\frac{y+1}{2}, 2, 5\right)$$

$$+ \frac{1}{2} Beta\left(\frac{x+1}{2}, 2, 2\right) * Beta\left(\frac{y+1}{2}, 2, 2\right)$$

where $Beta(x, \alpha, \beta)$ is the density of the one-dimensional Beta distribution with parameter α and β.

This distribution was chosen specifically to illustrate the developed framework in the case of compact support: it has compact support in its domain $[-1,1] \times [-1,1]$ and has non-independent components, but the 1D marginal density functions can be obtained in simple form:

$$\pi_X(x) = \frac{1}{2} Beta\left(\frac{x+1}{2}, 2, 5\right) + \frac{1}{2} Beta\left(\frac{x+1}{2}, 2, 2\right).$$

Using this probability distribution, we illustrate the estimates provided in previous sections, by expressing the L^2 and L^∞ distance between the estimator (using Griddy

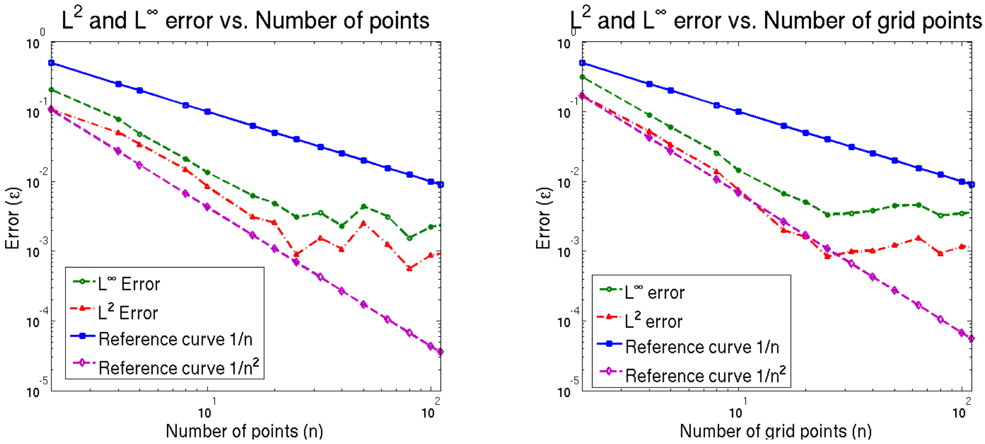

Figure 2. Left: Error of the 1D marginal empirical cumulative distribution function, and Right: error of the empirical cumulative distribution function, both as a function of the number of points used in the approximation grid.

Gibbs) and the true distribution of interest in terms of the number of points used in the grid of approximation. For various number number of points n used in the grid of approximation, we use standard linear interpolation on n equally spaced points in the interval $[-1, 1]$ to approximate the 1-dimensional conditional distributions. For each grid, a Griddy Gibbs chain of length 10^5 is generated. We then use the sampled points to estimate the empirical cumulative distribution function (ECDF) and the 1D marginal ECDF of the invariant distribution of the chains. Finally, the L^2 and L^∞ distance between the estimated ECDFs with different number of grid points and the true CDF are calculated.

We note that in the context of our example, it is more convenient to work with CDFs rather than with PDFs for two main reasons: (i) CDFs can be approximated using nonparametric estimators; and (ii) there is a well-developed theoretical machinery for the comparison of CDFs using such estimators. Moreover, it is well-known that the ECDF is a non-parametric, unbiased estimator that converges uniformly to the true CDF (a result known as the Glivenko - Cantelli theorem [34]).

The results are illustrated in Figure 2. The error of both ECDF and the marginal ECDF of the first variable decrease faster than $O(\frac{1}{n})$ and approximately as fast as $O(\frac{1}{n^2})$ when the number of the grids point n increases, until it reaches a level at which the error of the Griddy Gibbs sampling is dominated by the error of the Monte Carlo simulation. Since the accuracy of standard 1D linear approximation method is bounded by $O(\frac{1}{n})$, and can be as fast as $O(\frac{1}{n^2})$ if the function has bounded second derivative, this confirms our theoretical results about linear dependency between error of the 1D approximation, and the distance from the estimated distribution to the true distribution of interest.

6.2. *An example in systems biology.*

In this example, we consider a mathematical model of the T-cell signaling pathway proposed by Lipniacki et al. in [35]. The behaviour of the system is modelled as an ODE system controlled by 19 different parameters with 37 state variables and fixed initial conditions:

$$\begin{aligned} \dot{x} &= \alpha(\omega, x) && \text{(System of ODEs)} \\ x(0) &= x_0(\omega) && \text{(Initial conditions)} \\ y(t) = f(\omega, t) &= \beta(\omega, x(t)) && \text{(Output)} \end{aligned}$$

Here $x = (x_1, x_2, ..., x_{n_x}) \in M \subset \mathbb{R}^{n_x}$ is the state variable, with M a subset of $\mathbb{R}^{n_x}$ containing the initial state, and $f(\omega, t) \in \mathbb{R}$ is the output response (system dynamics). In the scope of this paper, we are interested in the dynamics of pZap, one of the state variables of the system. The vector of unknown parameters is denoted by $\omega = (\omega_1, ..., \omega_N) \in \mathbb{R}^N$ and is assumed to belong to a subset Ω of $\mathbb{R}^N$. These functions and initial conditions depend on the parameter vector $\omega \in \Omega$.

The traditional approach to study such a system is to estimate values of the parameters from observations. However, in the field of systems biology, usually it is not possible to estimate all parameters in a given model, in particular if the model is complex and the data is sparse and noisy. Thus, to represent explicitly the state of knowledge, it is best to consider not a single parameter valuation but the whole space of uncertain parameters. The uncertainty in parameter values is often characterized by a probability distribution $\pi(\omega)$ on the set of all possible parameter values, based on how the output of the system driven by a particular parameter valuation fits previous data. This gives a distribution with density

$$\hat{\pi}(\omega) = c_n \ \exp\left(- \sum_{i=1}^{n} |f(\omega, t_i) - d(t_i)|^2 \right), \tag{14}$$

where c_n is a normalizing constant, $(t_1, d_1), ...(t_n, d_n)$ is the set of previous data.

Inference about the system will be made based on π. For example, in [10, 36], the optimal experiment is chosen at the time point where the maximum value of the normalized variance of the outputs with respect to π is achieved. Another example was given in [11] where the expected dynamics estimator to recover the correct system dynamics is defined as the expected value of the system dynamics with respect to the distribution π.

This motivates the problem of sampling with respect to the distribution π. As noted in the introduction, the use of the standard Gibbs sampling method is hindered by two factors: first, there is no closed-form formula for the distribution π or for the corresponding one-dimensional conditional distributions; second, the evaluation of the unnormalized distribution at one point is computationally expensive (it is equivalent to solving a high dimensional system of differential equations). It is then necessary to approximate the conditional distribution by functions of simpler forms. The Griddy Gibbs method therefore is a suitable choice for this sampling process.

In this particular example, we restrict the analysis to the five most sensitive parameters with respect to perturbation. This choice is based on previous knowledge about the dynamics of the system and on the result of a global sensitivity analysis using sparse grid interpolation([37]).

To further reduce the computational cost, we also employ a sparse grid interpolant to approximate the output of the ODE system. That is, the output functions of the system of ODEs are evaluated on a sparse grids of 10^5 points on the parameter space, then the method of sparse grid interpolation is employed to approximate the outputs at other sets of parameter values. Moreover, the one-dimensional conditional distributions are then approximated by piecewise linear functions on grids of fineness $\delta = 0.2$ (which corresponds to a grid with 11 equally spaced points). It is worth noting that although this is a two-leveled approximation, it still fits into the framework developed in previous sections.

We will compare the performance of the Griddy Gibbs sampling with the variation of Gibbs sampling suggested by Tierney et al. in [24]. In Tierney's algorithm, a Metropolis chain is embedded to ensure that the equilibrium distribution is exactly π even on a coarse grid. The drawback is that the computational cost is at least twice as much as Griddy Gibbs sampling using the same grid. Moreover, the algorithm is more difficult to set up and is restricted to piecewise linear and piecewise constant approximations.

In Figure 3, we use samples from Griddy Gibbs and from Tierney's algorithm to compare the conditional and marginal distribution derived from the ECDFs. In the left panel, we compare the conditional distributions (using samples from Griddy Gibbs and Tierney's algorithm) of the second parameter on the first paramter, for various values of this parameter. In the right panel, the difference between the two marginal joint distributions of the first and the second variable are computed. We also compare the difference between the two marginal joint distributions of the first and the second variable while using various numbers of samples in Figure 4. The results from Figure 3 and Figure 4 suggest that the Griddy Gibbs sampling method is as effective as Tierney's algorithm (whose convergence is also guaranteed theoretically) in generating Markov chains with with respect to a given invariant measure: the difference between the two marginal distributions is of the same magnitude as the error of the Monte Carlo method itself.

We then investigate the performance of the Griddy Gibbs sampling in making inferences about dynamics. For this we consider the Expected Dynamics Estimator based on one single simulated data point. This generates a distribution π_1 as in (14) with $n = 1$, and we then use this distribution to estimate the system dynamics by

$$\hat{D}_1(t) = E_{\pi_1(\omega)}[f(\omega, t)].$$

The results are provided in Figure 5 (Left). The expected dynamics are calculated using the empirical mean of the output values on the previous two sets of samples. Once again, the performance of the Griddy Gibbs sampling is as good as Tierney's algorithm in computing the expected dynamics.

Finally, Figure 5 (Right) compares the auto-correlation coefficients of the Monte Carlo Markov chains generated by the two algorithms. To compute the auto-correlation coefficients, two Monte-Carlo Markov chains of length 10^5 were generated by the two algorithms, respectively. The figure illustrates the fact that not only is the computational cost of Tierney's algorithm (to generate one instance of the chain) higher, but also its auto-correlation function converges (to zero) at a much lower rate. In this particular example, if one wants to get two sets of i.i.d samples with the same number of points by both algorithms, the computational cost for Tierney's algorithm is at least ten times that of the cost for Griddy Gibbs.

7. Conclusion

We have shown, subject to some fairly natural conditions, that the Griddy Gibbs method has a unique, invariant measure. Moreover, we gave L^p estimates on the distance between this invariant measure and the corresponding measure obtained from Gibbs sampling. These results provide a theoretical foundation for the use of the Griddy Gibbs sampling method.

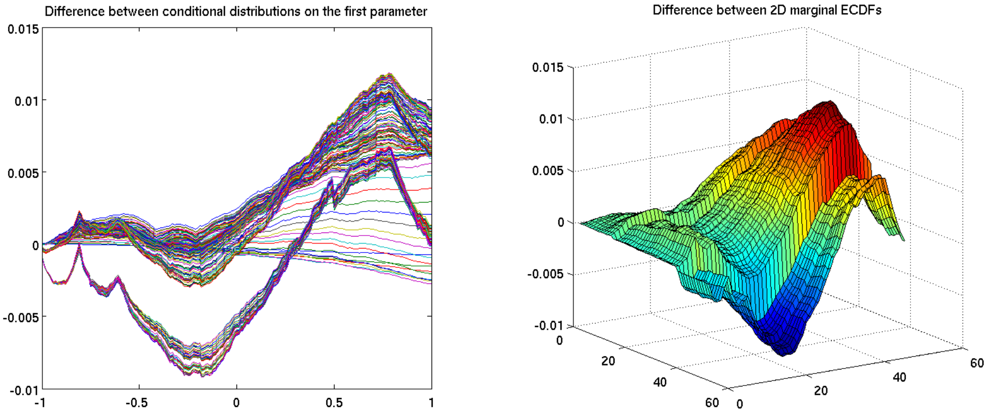

Figure 3. Conditional and marginal distribution for the T-cell model. Left: The difference between the conditional distributions on the first parameters (one curve for each value of this parameter). Right: The difference between the marginal joint distributions of the first two parameters, achieved from Griddy Gibbs and Tierney's algorithm. Figure 4 shows that the differences between corresponding ECDFs are of the same magnitude as the error of the Monte Carlo method ($O(\frac{1}{\sqrt{N}})$, where N is the number of samples)

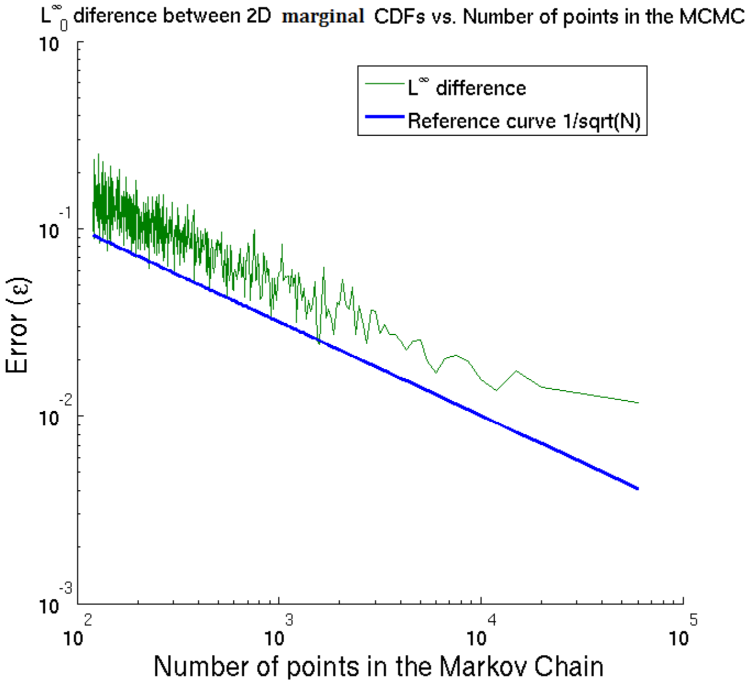

Figure 4. The difference between the marginal distributions computed by Griddy Gibbs and Tierney's algorithm is of the same magnitude as the error of the Monte Carlo method itself ($O(\frac{1}{\sqrt{N}})$, where N is the number of samples).

Moreover, using the theoretical framework developed to validate the Griddy Gibbs sampling method, we also successfully provided a more general result about the sensitivity of invariant measures under small perturbations on the transition probability. Our results imply that if we replace the transitional probability P of a Monte Carlo Markov chain by a different transitional probability Q that is close to P in L^p norm ($2 \leq p \leq \infty$), the distance between the two invariant measures (in L^p) is bounded by a constant times

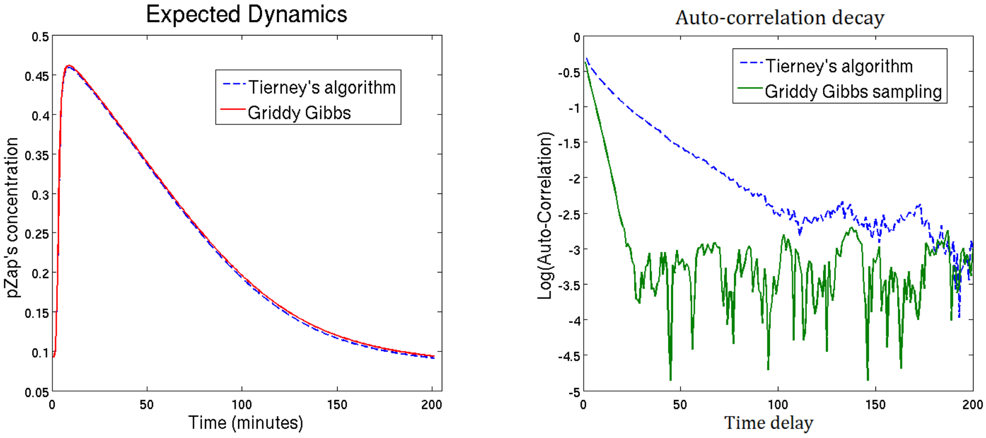

Figure 5. Left: The expected dynamics estimator based on one data point, generated by Griddy Gibbs and Tierney's samples. Right: Auto-Correlation coefficients of the Markov chains generated by Griddy Gibbs algorithm and Tierney's algorithm.

the L^p-distance between P and Q, provided that the approximation schemes satisfy a mild condition provided in the paper. This condition is very general and does not hinder the application of the Griddy Gibbs sampling method, since it can always be guaranteed simply by using additional cutoff functions on the approximation scheme, without significantly affecting its accuracy. The method can be generalized to validate other Monte Carlo Markov chain sampling methods that involve approximation.

We also gave numerical examples to illustrate our theoretical findings and demonstrate the utility of the method in different applications. The numerical results confirm the linear relation between the distance between the invariant measures and the accuracy of the approximation scheme derived in theory. Moreover, our examples illustrate that Griddy Gibbs performs as well as its variants in applications and that the algorithm is simpler to implement and less computationally expensive. Additionally, the Markov chains generated by this algorithm have significantly smaller auto-correlation coefficients than those of other variant algorithms. These features demonstrate that Griddy Gibbs is a simple and effective sampling method that can be employed in applications with confidence in its validity.

Acknowledgement

All authors of this paper are partially supported by NSF grant DMS-0900277.

References

[1] Ritter C and Tanner MA. Facilitating the Gibbs Sampler: the Gibbs Stopper and the Griddy-Gibbs Sampler. *J. Am. Stat. Assoc.* 1992; **87**: 861–868.
[2] Ray B, McCulloch R and Tsa R. Bayesian methods for change-point detection in long-range dependent processes. *Stat. Sinica.* 1997; **7**: 451–472.
[3] Barnard J, McCulloch R and Meng X. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sinica.* 2000; **10**: 1281–1311.
[4] Yang H et al.. Adaptive sampling for Bayesian geospatial models. *Statist. Comput.* 2014; 24.6: 1101–1110.

[5] Kundu S and David BD. Bayes variable selection in semiparametric linear models. *J. Am. Stat. Assoc.* 2014; 109.505: 437–447.

[6] Kundu S and David BD. Latent factor models for density estimation. Biometrika. 2014; 101.3: 641–654.

[7] Pham TH, Ormerod JT and Wand MP. Mean field variational Bayesian inference for nonparametric regression with measurement error. *Comput. Stat. Data. An.* 2013; 68: 375-387.

[8] Mansinghka Vikash et al.. Approximate bayesian image interpretation using generative probabilistic graphics programs. *Advances in Neural Information Processing Systems.* 2013.

[9] Ritter C. Statistical analysis of spectra from electron spectroscopy for chemical analysis. *J. Roy. Stat. Soc. D-Sta.* 1994; **43**: 111–127.

[10] Dong W et. al. Systems Biology of the Clock in Neurospora crassa. *PLoS ONE.* 2008; 3: e3105.

[11] Dinh V, Rundell AE and Buzzard GT. Experimental Design for Dynamics Identification of Cellular Processes. *Bull. Math. Biol.* 2014; 76.3: 597–626.

[12] Dinh V, Rundell AE and Buzzard GT. Effective sampling schemes for Behavior Discrimination in nonlinear systems. *International Journal for Uncertainty Quantification.* 2014; 4.6.

[13] J. Li et. al. A random-effects Markov transition model for Poisson-distributed repeated measures with non-ignorable missing values. *Stat. Med.* 2007; **26**: 2519–2532.

[14] Ardia D, Hoogerheide L and Dijk H. Adaptive mixture of Student-t distributions as a flexible candidate distribution for efficient simulation: The R Package AdMit. *J. Stat. Softw.* 2009; **29**: 1–32.

[15] Ausna M and Galeano P. Bayesian estimation of the Gaussian mixture GARCH model *Comput. Stat. Data. An.* 2007; **51**: 2636–2652.

[16] Boatwright P, McCulloch R, and Rossi P. Account-level modeling for trade promotion: an application of a constrained parameter hierarchical model. *J. Am. Stat. Assoc.* 1999; **94** : 1063–1073.

[17] Bauwens L and Rombouts J. Bayesian inference for the mixed conditional heteroskedasticity model. *Economet. J.* 2007;**10**: 408–425.

[18] Walkera D, Prez-Barberab F and Marion G. Stochastic modelling of ecological processes using hybrid Gibbs samplers. *Ecol. Model.* 2006;**198**: 40–52.

[19] Bauwens L and Rombouts J. Bayesian clustering of many GARCH models. *Economet. Rev.* 2007; **26**: 365–386.

[20] Michalopoulou Z and Picarelli M. Gibbs sampling for time-delay and amplitude estimation in underwater acoustics. *J. Acoust. Soc. Am.* 2005; **117**: 799–8087.

[21] Boatwright P, McCulloch R and Rossi P. A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. *J. Roy. Stat. Soc. C-App.* 2001; **50**: 187–200.

[22] Ray B and Tsa R. Bayesian methods for change-point detection in long-range dependent processes. *J. Time. Ser. Anal.* 2002; **23**: 687–705.

[23] Chen C and Wen Y. On goodness of fit for time series regression models *J. Stat. Comput. Sim.* 2001; **69**: 239–256.

[24] Tierney L. (1994). Markov chains for exploring posterior distributions. *Ann. Stats.* 1994; **22**: 1701–1728.

[25] Liu, S.J. et. al. The Multiple-Try Method and local optimization in Metropolis sampling. *J. Am. Stat. Assoc.* **95**, 121–134.

[26] Roberts GO, Rosenthal JS and Schwartz PO Convergence properties of perturbed Markov chains. *JJ. Appl. Probab.*. 1998: 1–11

[27] Mitrophanov AY. Sensitivity and convergence of uniformly ergodic Markov chains. *J. Appl. Probab.*, 2005: 1003–1014.

[28] Andrieu C and Roberts GO (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stats.* 2009; **37.2**, 697–725.

[29] Roberts GO and Rosenthal JS. General state space Markov chains and MCMC algorithms. *Probab. Surv.* 2004; **1**: 20–71.

[30] Athreya, K. et. al. On the convergence of the Markov Chain Simulation Method. *Ann. Stats.*1996; **24**: 69–100.

[31] Koralov L and Sinai Y. Theory of probability and random processes. Springer. 2007.

[32] Lasser R. Introduction to Fourier Analysis. CRC Press. 1996

[33] Pedersen M. Functional analysis in applied mathematics and engineering. CRC Press. 1999.

[34] Lillacci G and Khammash G. A distribution-matching method for parameter estimation and model selection in computational biology. *Int. J. Robust. Nonlinear Control* 2012; **22**: 1065–1081.

[35] Lipniacki T, Hat B, Faeder JR, Hlavacek WS. Stochastic effects and bistability in T cell receptor signaling. *J. Theoret. Biol.*. 2008; **254**: 110–122.

[36] Donahue MM, Buzzard GT and Rundell AE (2010). Experiment design through dynamical characterisation of non-linear systems biology models utilising sparse grids. *IET System Biology.* 2010; **4**: 249–262.

[37] Buzzard GT. Global sensitivity analysis using sparse grid interpolation and polynomial chaos. *Reliability Engineering and System Safety.* 2012; **17**: 82–89.

[38] Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics.* 1992; **1**: 337–48.