STOCHASTIC APPROXIMATION AND ITS APPLICATION IN MCMC

A Dissertation

by

YICHEN CHENG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Faming Liang
Committee Members,	Huiyan Sang
	Samiran Sinha
	Jianxin Zhou
Department Head,	Simon Sheather

August 2013

Major Subject: Statistics

ABSTRACT

Stochastic approximation has been widely used since first proposed by Herbert Robbins and Sutton Monro in 1951. It is an iterative stochastic method that attempts to find the zeros of functions that cannot be computed directly. In this thesis, we used the technique in several different aspects. It was used in the analysis of large geostatistical data, in the improvement of simulated annealing algorithm also, as well as for NMR protein structure determination.

1. We proposed a resampling based Stochastic approximation method for the analysis of large geostatistical data. The main difficulty that lies in the analysis of geostatistical data is the computation time is extremely long when the sample size becomes large. Our proposed method only use a small portion of the data at each iteration. Each time, we update our estimators based on a randomly selected subset of the data using stochastic approximation. In this way, we use the information from the whole data set while keep the computation time almost irrelevant to the sample size. We proved the consistency of our estimator and showed by simulation study that the computation time is much reduced compared to other existing methods.

2. Simulated Annealing algorithm has been widely used for optimization problems. However, it can not guarantee the global optima to be located unless a logarithmic cooling schedule is used. However, the logarithm rate is so slow that no one can afford such a long cpu time. We proposed a new stochastic optimization algorithm, the so-called simulated stochastic approximation annealing (SAA) algorithm, which is a combination of simulated annealing and the stochastic approximation Monte Carlo (SAMC) algorithm. It is shown that the new algorithm can work with a cooling schedule that decreases much faster than in the logarithmic cooling schedule

while guarantee the global optima to be reached when temperature tends to zero.

3. Protein Structure determination is a very important topic in computational biology. It aims to determine different conformations for each protein, which helps to understand biological functions such as protein-protein interactions, protein-DNA interactions and so on. Protein structure determination consists of a series of steps and peak picking is a very important step. It is the prerequisite for all other steps. Manually pick the peaks is very time consuming. To automate this process, several methods have been proposed. However, due to the complexity of NMR spectra, the existing method is hard to distinguish false peaks and true peaks perfectly. The main difficulty lies in identifying true peaks with low intensity and overlapping peaks.

We propose to model the spectrum as a mixture of bivariate Gaussian densities and used stochastic approximation Monte Carlo (SAMC) method as the computational approach to solve this problem. Essentially, by putting the peak picking problem into a Bayesian framework, we turned it into a model selection problem. Because Bayesian method will automatically penalize including too much component into the model, our model will distinguish true peaks from noises without pre-process of the data.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my committee chair, Professor Faming Liang, who has the attitude and the substance of a genius. He has a deep and continuous passion in research and is very efficient in terms of his brilliant ideas. He gave a lot of great ideas that guided my search direction. It is really an honor that I could work with him and share the experience of how a good research project can be done.

I would like to thank my committee members, Professor Sang, Sinha and Zhou. They have devoted a lot of time and effort to make this dissertation better. Their advise and comments really helped a lot in the whole process.

Last but not least, I would like to thank all my friends and families that support me with their heart and soul. They make the past 5 years a joyful and memorable moment for me to cherish for the rest of my life.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION: STOCHASTIC APPROXIMATION

1.1 Stochastic Approximation

Stochastic approximation methods are a series of methods that tries to find the minima or zeros of a integration function. The problem can be written as finding the global minima of the following expectation

$$\min_{\theta \in \Theta} v(\theta) \triangleq E\{V(\theta, x)\}. \tag{1.1}$$

Or, equivalently, finding the zeros of the integration equation

$$h(\theta) = \int_{\mathcal{X}} H(\theta, x) g_\theta(x) dx = 0, \tag{1.2}$$

if we define $h(\theta) = v'(\theta)$ and $H(\theta, x) = \frac{\partial V(\theta, x)}{\partial \theta}$. Here, θ is the parameter vector, x is the random variable, and $g_\theta(x)$ is the density function for x that depends on parameter θ. The important thing here is that we don't get to observe the function $v(\theta)$ or $h(\theta)$ directly, instead, we observed their noisy version $V(\theta)$ or $H(\theta)$ respectively. In the literature of stochastic approximation, $h(\theta)$ is known as the mean field function and the difference between $h(\theta)$ and its noisy version $H(\theta, x)$ is known as observational noise, which is defined as follows:

$$\xi_{t+1} = H_{\tau_{t+1}}(\theta_t, x_{t+1}) - h_{\tau_{t+1}}(\theta_t),$$

In 1951, Robin and Monro introduced the so-called Robbins - Monro algorithm (1951) to solve the integration equation and the algorithm works as follows:

Algorithm 1.1.1. *Stochastic Approximation*

a. Generate $X_{t+1} \sim g_{\theta_t}(x)$, where t indexes the iteration.

b. Set $\theta_{t+1} = \theta_t + a_t H(\theta_t, X_{t+1})$, where a_t is the gain factor.

This Robbins â Monro algorithm is the most popular stochastic approximation method used. There is also another stochastic approximation method called Kiefer-Wolfowitz algorithm proposed by Kiefer and Wolfowitz in 1952, which is applied to the problem of finding minima. In this thesis, we will focus the stochastic approximation method proposed by Robbins and Monro.

In the case where it is not easy to directly sample from the density function $g_{\theta_t}(x)$, people suggest to substitute step a. by the following:

a'. Generate X_{t+1} from a Markov transition kernel P_{θ_t} that admit $g_{\theta_t}(x)$ as the invariant distribution.

One basic criteria for the above described algorithms to converge is that the gain factor satisfies the following condition

$$\sum_{t}^{\infty} a_t = \infty, \text{ and } \sum_{t}^{\infty} a_t^2 < \infty. \tag{1.3}$$

The first part of 1.3 says that the algorithm will not be influenced by the start point and the second part of 1.3 says that the method is bounded in variance which makes the algorithm converge.

1.2 Varying Truncation Stochastic Approximation MCMC

To ensure the convergence of Stochastic Approximation algorithm, some strong conditions need to be put on the mean field function, which are usually not very easy to verify. To loosen the conditions, several authors have proposed truncated version of stochastic approximation. A varying truncation stochastic approximation algorithm (Andrieu et al, 2005) works as follows:

Algorithm 1.2.1. *Varying Truncation Stochastic Approximation MCMC*

(i) Generate X_{t+1} that admit $g_{\theta_t}(x)$ as the invariant distribution, where t indexes the iteration.

(ii) Set $\theta_{t+\frac{1}{2}} = \theta_t + a_t H(\theta_t, X_{t+1})$, where a_t is the gain factor.

(iii) If $\|\theta_{t+\frac{1}{2}} - \theta_t\| \leq b_t$ and $\theta_{t+\frac{1}{2}} \in \mathcal{K}_{\pi_k}$, then set $\theta_{t+1} = \theta_{t+\frac{1}{2}}$ and $\pi_{t+1} = \pi_t$; otherwise, set $\theta_{t+1} = \mathbb{T}(\theta_t)$ and $\pi_{t+1} = \pi_t + 1$. Here $\mathbb{T}$ and π_t are defined as in algorithm 2.2.1.

Conditions for Convergence of Algorithm 1.2.1

Theoretical properties of algorithm 1.2.1 are studied under the following conditions:

(A_1) The function $h : \Theta \mapsto \mathbb{R}^d$ is continuous, and there exists a continuously differentiable function $v : \Theta \mapsto [0, \infty)$ such that:

(i) There exists $C_0 > 0$ such that

$$\mathcal{L} = \{\theta \in \Theta, \langle \nabla v(\theta), h(\theta) \rangle = 0\} \subset \{\theta \in \Theta, v(\theta) < C_0\}, \tag{1.4}$$

where $\langle x, y \rangle$ denotes the Euclidean inner product.

(ii) There exists $C_1 \in (C_0, \infty]$ such that $\mathcal{V}_{C_1}$ is a compact set, where $\mathcal{V}_C = \{\theta \in \Theta, v(\theta) \leq C\}$.

(iii) For any $\theta \in \Theta \setminus \mathcal{L}$, $\langle \nabla v(\theta), h(\theta) \rangle < 0$.

(iv) The closure of $v(\mathcal{L})$ has an empty interior.

(A_2) There exists a function $V : \mathcal{X} \to [1, \infty)$ such that for any compact subset $\mathcal{K} \subset \Theta$, there exists a constant c such that

(i) $\sup_{\theta\in\mathcal{K}} \|H(\theta,\cdot)\|_V \le c$;

(ii) $\sup_{(\theta,\theta')\in\mathcal{K}\times\mathcal{K}} \|H(\theta,\cdot) - H(\theta',\cdot)\|_V \le c\|\theta-\theta'\|$.

(A_3) The mean field function $h(\theta)$ is measurable and locally bounded. There exist a stable matrix F (i.e., all eigenvalues of F are with negative real parts), $\rho > 0$, and a constant c such that, for any $\theta_* \in \mathcal{L}$ (defined in (1.4)),

$$\|h(\theta) - F(\theta-\theta_*)\| \le c\|\theta-\theta_*\|^2, \quad \forall\theta \in \{\theta : \|\theta-\theta_*\| \le \rho\}.$$

(A_4) The sequences $\{a_t\}$ and $\{b_t\}$, which are defined to be $a(t)$ and $b(t)$ as functions of t and are exchangeable with $a(t)$ and $b(t)$, respectively, are non-increasing, positive, and satisfy the conditions:

$$\begin{aligned} &\lim_{t\to\infty} a_t = 0, \quad \sum_{t=0}^{\infty} a_t = \infty, \quad \frac{a_{t+1}-a_t}{a_t} = O(a_{t+1}^{\tau_1}), \\ &\lim_{t\to\infty} b_t = 0, \quad \sum_{t=1}^{\infty} \{a_t^{\tau_2} + (a_t/b_t)^{\tau_3} + a_t b_t^{\tau_4}\} < \infty, \end{aligned} \tag{1.5}$$

for some values of $\tau_1 \in (1,2]$, $\tau_2 \in (1,2]$, $\tau_3 \in [2,\infty)$ and $\tau_4 \in (0,1]$.

Moreover, we assume that the function $a(t)$ is differentiable, with either (i) or (ii) holding:

(i) $a(t)$ varies regularly with exponent $(-\beta)$, $\frac{1}{2} < \beta < 1$; that is, for any $z > 0$, $a(zt)/a(t) \to z^{-\beta}$ as $t \to \infty$.

(ii) For $t \ge 1$, $a(t) = t_0/t$ with $t_0 > -1/(2\lambda_F)$, where λ_F denotes the largest real part of the eigenvalue of the matrix F (defined in condition A_3) with $\lambda_F < 0$.

Condition (A_4) can be applied to the usual gains $a_t = t_0/t^{\beta}$ and $b_t = t_0'/t^{\beta'}$ by

choosing $\beta \in (\frac{1}{2}, 1]$, $\beta' \in (\frac{1}{2}, \beta - \frac{1}{\tau_3})$, $\tau_3 \in (2, \infty)$ and $\tau_4 = 1$. Following Pelletier (1998), we deduce that

$$\left(\frac{a_t}{a_{t+1}}\right)^{1/2} = 1 + \frac{\beta}{2t} + o(\frac{1}{t}). \tag{1.6}$$

In terms of a_t, (1.6) can be rewritten as

$$\left(\frac{a_t}{a_{t+1}}\right)^{1/2} = 1 + \zeta a_t + o(a_t), \tag{1.7}$$

where $\zeta = 0$ for the case (i) of (A_4) and $\zeta = \frac{\beta}{2t_0}$ for the case (ii) of (A_4). Clearly, the matrix is $F + \zeta I$ is still stable.

1.3 Stochastic Approximation Monte Carlo

Given a positive integrable function $f(x) : x \in \mathcal{X}, f \in L_1$, the corresponding energy function is defined as $U(x) = -log(f(x))$. We partition the sample space $\mathcal{X}$ into m disjoint subregions according to the energy function, and they can be written as: $E_1 = \{x : U(x) < u_1\}$, $E_2 = \{x : u_1 < U(x) < u_2\}$,$\cdots$, $E_{m-1} = \{x : u_{m-2} < U(x) < u_{m-1}\}$ and $E_m = \{x : U(x) > u_{m-1}\}$. Here $u_1, u_2, \cdots, u_{m-1}$ are pre-specified values. If we can find two values u_{min} and u_{max} such that $u_{min} < U(x) < u_{max}$ for all $x \in \mathcal{X}$. Then usually we set $u_1, u_2, \cdots, u_{m-1}$ to be equally spaced between u_{min} and u_{max}, so we have $U_i = \frac{i}{m}u_{max} + \frac{m-i}{m}u_{min}$ for $i = 1, \cdots, m-1$. SAMC algorithm (Liang et al., 2007) aims to sample from the following distribution:

$$p_\theta(x) \propto \sum_{i=1}^{m} \frac{f(x)}{e^{\theta_i}} I(x \in E_i), \tag{1.8}$$

where $\theta = (\theta_1, \cdots, \theta_m)$ and $\theta_i = log \int_{E_i} \phi(x)dx$.

Because $\int_{E_i} \phi(x)dx$ usually does not have an explicit form so we need to estimate it. We let θ_{ti} be the estimate of $log \int_{E_i} \phi(x)dx$ at iteration t. Then at time t, the

distribution can be estimated as:

$$p_{\theta_{\mathbf{t}}}(x) \propto \sum_{i=1}^{m} \frac{f(x)}{e^{\theta_{ti}}} I(x \in E_i), \tag{1.9}$$

where $\theta_{\mathbf{t}} = (\theta_{t1}, \cdots, \theta_{tm})$. Then a general SAMC works as follows:

1) At iteration t, simulate a sample $x^{(t+1)}$ from the proposal distribution $q(x^{(t)}, \cdot)$ that admits equation (1.9) as the invariant distribution

2) Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \gamma_{t+1}(\mathbf{e}_{t+1} - \pi)$. Where $\mathbf{e_{t+1}} = (e_{t+1,1}, \cdots, e_{t+1,m})$, $e_{t+1,i} = 1$ if $x^{(t)} \in E_i$ and 0 otherwise. γ_{t+1} is called the gain factor and it is a positive non-decreasing sequence satisfying $\sum \gamma_t = \infty$ and $\sum \gamma_t^{\zeta} < \infty$ for some $\zeta \in (1, 2)$.

2. A RESAMPLING-BASED STOCHASTIC APPROXIMATION APPROACH FOR ANALYSIS OF LARGE GEOSTATISTICAL DATA *

In this chapter, we will introduce the method for analysis of large geostatistical data, especially for Gaussian geostatistical model. This is a quite general method which can be applied to any large data set that we need to model the dependency structure between data points. we will first introduce the Gaussian geostatistical model, then describe in detail about what does our algorithm do and why will it work. We give theoretical prove of the asymptotic properties of our estimator. Finally, we will show the power of this method using both simulation studies and real data examples.

2.1 Background

A Gaussian geostatistical model can be written as follows:

$$Y(s_i) = \mu(s_i) + X(s_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N\left(0, \tau^2\right). \tag{2.1}$$

Here, $s_i, i = 1, \cdots, n$ are the locations on a spatial region, $s_i \in \mathbb{R}^2$. $Y(s_i)$ denotes the observation at location s_i, $\mu(s_i)$ denotes the mean of $Y(s_i)$, $\{X(s_i)\}$ denotes a spatial Gaussian process with $E(X(s_i)) = 0$, $Var(X(s_i)) = \sigma^2$, and $corr(X(s_i), X(s_j)) = \rho(\|s_i - s_j\|)$ Basically, this means that the observation can be decomposed into a spatial process and some observational noise. And if we assume the noises follow normal distribution, then its called Gaussian geostatistical model.

The Gaussian geostatistical model is a very popular choice for modeling spatial

*Parts of this chapter are reprinted with permission from "A Resampling-based Stochastic Approximation Method for Analysis of Large Geostatistical Data" by Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P., 2013. J. Amer. Statist. Assoc., 108, 325-339. Copyright [2013] by American Statistical Association.

data. The key characteristic is that we assume the observations from different location to be correlated with each other. So a very important part of the model is to model the structure of covariance matrix. When the sample size is n, the covariance matrix will be a $n \times n$ matrix. The computation becomes very time consuming for large data, because we need to calculate the determinant as well as the inverse of such a big matrix.

There are some existing methods that try to alleviate the computational burden by finding a good approximation. For example, the covariance tapering tries to approximate the covariance matrix by a sparse matrix with lots of zeros, see, for example, Furrer et al. (2006), Kaufman et al. (2008) and Du et al. (2009). The lower dimensional space process approximation tries to approximate the underlying spatial process by find a lower dimensional representation of the spatial process, for example, using smoothing techniques. The likelihood approximation start from the likelihood function and find approximation directly for the likelihood function instead of the covariance structure or the spatial process, see, for example, Fuentes (2007), Matsuda et al. (2009) and Stein et al. (2004). And some others propose to approximate using Markov Random Field (Rue and Tjelmeland, 2002 and Rue and Held, 2005). However, one concern is with how good the approximation is and how much dimension it will reduce. Secondly, even if we assume the approximation is good and it reduces the computational time a lot, still, it will introduce loss of information.

In order to overcome this bottleneck, we propose to use stochastic approximation. So in order to get the parameter estimator, we only use part of the information at each iteration. At each step, we sample a small subset of the large data set. Then we do update of the estimator using stochastic approximation based on that subset we sampled.

Another problem is about the asymptotic behavior of the estimators. A well known fact is that the model is non-identifiable for Gaussian geostatistical data. That is, there exists equivalent probability measures. (Stein, 2004; Zhang, 2004) This means some parameters of the model can not be consistently estimated. This may become an obstacle when studying the asymptotics. For geostatistical data, there are two types of asymptotics, the expanding-domain asymptotics and the infill asymptotics. The former describes the case when we increase the number of samples, we also expand the region to sample from. That is, we keep the sampling density as constant. Infill asymptotics describes the case where we fix the region to sample from. So as we increase the sample size, the sampling density is increased as well. Theories have been established under the expanding-domain situation. The maximum likelihood estimator was shown to be consistent and asymptotically normal for different covariance models. However, the asymptotic behavior is different under infill asymptotics. First of all, not all parameters related with covariance structure are consistently estimated, although they will be consistently estimated after some reparameterization. Secondly, it is not clear about the asymptotic behaviors under infill asymptotics even after the reparameterization (Lahiri, 1996).

In this paper, we set up a series of conditions and showed that under those conditions, the RSA estimator will be consistent and normally distributed after a reparameterization. We achieved this goal by studying two fold of approximations. Firstly, we studied the properties of the stochastic approximation estimator when t goes to infinity. This can be done by following the proof of stochastic approximation. Secondly, we examined the asymptotic properties for estimators of our estimating equation when the sample size goes to infinity. This is done by observing our estimating equation has the form of U statistic.

2.2 Method

We proposed the so-called resampling-based stochastic approximation (RSA) method to solve the estimating problem. We used $Y(s_i), i = 1, \cdots, n$ to denote the complete data with n observation. Suppose each time, we sample m locations from the complete data and write it as $\boldsymbol{S} = (s_1^*, \ldots, s_m^*)$. Then write the observation for those m locations as $\boldsymbol{Z}(\boldsymbol{s}) = (Y(s_1^*), \ldots, Y(s_m^*))^T$. Since $\boldsymbol{Y}$ follows a multivariate normal distribution, we have:

$$\boldsymbol{Z}|\boldsymbol{S} \sim N_m(\boldsymbol{\mu}_z, \Sigma_z). \tag{2.2}$$

Here, $\boldsymbol{\mu}_z = (\mu(s_1^*), \ldots, \mu(s_m^*))^T$, $\Sigma_z = \sigma^2 R_z + \tau^2 I$, and R_z is an $m \times m$ correlation matrix with the (i, j)-th element given by a correlation function $\rho(\|s_i^* - s_j^*\|)$. For simplicity we will just assume that

$$\rho(h) = \exp(-h/\phi), \tag{2.3}$$

$\mu(s_i^*) = \beta_0$, where $\phi > 0$ and β_0 are some unknown parameters. The parameter *phi* determines the strength of the correlation. The bigger *phi* is, the stronger the correlation is. It can be easily extend to the case where we have some covariates that is believed to affect the mean trend over locations. Assume we have p covariate $c_1, \cdots, c_p$, then we can model the mean as

$$\boldsymbol{\mu}_z = \beta_0 \mathbf{1}_m + \sum_{j=1}^{p} \beta_j \boldsymbol{c}_j, \tag{2.4}$$

where $\mathbf{1}_m$ denotes a m-vector of 1's.

The goal is the find the solution of the following estimating equation:

$$\binom{n}{m}^{-1}\sum_{i=1}^{\binom{n}{m}} H(\theta, \boldsymbol{z}_i, \boldsymbol{s}_i) \triangleq \binom{n}{m}^{-1}\sum_{i=1}^{\binom{n}{m}} \frac{\partial \log f_\theta(\boldsymbol{z}_i|\boldsymbol{s}_i)}{\partial\theta} = 0, \tag{2.5}$$

where $f_\theta(\boldsymbol{z}|\boldsymbol{s})$ is a multivariate normal density given by (2.2). The above estimating equation can be viewed as derivative of the mean maximum log-likelihood function defined as follows:

$$\binom{n}{m}^{-1}\sum_{i=1}^{\binom{n}{m}} \log f_\theta(\boldsymbol{z}_i|\boldsymbol{s}_i). \tag{2.6}$$

Or, it can also be defined as derivative of the kullback-Leibler divergence,

$$\mathrm{KL}(f_\theta, g) = -\int\int \log\left(\frac{f_\theta(\boldsymbol{z}|\boldsymbol{s})}{g(\boldsymbol{z}|\boldsymbol{s})}\right) g(\boldsymbol{z}|\boldsymbol{s})g(\boldsymbol{s})d\boldsymbol{z}d\boldsymbol{s}, \tag{2.7}$$

Note that equation (2.5) forms a U statistics with kernel $H(\theta, \boldsymbol{z}_i, \boldsymbol{s}_i)$, which is a vector of dimension that is the same as the number of parameters. This gives very nice properties for the estimator as we will discuss in the following section.

The respective components of $H(\theta, \boldsymbol{z}, \boldsymbol{s})$ in (2.5) are given by

$$\begin{cases} H^{\beta_0}(\theta, \boldsymbol{z}, \boldsymbol{s}) & = \mathbf{1}_m^T\Sigma_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z), \\ H^{\phi}(\theta, \boldsymbol{z}, \boldsymbol{s}) & = -\frac{1}{2}\mathrm{tr}(\Sigma_z^{-1}\sigma^2\frac{dR_z}{d\phi}) + \frac{\sigma^2}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)^T\Sigma_z^{-1}\frac{dR_z}{d\phi}\Sigma_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z), \\ H^{\sigma^2}(\theta, \boldsymbol{z}, \boldsymbol{s}) & = -\frac{1}{2}\mathrm{tr}(\Sigma_z^{-1}R_z) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)'\Sigma_z^{-1}R_z\Sigma_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z), \\ H^{\tau^2}(\theta, \boldsymbol{z}, \boldsymbol{s}) & = -\frac{1}{2}\mathrm{tr}(\Sigma_z^{-1}) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)^T\Sigma_z^{-2}(\boldsymbol{z} - \boldsymbol{\mu}_z), \end{cases} \tag{2.8}$$

where $\frac{dR_z}{d\phi}$ is a $m \times m$-matrix with the (i, j)-th element given by

$$\left(\frac{dR_z}{d\phi}\right)_{ij} = \frac{h_{ij}}{\phi^2}e^{-h_{ij}/\phi},$$

h_{ij} denotes the Euclidean distance between site i and site j and $H^{\beta_0}(\theta, \boldsymbol{z}, \boldsymbol{s})$ denotes

the element in $H(\theta, \boldsymbol{z}, \boldsymbol{s})$ with respect to β_0

We can rewrite equation (2.5) as follows:

$$h(\theta) \triangleq E\{H(\theta, \boldsymbol{z}, \boldsymbol{s})\} = 0, \tag{2.9}$$

Then it has the same form as for the stochastic approximation. Here, each time we draw a subset of size m from the complete data set, we observed one realization of $H(\theta, \boldsymbol{z}, \boldsymbol{s})$.

So following the varying truncation stochastic approximation method, the RSA algorithm follows:

Algorithm 2.2.1. *Resampling-based Stochastic Approximation (RSA) Algorithm*

(i) Draw $(\boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1})$ *from the set* $\{Y(s_1), \ldots, Y(s_n)\}$ *at random and without replacement.*

(ii) Update each component of θ_t *in the following equations:*

$$\begin{cases} \xi_0^{(t+\frac{1}{2})} & = \xi_0^{(t)} + a_{t+1} H_{\xi_0}(\theta_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ \xi_1^{(t+\frac{1}{2})} & = \xi_1^{(t)} + a_{t+1} H_{\xi_1}(\theta_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ \vdots & \\ \xi_p^{(t+\frac{1}{2})} & = \xi_p^{(t)} + a_{t+1} H_{\xi_p}(\theta_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ \phi^{(t+\frac{1}{2})} & = \phi^{(t)} + a_{t+1} H_{\phi}(\theta_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ (\sigma^2)^{(t+\frac{1}{2})} & = (\sigma^2)^{(t)} + a_{t+1} H_{\sigma^2}(\theta_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ (\tau^2)^{(t+\frac{1}{2})} & = (\tau^2)^{(t)} + a_{t+1} H_{\tau^2}(\theta_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}). \end{cases}$$

(iii) If $\|\theta_{t+\frac{1}{2}} - \theta_t\| \leq b_t$ *and* $\theta_{t+\frac{1}{2}} \in \mathcal{K}_{\pi_k}$, *then set* $\theta_{t+1} = \theta_{t+\frac{1}{2}}$ *and* $\pi_{t+1} = \pi_t$;

otherwise, set $\theta_{t+1} = \mathbb{T}(\theta_t)$ *and* $\pi_{t+1} = \pi_t + 1$.

Here, $\|\cdot\|$ denote the Euclidean norm of a vector, a_t and b_t are two non-increasing positive sequences that goes to zero. $\{\mathcal{K}_s, s \geq 0\}$ be a sequence of compact subsets of Θ that satisfies

$$\bigcup_{s\geq 0} \mathcal{K}_s = \Theta, \quad \text{and} \quad \mathcal{K}_s \subset \text{int}(\mathcal{K}_{s+1}), \quad s \geq 0, \tag{2.10}$$

where $\text{int}(A)$ denotes the interior of set A.

For the RSA Algorithm described above, we adopted a slightly different version of stochastic approximation method, the so-called varying truncation stochastic approximation algorithm proposed by Andrieu et. al. in 2005. The main difference is that at iteration t, we preset a certain boundary that is conditional on the previous estimator θ_{t-1}. If θ_t is out of that boundary, then we set θ_t as a project of θ_{t-1}. And this is called a truncation. So using this method, if we can show that the number of truncation is finite, then the series θ_t is bounded by its definition. In practice, we can set the boundary to be very large such that when do computation, there is no need for truncation.

2.3 Theoretical Results

To show the RSA algorithm will be consistent estimator, we need to do it two fold. Firstly, it is necessary to show that equations (2.5) will give consistent estimator. So if we denote the estimator by $\tilde{\theta}_n$ and the true parameter is θ_0 then we would like to show that as the sample size n increases, $\tilde{\theta}_n$ will converge to θ_0 and study what is the convergence rate. Secondly, let $\hat{\theta}_n^{(t)}$ be the estimate obtained at step t using RSA. Then we would like to show that as the number of iterations t increases, $\hat{\theta}_n^{(t)}$ will converge to $\tilde{\theta}_n$ and we are interested at the convergence rate. In the following

subsections we will give theorems to show the asymptotic properties for $\tilde{\theta}_n$ and $\hat{\theta}_n^{(t)}$ respectively. All proofs for this section is given in Appendix A.

2.3.1 Infill Asymptotics of $\tilde{\theta}_n$

In this subsection we show the theoretical properties of $\tilde{\theta}_n$. The motivation for the prove is based on the observation that estimating equation (2.5) has the form of a U statistics. So we give general results for U statistics in lemma 2.3.1 - lemma 2.3.4 and results for $\tilde{\theta}_n$ are given in theorem 2.3.1 - theorem 2.3.2.

Let

$$U_n = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \psi(X_1^{(i)}, \ldots, X_m^{(i)}), \tag{2.11}$$

be a U-statistic defined on the random sample $\{X_1, \ldots, X_n\}$, where $\psi(\cdot)$ is called the kernel of the U-statistic.

The following lemma shows the convergence of the U-statistic when $X_1, \cdots, X_m$ are dependent.

Lemma 2.3.1. *Let $\{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded, stationary random field. If the mapping $(x_1, \ldots, x_m) \mapsto \psi(x_1, \ldots, x_m)$ is continuous (a.e.) and $E|\psi(X_1, \ldots, X_m)|^2 < \infty$, then, as $n \to \infty$,*

$$U_n \to E(\psi(X_1, \ldots, X_m)) \quad \textit{in probability.}$$

For RSA, the U statistic depends on a set of parameters $\boldsymbol{\theta}$, so in order to show the dependence on $\boldsymbol{\theta}$, we define the following:

$$U_n(\theta) = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \psi_\theta(X_1^{(i)}, \ldots, X_m^{(i)}) \quad \text{and} \quad U(\theta) = E(\psi_\theta(X_1, \ldots, X_m)). \tag{2.12}$$

Lemma 2.3.1 shows that $U_n(\boldsymbol{\theta}) \to U(\boldsymbol{\theta})$ in probability for each fixed $\boldsymbol{\theta}$. We are

going to show that the value that minimizes $U_n(\boldsymbol{\theta})$ will converge to that of $U(\boldsymbol{\theta})$ under some mild conditions.

Lemma 2.3.2. *Let $\{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded, stationary random field. Let $\Theta_0 = \{\theta^* \in \Theta : U(\theta^*) = \sup_\theta U(\theta)\}$ denote the set of global maximizers of $U(\theta)$. Assume the following conditions hold:*

(i) The mapping $\theta \mapsto \psi_\theta(X_1, \ldots, X_m)$ is continuous for almost all $(X_1, \ldots, X_m)$ and satisfies

$$E|\psi_\theta(X_1, \ldots, X_m)|^2 < \infty. \tag{2.13}$$

(ii) The mapping $(x_1, \ldots, x_m) \mapsto \sup_{\theta \in O} \psi_\theta(x_1, \ldots, x_m)$ is measurable for every sufficiently small ball $O \subset \Theta$ and satisfies

$$E|\sup_{\theta \in O} \psi_\theta(X_1, \ldots, X_m)|^2 < \infty. \tag{2.14}$$

Then for any estimators $\tilde{\theta}_n$ such that $U_n(\tilde{\theta}_n) \geq U_n(\theta^*) + o_p(1)$ for some $\theta^* \in \Theta_0$, for every $\epsilon > 0$ and every compact set $\mathcal{K} \subset \Theta$,

$$P(d(\tilde{\theta}_n, \Theta_0) \geq \epsilon \text{ and } \tilde{\theta}_n \in \mathcal{K}) \to 0,$$

where $d(\cdot, \cdot)$ denotes a distance metric.

To study the infill asymptotics of $\tilde{\theta}_n$, we define

$$l_\theta(\boldsymbol{z}, \boldsymbol{s}) = \log f_\theta(\boldsymbol{z}|\boldsymbol{s}), \quad M(\theta) = E[l_\theta(\boldsymbol{z}, \boldsymbol{s})], \quad M_n(\theta) = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} l_\theta(\boldsymbol{z}_i, \boldsymbol{s}_i). \tag{2.15}$$

Thus, $M_n(\theta)$ forms a U-statistic estimator of $M(\theta)$ with the kernel $l_\theta(\boldsymbol{z})$, and minimizing (2.5) is equivalent to maximizing $M_n(\theta)$.

The following theorem shows that $\tilde{\theta}_n$ will converge to the set $\Theta_0 = \{\theta^* : El_{\theta^*}(\boldsymbol{Z}, \boldsymbol{S}) = \sup_{\theta\in\Theta} El_\theta(\boldsymbol{Z}, \boldsymbol{S})\}$ in probability. Not that it is a result that is independent of the subset sample size m.

Theorem 2.3.1. *Let $\{Y(s_1), \ldots, Y(s_n)\}$ denote a random sample drawn from the spatial Gaussian model (2.1) defined on a bounded region, let $\tilde{\theta}_n$ denote a solution to (2.5), and let $\Theta_0 = \{\theta^* \in \Theta : El_{\theta^*}(\boldsymbol{Z}, \boldsymbol{S}) = \sup_{\theta\in\Theta} El_\theta(\boldsymbol{Z}, \boldsymbol{S})\}$, where $(\boldsymbol{Z}, \boldsymbol{S})$ denotes a random sample of size m drawn from model (2.1). Assume Θ is compact, then for every $\epsilon > 0$,*

$$P(d(\tilde{\theta}_n, \Theta_0) \geq \epsilon) \to 0$$

as $n \to \infty$, where $d(\cdot, \cdot)$ denotes a distance metric.

The following lemma shows normality of the U-statistic when $X_1, \cdots, X_m$ are dependent. To prove this lemma, we assume that the function $\psi(x_1, \ldots, x_m)$ is continuous (a.e.) and $E|\psi(X_1, \ldots, X_m)|^2 < \infty$. In addition, we impose some constraints on the sampling procedure of $\mathcal{S}_n = \{X_1, \ldots, X_n\}$: $\mathcal{S}_n$ is drawn through a procedure which ensures that for $1 \leq k \leq m-1$ and any $\alpha > 0$,

$$E|\psi_{k,n}(X_1, \ldots, X_k)|^2 \text{ is uniformly bounded w.r.t. } n \text{ and } n^\alpha \sigma_{k,n}^2 \to \infty \text{ as } n \to \infty, \tag{2.16}$$

where $\psi_{k,n}(x_1, \ldots, x_k) = E\{\psi(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_k = x_k, \mathcal{S}_n\}$ is the conditional expectation of $\psi(X_1, \ldots, X_m)$ based on the finite population $\mathcal{S}_n$, and $\sigma_{k,n}^2 = \mathrm{Var}(\psi_{k,n}(X_1, \ldots, X_k))$. Let $\psi_k(x_1, \ldots, x_k) = E\{\psi(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_k = x_k\}$. Be aware that $E(|\psi_{k,n}(X_1, \ldots, X_k)|^2)$ is actually the second-order sample moments of ψ_k and $\sigma_{k,n}^2$ the sample variance of ψ_k. This assumption essentially requires that the sample $\{X_1, \ldots, X_n\}$ resembles the underlying random field such that $\sigma_{k,n}^2$ converges to a constant as $n \to \infty$. This assumption is satisfied except that the sam-

pling procedure is degenerated to drawing samples from a single site or the function $\psi_k(\cdot)$ is degenerated to taking a constant value.

Lemma 2.3.3. *Let $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded, stationary random field. Consider the U-statistic defined in (2.11). Assume the following conditions hold:*

(i) The function $\psi(x_1, \ldots, x_m)$ is continuous (a.e.), and $E|\psi(X_1, \ldots, X_m)|^2 < \infty$.

(ii) $\mathcal{S}_n$ satisfies the condition (2.16).

Then, as $n \to \infty$,

$$(U_n - E(\psi(X_1, \ldots, X_m)))/\sqrt{Var(U_n)} \Rightarrow N(0, 1),$$

where $\Rightarrow$ denotes the convergence in distribution, and $N(0, 1)$ denotes the standard normal distribution.

Lemma 2.3.4 shows the asymptotic normality of the estimator $\tilde{\theta}_n$, which maximizes $U_n(\theta)$ defined in (2.12).

Lemma 2.3.4. *Let $\{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded stationary random field. Assume the following conditions hold:*

(i) The parameter space Θ is compact.

(ii) The kernel $\psi_\theta(\cdot)$ is twice continuously differentiable on the interior of Θ, and satisfies

$$E|\psi_\theta(X_1, \ldots, X_m)|^2 < \infty, \quad E\|\frac{\partial}{\partial\theta}\psi_\theta(X_1, \ldots, X_m)\|^2 < \infty,$$
$$E\|\frac{\partial^2}{\partial\theta^2}\psi_\theta(X_1, \ldots, X_m)\|^2 < \infty. \tag{2.17}$$

(iii) The mapping $(x_1, \ldots, x_m) \mapsto \sup_{\theta \in O} \psi_\theta(x_1, \ldots, x_m)$ *is measurable for every sufficiently small ball* $O \subset \Theta$ *and satisfies*

$$E|\sup_{\theta \in O} \psi_\theta(X_1, \ldots, X_m)|^2 < \infty. \tag{2.18}$$

(iv) $\mathcal{S}_n$ *satisfies the condition (2.16); that is, there exists a constant* C *such that for* $1 \leq k \leq m-1$,

$$\sup_n E(\|\frac{\partial}{\partial \theta}\psi_{\theta,k}(X_1, \ldots, X_k)\|^2 | \mathcal{S}_n) < C, \quad a.s.,$$

where $\frac{\partial}{\partial \theta}\psi_{\theta,k}(x_1, \ldots, x_k) = E\left\{\frac{\partial}{\partial \theta}\psi_\theta(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_k = x_k)\right\}$. *In addition, for any* $\alpha > 0$ *and* $1 \leq k \leq m-1$, $n^\alpha \|\Sigma_{k,n}\| \to \infty$ *as* $n \to \infty$, *where* $\Sigma_{k,n}$ *denotes the sample covariance matrix of* $\frac{\partial}{\partial \theta}\psi_{\theta,k}(X_1, \ldots, X_k)$.

Then for any estimators $\tilde{\theta}_n$ *such that* $U_n(\tilde{\theta}_n) \geq U_n(\theta^*) + o_p(1)$ *for some* $\theta^* \in \Theta_0$,

$$\tilde{\theta}_n - \theta^* \Rightarrow N(0, H_*^{-1} \Sigma H_*^{-1}),$$

where $H_* = E\left\{\frac{\partial^2 \psi_\theta(X_1, \ldots, X_m)}{\partial \theta \partial \theta'}|_{\theta=\theta^*}\right\}$ *is the expected Hessian of* $\psi_\theta(X_1, \ldots, X_m)$ *at* θ^*, *and* Σ *is the covariance matrix of the* U*-statistic defined by the kernel* $\frac{\partial \psi_\theta(X_1, \ldots, X_m)}{\partial \theta}|_{\theta=\theta^*}$.

Theorem 2.3.2 concerns the asymptotic normality of the minimizer of the Kullback-Leibler divergence.

Theorem 2.3.2. *Let* $\{Y(s_1), \ldots, Y(s_n)\}$ *be a random sample drawn from the spatial Gaussian model (2.1) defined on a bounded region, let* $\tilde{\theta}_n$ *denote a solution to (2.5), and let* $\Theta_0 = \{\theta^* \in \Theta : El_{\theta^*}(\boldsymbol{Z}, \boldsymbol{S}) = \sup_{\theta \in \Theta} El_\theta(\boldsymbol{Z}, \boldsymbol{S})\}$, *where* $(\boldsymbol{Z}, \boldsymbol{S})$ *denotes a random sample of size* m *drawn from model (2.1). Assume that* Θ *is compact, the model is identifiable, and the sampling procedure of* $\{Y(s_1), \ldots, Y(s_n)\}$ *satisfies the*

condition (iv) of Lemma 2.3.4 (with $\psi_\theta(\cdot) = l_\theta(\cdot)$). Then

$$\tilde{\theta}_n - \theta^* \Rightarrow N(0, H_*^{-1}\Sigma H_*^{-1}), \tag{2.19}$$

where $H_* = E\left\{\frac{\partial^2 l_\theta(\boldsymbol{Z},\boldsymbol{S})}{\partial\theta\partial\theta'}|_{\theta=\theta^*}\right\}$ is the expected Hessian of $l_\theta(\boldsymbol{z})$ at θ^*, and Σ is the covariance matrix of the U-statistic defined by the kernel $\frac{\partial l_\theta(\boldsymbol{z},\boldsymbol{s})}{\partial\theta}|_{\theta=\theta^*}$.

2.3.2 Stochastic Approximation Asymptotics of $\hat{\theta}_n^{(t)}$

Algorithm 2.3.1. *Varying Truncation Stochastic Approximation*

(i) Generate $X_{t+1} \sim g_{\theta_t}(x)$, where t indexes the iteration.

(ii) Set $\theta_{t+\frac{1}{2}} = \theta_t + a_t H(\theta_t, X_{t+1})$, where a_t is the gain factor.

(iii) If $\|\theta_{t+\frac{1}{2}} - \theta_t\| \leq b_t$ and $\theta_{t+\frac{1}{2}} \in \mathcal{K}_{\pi_k}$, then set $\theta_{t+1} = \theta_{t+\frac{1}{2}}$ and $\pi_{t+1} = \pi_t$; otherwise, set $\theta_{t+1} = \mathbb{T}(\theta_t)$ and $\pi_{t+1} = \pi_t + 1$. Here $\mathbb{T}$ and π_t are defined as in algorithm 2.2.1.

The varying truncation stochastic approximation algorithm can been seen as a special case of varying truncation stochastic approximation MCMC algorithm given in Andrieu et al. (2005). Since the only difference is that at each iteration the new sample X_{t+1} is generated through an exact sampler instead of a MCMC sampler. The following two lemmas are a restatement of what's given in their paper that can be applied to the above algorithm.

Lemma 2.3.5. *Assume the conditions (A_1), (A_2) and (A_4) (given in Introduction) hold. Let k_π denote the iteration number at which the π-th truncation occurs in the simulation. Let $\mathcal{X}_0 \subset \mathcal{X}$ be such that $\sup_{x\in\mathcal{X}_0} V(x) < \infty$ and $\mathcal{K}_0 \subset \mathcal{V}_{C_0}$, where $\mathcal{V}_{C_0}$ is defined in (A_1). Let $\{\theta_t\}$ be given by Algorithm 2.3.1. Then there exists almost*

surely a number, denoted by π_s, such that $k_{\pi_s} < \infty$ and $k_{\pi_s+1} = \infty$; that is, $\{\theta_t\}$ can be kept in a compact set almost surely. In addition,

$$d(\theta_t, \mathcal{L}) \to 0, \quad a.s.,$$

where $\mathcal{L}$ is defined in (A_1), and $d(\theta, \mathcal{L}) = \inf_{\theta'}\{\|\theta - \theta'\| : \theta' \in \mathcal{L}\}$ denotes a distance measure induced by the Euclidian norm.

Lemma 2.3.6. *Assume the conditions (A_1), (A_2), (A_3), and (A_4) (given in Appendix B) hold. Let the simulation start with a point $(\theta_0, X_0) \in \mathcal{K}_0 \times \mathcal{X}$, where $\mathcal{K}_0 \subset \mathcal{V}_{C_0}$ (defined in (A_1)) and $\sup_{X\in\mathcal{X}} V(X) < \infty$. Let $\{\theta_t\}$ be given by Algorithm 2.3.1. Conditioned on $\Lambda(\theta_*) = \{\theta_t \to \theta_*\}$,*

$$\frac{\theta_t - \theta_*}{\sqrt{a_t}} \Rightarrow \mathbb{N}(0, \Sigma_{sa}), \tag{2.20}$$

where $\theta_* \in \mathcal{L}$ as defined in (A_1), $\mathbb{N}(\cdot, \cdot)$ denotes the Gaussian distribution and

$$\Sigma_{sa} = \int_0^\infty e^{(F'+\zeta I)t}\Gamma e^{(F+\zeta I)t}dt, \tag{2.21}$$

where F is defined in (A_3), ζ is defined in (A.4), and Γ is defined by

$$\frac{1}{N}\sum_{t=1}^{N} E(\epsilon_{t+1}\epsilon_{t+1}^T|\mathcal{F}_t) \to \Gamma,$$

with $\epsilon_{t+1} = H(\theta_t, X_{t+1}) - h(\theta_t)$, and $\mathcal{F}_t = \sigma\{\theta_0, X_0, \ldots, \theta_t, X_t\}$ being a σ-algebra formed by $\{\theta_0, X_0, \ldots, \theta_t, X_t\}$.

Based on the above results for algorithm 2.3.1. It is easily to get the following theorems by checking that the conditions (A_1), (A_2), (A_3), and (A_4) hold for RSA algorithm.

Theorem 2.3.3. *Let $\{Y(s_1), \ldots, Y(s_n)\}$ be a random sample drawn from a spatial Gaussian model (2.1), which is defined on a bounded region and has an exponential correlation function. Let $\mathcal{L} = \{\theta : \partial KL(f_\theta, \tilde{g})/\partial\theta = 0\}$ denote the set of solutions to the system of equations (2.5). Assume Θ is compact and let $\{\hat{\theta}_n^{(t)}\}$ be given by Algorithm 2.2.1. Then $\lim_{t\to\infty} d(\hat{\theta}_n^{(t)}, \mathcal{L}) = 0$ a.s. as $t \to \infty$.*

Theorem 2.3.4. *Let $\{Y(s_1), \ldots, Y(s_n)\}$ be a random sample drawn from a spatial Gaussian model (2.1), which is defined on a bounded region and has an exponential correlation function. Let $\mathcal{L} = \{\theta : \partial KL(f_\theta, \tilde{g})/\partial\theta = 0\}$ denote the set of solutions to the system of equations (2.5). Assume the model (2.1) is identifiable and Θ is compact. Let $\{\hat{\theta}_n^{(t)}\}$ be given by Algorithm 2.2.1. Then, given $\Lambda(\theta_*) = \{\hat{\theta}_n^{(t)} \to \theta_*\}$,*

$$\frac{\hat{\theta}_n^{(t)} - \theta_*}{\sqrt{a_t}} \Rightarrow \mathbb{N}(0, \Sigma_{sa}), \tag{2.22}$$

where $\theta_* \in \mathcal{L}$ and Σ_{sa} is as defined in Lemma 2.3.6.

As a summary of Theorem 2.3.2 and Theorem 2.3.4, we note that $\hat{\theta}_n^{(t)}$ is asymptotically normally distributed, and its asymptotic distribution is given by

$$\hat{\theta}_n^{(t)} \Rightarrow N(\theta^*, a_t\Sigma_{sa} + H_*^{-1}\Sigma H_*^{-1}),$$

where H_* and Σ are given in Theorem 2.3.2 and Σ_{sa} is given in Theorem 2.3.4. The term $a_t\Sigma_{sa}$ of the covariance matrix represents the part of Monte Carlo error in $\hat{\theta}_n^{(t)}$.

2.4 Simulation Examples

2.4.1 A Comparison with MLE

In this example, we consider a geostatistical model with measurement errors. The model is specified by (2.1) with $\beta_0 = \beta_1 = 1$, $\phi = 25$, $\sigma^2 = 1$, $\tau^2 = 1.0$, and

Table 2.1: A comparison with MLE for 50 simulated datasets with nugget effect.

Estimator	m	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}/\hat{\sigma}^2$	$\hat{\tau}^2$	CPU(m)
RSA	100	1.022(0.068)	0.998(0.009)	19.278(0.768)	0.939(0.010)	0.3
	300	1.016(0.065)	1.000(0.007)	22.046(0.684)	0.974(0.009)	6.4
	500	1.013(0.064)	1.001(0.007)	23.084(0.675)	0.977(0.008)	29.3
	700	0.997(0.063)	0.999(0.006)	24.023(0.659)	0.993(0.007)	81.5
MLE	—	1.000(0.061)	1.000(0.006)	25.269(0.72)	0.999(0.007)	19.4
True	—	1.000	1.000	25.000	1.0	—

the explanatory variable $\boldsymbol{c}_1$ is generated from a Gaussian distribution with mean 0 and standard deviation 0.5. Using the package *geoR* (Ribeiro Jr and Diggle, 2010), we simulated 50 datasets of size $n = 2000$ with the sampling sites being uniformly distributed in a bounded region of $[0, 100] \times [0, 100]$. We use this example to illustrate how the RSA estimator is related to the MLE.

For each dataset, RSA was run four times with $m = 100$, 300, 500 and 700, respectively. We set $a_0 = 0.01$ for the runs with $m = 100$, 300 and 500 and $a_0 = 0.001$ for the run with $m = 700$. Each run consisted of 2500 iterations. The numerical results are summarized in Table 2.1.

2.5 Data Examples

We consider the precipitation data from the National Climatic Data Center for the years 1895 to 1997. It available at `/www.image.ucar.edu/GSP/Data/US.monthly.met/`. In this analysis, we analyze the monthly total precipitation anomalies, which are defined as the monthly totals standardized by the long-run mean and standard deviation for each station. The data we considered is the precipitation anomalies of April 1948. The reason why we choose to work on this dataset is two fold. Firstly, the dataset is large, consisting of 11,918 stations. Note that part of the data was

Table 2.2: Numerical results of RSA for monthly precipitation in April 1948.

Method	m	$\hat{\beta}_0$	$\hat{\phi}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	CPU(m)
RSA	500	0.163	183.71	0.825	0.059	29.6
		(0.000)	(0.45)	(0.003)	(0.000)	
	700	0.161	179.38	0.829	0.057	84.1
		(0.001)	(1.15)	(0.001)	(0.000)	
MLE		0.138	164.20	0.807	0.057	10,340.4

imputed by Johns *et al.* (2003), but for the purpose of illustration, we follow Furrer (2006) to treat all data as real observations. Secondly, the data show no obvious non-stationarity or anisotropy. Otherwise, it would require a more complicated model, such as a mixture spatial model, than is considered here.

In our analysis, we first divide the data into two parts, a random subset of 11,000 observations as the training set and the remaining 918 observations as the test set. RSA was applied to the training data with $m = 500$ and $m = 700$. For each setting of m, RSA was run for 5 times with $a_0 = 0.001$ and each run consisted of 2500 iterations. The results are summarized in Table 2.2. It indicates that RSA works very stable for this example. The standard deviations of all parameters are quite small. We also used krigging (Stein, 1999) to do prediction on the test set, and the prediction performance is quite well.

3. SIMULATED STOCHASTIC APPROXIMATION ANNEALING FOR GLOBAL OPTIMIZATION WITH A SQUARE ROOT COOLING SCHEDULE

In this chapter, We will introduce a new global optimization method, the so-called simulated stochastic approximation annealing (SAA) algorithm. This is a general optimization method, which can be applied to functions with high dimensional parameters. We will first describe the simulated annealing algorithm (Kirkpatrik et al., 1983 and Cerny, 1985), then introduce the proposed new method. We will discuss intuitively why our method is valid and show the theoretical results. Finally, We will show some numerical results to demonstrate the effect and advantages of SAA algorithm.

3.1 Background

Simulated annealing is a Monte Carlo method that aims to find the global optima. It was independently described by Scott Kirkpatrick, C. Daniel Gelatt and Mario P. Vecchi in 1983 and by Vlado Cerny in 1985. After its introduction, it has been widely used in many different area. Given a function $U(x)$ that we want to minimize, simulated annealing method aims to sample from the following distribution

$$f_\tau(x) \propto \exp\{-\frac{U(x)}{\tau}\} \tag{3.1}$$

with τ being a changing parameter that is nonincreasing. The function $U(x)$ is called the energy function. The parameter τ is called temperature and the nonincreasing path it follows is called a cooling schedule. If we make the temperature to be a positive number close to 0, then essentially sampling from distribution 3.1 is equivalent

to locating the global minima of function $U(x)$. The simulated annealing algorithm works as follows:

Algorithm 3.1.1. *(Simulated Annealing)*

1. *Initialize the simulation at temperature τ_1 and an arbitrary sample $x_0 \in \mathcal{X}$.*

2. *At each temperature τ_i, simulate the distribution $f_{\tau_i}(x)$ for n_i iterations using the MH sampler. Pass the final sample to the next lower temperature level as the initial sample.*

This method is easy to work with. And when the temperature decreases slowly, it is guaranteed that the global minima will be reached. However, as already discussed, the problem lies in the strict restriction put on the cooling schedule to ensure convergence.

3.2 The Simulated Stochastic Approximation Annealing Algorithm

Now let's introduce the set up for simulated stochastic approximation annealing (SAA) algorithm. SAA algorithm tries to solve the exact same problem as simulated annealing, that is, to minimize the energy function $U(x)$. The difference is that we proposed a slightly different sampling scheme.

Be reminded that when the temperature is very close to zero, the shape for density function $f_\tau(x)$ described in equation 3.1 will be very spiky so them sample can easily get trapped. So inspired by stochastic approximation monte carlo (SAMC) method, we devide the sample space $\mathcal{X}$ into m disjoint subregions and adjust the density function based on its volume within each subregions. Denote the subregions as $E_1, \cdots, E_m$, and define them as follows: Let $E_1, ..., E_m$ denote a partition of the

sample space $\mathcal{X}$, which are made according to the energy function as follows:

$$E_1 = \{x : U(x) \leq u_1\}, E_2 = \{x : u_1 < U(x) \leq u_2\}, \ldots, E_{m-1} = \{x : u_{m-2} < U(x) \leq u_{m-1}\}, E_m = \{x : U(x) > u_{m-1}\}, \tag{3.2}$$

where $u_1 < u_2 < \ldots < u_{m-1}$ are pre-specified numbers. We define the weight vector $w_\tau = (w_\tau^{(1)}, \ldots, w_\tau^{(m)})$, $w_\tau^{(i)} = \int_{E_i} e^{-U(x)/\tau} dx$. Then SAA algorithm aims to sample from the following distribution.

$$f_{w_\tau,\tau}(x) \propto \sum_{i=1}^{m} \frac{\pi_i e^{-U(x)/\tau}}{w_\tau^{(i)}} I(x \in E_i), \tag{3.3}$$

Here π_i's satisfy the constraints: $\pi_i > 0$ for all i and $\sum_{i=1}^m \pi_i = 1$. If we integrate equation 3.3 on each subregion, then it is easy to see that the sampling frequency on subregion E_i will be equal to π_i. This gives SAA very good properties even when the temperature is very close to zero. So if we know the weight vector $\mathbf{w}_\tau$, then we can simply use a NH sampling scheme and get the global optima. Although the weight vector is unknown to us, in the process of sampling, we do know the actual sampling frequencies for each subregion, which is somehow an indicator for the area of each subregion. In order to accommodate the fact that all w_i's are greater than zero, we set $\theta_\tau^{(i)} = \log(w_\tau^{(i)}/\pi_i)$ for $i = 1, \ldots, m$, let $\theta_\tau = (\theta_\tau^{(1)}, \ldots, \theta_\tau^{(m)})$, let θ_t denote the working estimator of θ_τ at iteration t, and let Θ denote the space of θ_t. Then the SAA algorithm works as follows:

Let $\{\mathcal{M}_k, k = 0, 1, \ldots\}$ be a sequence of positive numbers increasingly diverging to infinity, which work as truncation bounds of $\{\theta_t\}$. Let σ_t be a counter for the number of truncations up to iteration t, and $\sigma_0 = 0$. Let $\tilde{\theta}_0$ be a fixed point in Θ. Fix an arbitrary initial value θ_0, then SAA iterates as follows:

Algorithm 3.2.1. *(SAA Algorithm)*

1. *(Sampling) Simulate a sample X_{t+1} with a single MH update, which starts with X_t and leaves the following distribution invariant:*

$$f_{\theta_t,\tau_{t+1}}(x) \propto \sum_{i=1}^{m} \exp\left\{-U(x)/\tau_{t+1} - \theta_t^{(i)}\right\} I(x \in E_i), \tag{3.4}$$

   where $I(\cdot)$ is the indicator function.

2. *(θ-updating) Set*

$$\theta_{t+\frac{1}{2}} = \theta_t + \gamma_{t+1} H_{\tau_{t+1}}(\theta_t, x_{t+1}), \tag{3.5}$$

   where $H_{\tau_{t+1}}(\theta_t, x_{t+1}) = \boldsymbol{e}_{t+1} - \boldsymbol{\pi}$, $\boldsymbol{e}_{t+1} = (I(x_{t+1} \in E_1), ..., I(x_{t+1} \in E_m))$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$.

3. *(Truncation) If $\|\theta_{t+\frac{1}{2}}\| \leq \mathcal{M}_{\sigma_t}$, set $\theta_{t+1} = \theta_{t+\frac{1}{2}}$; otherwise, set $\theta_{t+1} = \tilde{\theta}_0$ and $\sigma_{t+1} = \sigma_t + 1$.*

Using SAA algorithm described above, we achieved two goals at the same time.

1. We used samples to estimate weights or $\boldsymbol{\theta}$ for each subregion.

2. Based on the estimates of θ_i's we sampled from the desired distribution 3.3

It is reasonable to believe that under some mild condition, SAA algorithm gives estimator $\boldsymbol{\theta}$ that converges to the true value and provide a sampling scheme that helps to avoid local trap problem. We will give formal statement and proof for the above assertion.

3.3 Convergence

We can view SAA in a different perspective. It can be views as a SAMCMC algorithm that solves the following integration equation:

$$h_{\tau_*}(\theta) = \int H_{\tau_*}(\theta, x) f_{\theta,\tau_*}(x) dx = 0, \tag{3.6}$$

where $f_{\theta,\tau_*}(x)$ denotes a density function dependent on θ and the limiting temperature τ_*. And we use θ_* to denote a solution to the target equation (3.6). Because the temperature is changing over time, so actually, we can write the working target sampling distribution at iteration t as:

$$h_{\tau_t}(\theta) = \int H_{\tau_t}(\theta, x) f_{\theta,\tau_t}(x) dx = 0, \quad t = 1, 2, \ldots. \tag{3.7}$$

Here $f_{\theta,\tau_t}(x)$ is a density function dependent on θ and the temperature τ_t.

In order to show the convergence of SAA algorithm, we follow the general procedure for any stochastic approximation algorithm. That is, we want to show the mean field function, the observational noise and the step size to have some nice properties.

1. We want to make sure the mean function is stable enough that if our sample is at a place that is close to the true values, then there is a big chance that it will go toward the right direction.

2. The observational noise can be canceled out during iterations in some sense.

3. The gain factor is large at the beginning that make the algorithm to be able to search through the whole parameter space. Also, it should be small enough when number of iterations become large such that the estimator will not jump around towards the end of the iteration.

We will describe each of the conditions sufficient for the SAA algorithm to converge. Those conditions are not the necessary conditions.

It is easy to get that the mean field function of SAA is given by

$$h_\tau(\theta) = \int H_\tau(\theta, x) f_{\theta,\tau}(x) dx = \left(\frac{S_\tau^{(1)}(\theta)}{S_\tau(\theta)} - \pi_1, \ldots, \frac{S_\tau^{(m)}(\theta)}{S_\tau(\theta)} - \pi_m \right), \tag{3.8}$$

where $S_\tau^{(i)}(\theta) = \int_{E_i} e^{-U(x)/\tau} dx / e^{\theta^{(i)}}$, and $S_\tau(\theta) = \sum_{i=1}^m S_\tau^{(i)}(\theta)$. And following Liang et. al. (2007), we define

$$v_\tau(\theta) = \frac{1}{2} \sum_{i=1}^m \left(\frac{S_\tau^{(i)}(\theta)}{S_\tau(\theta)} - \pi_i \right)^2, \tag{3.9}$$

which is the same function as in equation (1.1). Again, $h_\tau(\theta)$ is the derivative of $v_\tau(\theta)$ It is known as Lyapunov function in the literature of stochastic approximation. In the setting of SAA algorithm, both $h_\tau(\theta)$ and $v_\tau(\theta)$ have very nice properties by noticing that they are both bounded on the space $\Theta \times \mathbb{T}$

Conditions on mean field function

(A_1') (Stability conditions)

(i) The function $h_\tau(\theta)$ is bounded and continuously differentiable with respect to both θ and τ, and there exists a non-negative, upper bounded, and continuously differentiable function $v_\tau(\theta)$ such that for any $\Delta > \delta > 0$,

$$\sup_{\delta \le d((\theta,\tau),\mathcal{L}) \le \Delta} \nabla_\theta^T v_\tau(\theta) h_\tau(\theta) < 0, \tag{3.10}$$

where $\mathcal{L} = \{(\theta, \tau) : h_\tau(\theta) = 0, \theta \in \Theta, \tau \in \mathbb{T}\}$ is the zero set of $h_\tau(\theta)$, and $d(z, S) = \inf_y \{\|z - y\| : y \in S\}$. Further, the set $v(\mathcal{L}) = \{v_\tau(\theta) : (\theta, \tau) \in \mathcal{L}\}$ is nowhere dense.

(ii) Both $\nabla_\theta v_\tau(\theta)$ and $\nabla_\tau v_\tau(\theta)$ are bounded over $\Theta \times \mathbb{T}$. Here ∇_θ denotes the gradient with respect to *theta* and ∇_τ denotes the gradient with respect to *tau* In addition, for any compact set $\mathcal{K} \subset \Theta$, there exists a constant $0 < c < \infty$ such that

$$\begin{aligned} &\sup_{(\theta,\theta')\in\mathcal{K}\times\mathcal{K},\tau\in\mathbb{T}} \|\nabla_\theta v_\tau(\theta) - \nabla_\theta v_\tau(\theta')\| \le c\|\theta - \theta'\|, \\ &\sup_{\theta\in\mathcal{K},(\tau,\tau')\in\mathbb{T}\times\mathbb{T}} \|\nabla_\theta v_\tau(\theta) - \nabla_\theta v_{\tau'}(\theta)\| \le c|\tau - \tau'|, \\ &\sup_{\theta\in\mathcal{K},(\tau,\tau')\in\mathbb{T}\times\mathbb{T}} \|h_\tau(\theta) - h_{\tau'}(\theta)\| \le c|\tau - \tau'|. \end{aligned} \tag{3.11}$$

Conditions on observation noise

In the literature, conditions put on observation noise can be categorized into two approaches. One approach put conditions on the observation noise directly and the other approach put conditions through the Markov transition kernel.

In this paper, we assume that for any $\theta \in \Theta$ and $\tau \in \mathbb{T}$, the Markov transition kernel $P_{\theta,\tau}$ satisfies the Doeblin condition, which is equivalent to assuming that the resulting Markov chain is uniformly ergodic (Nummelin, 1984, Theorem 6.15).

(A_2') (Doeblin condition) For any given $\theta \in \Theta$ and $\tau \in \mathbb{T}$, the Markov transition kernel $P_{\theta,\tau}$ is irreducible and aperiodic. In addition, there exist an integer l, $0 < \delta < 1$, and a probability measure ν such that for any compact subset $\mathcal{K} \subset \Theta$,

$$\inf_{\theta\in\mathcal{K},\tau\in\mathbb{T}} P^l_{\theta,\tau}(x, A) \ge \delta\nu(A), \quad \forall x \in \mathcal{X},\ \forall A \in \mathcal{B}_\mathcal{X},$$

where $\mathcal{B}_\mathcal{X}$ denotes the Borel set of $\mathcal{X}$; that is, the whole support $\mathcal{X}$ is a *small* set for each kernel $P_{\theta,\tau}$, $\theta \in \mathcal{K}$ and $\tau \in \mathbb{T}$.

Note that if the drift function $V(x) \equiv 1$, then V-uniform ergodicity is reduced

to uniform ergodicity. To verify (A_2'), one may assume that $\mathcal{X}$ is compact, $U(x)$ is bounded in $\mathcal{X}$, and the proposal distribution $q(x, y)$ satisfies the *local positive* condition:

(Q) There exists $\delta_q > 0$ and $\epsilon_q > 0$ such that, for every $x \in \mathcal{X}$, $|x - y| \leq \delta_q \Rightarrow q(x, y) \geq \epsilon_q$.

Then the condition (A_2') holds following from Roberts and Tweedie (1996, Theorem 2.2), where it is shown that if the target distribution is bounded away from 0 and ∞ on every compact set of its support $\mathcal{X}$, then the MH chain with a proposal satisfying (Q) is irreducible and aperiodic, and the every non-empty compact set is a *small* set.

The proposals satisfying the local positive condition can also be easily designed for both continuous and discrete systems. For continuous systems, $q(x, y)$ can be set to a random walk Gaussian proposal, $y \sim N(x, \sigma^2 I_{d_x})$, where σ^2 can be calibrated to have a desired acceptance rate, e.g., $0.2 \sim 0.4$. For discrete systems, $q(x, y)$ can be set to a discrete distribution defined on a neighborhood of x. Besides the single-step MH move, the multiple-step MH move, the Gibbs sampler, and the Metropolis-within-Gibbs sampler can also be shown to satisfy condition (A_2) under appropriate conditions, see e.g. Rosenthal (1995; Lemma 7) and Liang (2009b) for the proofs. Note that to satisfy (A_2), $\mathcal{X}$ is not necessarily compact. Rosenthal (1995) gave one example for which the sample space is unbounded, yet the Markov chain is uniformly ergodic.

Conditions on gain factor and temperature sequences

(A_3') (*Conditions on* $\{\gamma_t\}$ *and* $\{\tau_t\}$)

(i) The sequence $\{\gamma_t\}$ is positive, non-increasing and satisfies the following

conditions:

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \quad \frac{\gamma_{t+1} - \gamma_t}{\gamma_t} = O(\gamma_{t+1}^{\iota}), \quad \sum_{t=1}^{\infty} \frac{\gamma_t^{(1+\iota')/2}}{\sqrt{t}} < \infty, \tag{3.12}$$

for some $\iota \in [1, 2)$ and $\iota' \in (0, 1)$.

(ii) The sequence $\{\tau_t\}$ is positive and non-increasing and satisfies the following conditions:

$$\lim_{t\to\infty} \tau_t = \tau_*, \quad \tau_t - \tau_{t+1} = o(\gamma_t), \quad \sum_{t=1}^{\infty} \gamma_t |\tau_t - \tau_{t-1}|^{\iota''} < \infty, \tag{3.13}$$

for some $\iota'' \in (0, 1)$, and

$$\sum_{t=1}^{\infty} \gamma_t |\tau_t - \tau_*| < \infty, \tag{3.14}$$

As shown in Chen (2002, p.134), the condition $\sum_{t=1}^{\infty} \frac{\gamma_t^{(1+\iota')/2}}{\sqrt{t}} < \infty$ implies

$$\sum_{t=1}^{\infty} \gamma_t^{1+\iota'} < \infty, \tag{3.15}$$

which is often assumed in studying the convergence of stochastic approximation algorithms. The condition (3.13) implies that $\{\tau_t\}$ cannot decrease too fast, and it should be set according to the gain factor sequence $\{\gamma_t\}$. The condition (3.14) also rules out the settings that $\{\tau_t\}$ converges to a point with a big gap to τ_*. For the sequences $\{\gamma_t\}$ and $\{\tau_t\}$, one can typically set

$$\gamma_t = \frac{C_1}{t^{\varsigma}}, \qquad \tau_t = \frac{C_2}{\sqrt{t}} + \tau_*, \tag{3.16}$$

for some constants $C_1 > 0$, $C_2 > 0$, and $\varsigma \in (0.5, 1]$. Then it is easy to verify that

(3.16) satisfies (A_3').

Under the above conditions, we have the following theorems concerning the convergence of $\{\theta_t\}$. Theorem 3.3.1 shows that $\{\theta_t\}$ remains in a compact subset of Θ.

Theorem 3.3.1. *Assume that $\mathbb{T}$ is compact and the conditions (A_1')-(A_3') holds. If $\tilde{\theta}_0$ used in the SAA algorithm is such that $\sup_{\tau\in\mathbb{T}} v_\tau(\tilde{\theta}_0) < \inf_{\|\theta\|=c_0,\tau\in\mathbb{T}} v_\tau(\theta)$ for some $c_0 > 0$ and $\|\tilde{\theta}_0\| < c_0$, then the number of truncations in SAA is almost surely finite; that is, $\{\theta_t\}$ remains in a compact subset of Θ almost surely.*

The proof of Theorem 3.3.1 follows the proof of Theorem 2.2.1 of Chen (2002) but with some modifications related with the observation noise, mean field function, and Lyapunov function. The details of the proof can be found in the Appendix B.

Theorem 3.3.2. *Assume the conditions of Theorem 3.3.1 hold. Then, as $t \to \infty$,*

$$d(\theta_t, \mathcal{L}_{\tau_*}) \to 0, \quad a.s.,$$

where $\mathcal{L}_{\tau_*} = \{\theta \in \Theta : h_{\tau_*}(\theta) = 0\}$ and $d(z, S) = \inf_y\{\|z - y\| : y \in S\}$.

The proof of theorem 3.3.2 is a reproduction of the proof of Theorem 5.5 of Andrieu *et al.* (2005) but with some modifications for accommodating the temperature sequence $\{\tau_t\}$. The details of the proof can be found in Appendix B. As one can see from the proofs of these two theorems, the expanding truncation weakens the condition of the Markov transition kernel for SAA. Without this technique, a more restrictive condition may need to be assumed. For example, one may assume that Θ is compact or the Doeblin condition holds uniformly over the space $\Theta \times \mathbb{T}$. The former is usually less acceptable and the latter is usually difficult to verify.

Let $\tilde{\theta}_* \in \mathcal{L}_{\tau_*}$ be the convergence point of $\{\theta_t\}$ in a run of SAA. The value of $\tilde{\theta}_*$ may be different from the true value θ_* by a constant vector. Since the probability density/mass function $f_{\theta,\tau}(x)$ is invariant to the transformation $\theta_* \leftarrow \theta_* + \boldsymbol{c}$ for a constant vector $\boldsymbol{c}$, in what follows we will denote by θ_* the point that θ_t converges to. Since $f_{\theta,\tau}(x)$ is a continuous function of θ and τ, Theorem 3.3.2 implies that as $t \to \infty$,

$$f_{\theta_t,\tau_{t+1}}(x) \to f_{\theta_*,\tau_*}(x), \quad a.s. \tag{3.17}$$

However, for stochastic optimization problems, the above convergence is not enough. Moreover, since SAA falls into the class of adaptive MCMC algorithms, it is unclear if $X_{t+1} \sim f_{\theta_t,\tau_{t+1}}(x)$ holds. To address this issue, we establish the following strong law of large numbers.

Theorem 3.3.3. *(SLLN) Assume the conditions of Theorem 3.3.1 hold. Let $x_1, \ldots, x_n$ denote a set of samples simulated by SAA in n iterations. Let g: $\mathcal{X} \to \mathbb{R}$ be a measurable function such that it is bounded and integrable with respect to $f_{\theta,\tau}(x)$. Then*

$$\frac{1}{n}\sum_{k=1}^{n} g(x_k) \to \int_{\mathcal{X}} g(x) f_{\theta_*,\tau_*}(x)dx, \quad a.s.$$

The proof of this theorem can be found in the Appendix B. Let $u_i^* = \min_{x \in E_i} U(x)$ denote the minimum of $U(x)$ on the subregion E_i. Then u_1^* corresponds to the global minimum value of $U(x)$ over $\mathcal{X}$, provided that E_1 is nonempty. Let $J(x_k)$ denote the index of the subregion that the sample x_k belongs to, i.e., $J(x_k) = i$ if $x_k \in E_i$. Then we have the following corollary.

Corollary 3.3.1. *Assume the conditions of Theorem 3 hold. Let $x_1, \ldots, x_t$ denote*

a set of samples simulated by SAA in t iterations. Then, for any $\epsilon > 0$, as $t \to \infty$,

$$\frac{1}{\sum_{k=1}^{t} I(J(x_k) = i)} \sum_{k=1}^{t} I(U(x_k) \leq u_i^* + \epsilon \;\&\; J(x_k) = i) \to \frac{\int_{\{x:U(x)\leq u_i^*+\epsilon\}\cap E_i} e^{-U(x)/\tau_*} dx}{\int_{E_i} e^{-U(x)/\tau_*} dx}, \tag{3.18}$$

almost surely for $i = 1, \ldots, m$, where $I(\cdot)$ denotes an indicator function.

3.4 Examples

Consider the function $U(\boldsymbol{x}) = -\{x_1 \sin(20x_2) + x_2 \sin(20x_1)\}^2 \cosh\{\sin(10x_1)x_1\} - \{x_1 \cos(10x_2) - x_2 \sin(10x_1)\}^2 \cosh\{\cos(20x_2)x_2\}$ that we want to minimize, where $\boldsymbol{x} = (x_1, x_2) \in [-1.1, 1.1]^2$. This example is modified from Example 5.3 of Robert and Casella (2004). Figure 3.1(a) shows that $U(\boldsymbol{x})$ has a multitude of local energy minima separated by high-energy barriers. The global minimum energy value is -8.12465, which is located at (-1.0445, -1.0084) and (1.0445, -1.0084).

To apply SAA to this example, the sample space was partitioned as in (3.2) with $m = 41$, where u_i's form an arithmetic sequence with $u_1 = -8.0$ and $u_{40} = -0.2$. The proposal distribution is a Gaussian random walk $q(\boldsymbol{x}_t, \cdot) = N_2(\boldsymbol{x}_t, 0.25^2 I_2)$. The gain factor sequence is set with $T_0 = 2000$ and $\varsigma = 1.0$, and the temperature sequence is set with $\tau_h = 0.5$, $T_0' = 200$, and $\tau_* = 0.01$. To make the problem more difficult, τ_h was set to a very small value. SAA was initialized at (1.0,1.0), which is close to a local minimum of $U(x)$, and run for 10^5 iterations. After thinning by a factor of 100, 1000 samples were collected from the run. Figure 3.1(b) shows the evolving path of the 1000 samples.

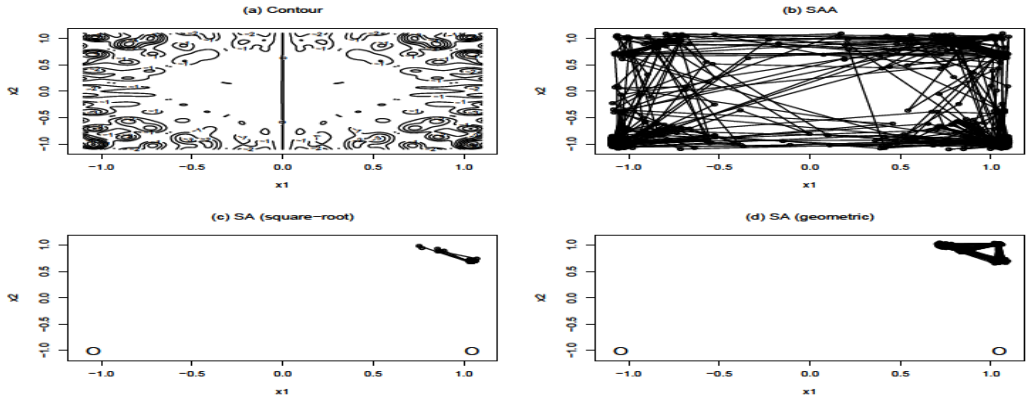
For comparison, simulated annealing was also applied to this example. Two different cooling schedules were tried, the square-root and geometric cooling schedules. The former was exactly the same as the one used in SAA. With this cooling schedule, the temperature ladder consisted of 10^5 levels and there was only one iteration

performed at each temperature level. The run started at the same point (1.0,1.0) as SAA. Figure 3.1(c) shows the evolving path of 1000 samples collected at equally spaced time points from the run. For the geometric cooling schedule, the temperature ladder was set as follows:

$$\tau_{i+1} = \varrho\tau_i, \quad i = 1, 2, \ldots, m,$$

where $\tau_1 = \tau_h$, $\varrho = 0.997244$ and the number of temperature levels $m = 1000$. This is a rather common setting for simulated annealing, especially for the value of ϱ. The resulting lowest temperature from this schedule is the same as in the square-root cooling schedule. The algorithm started at the same point (1.0,1.0) as SAA and then iterated for 100 iterations at each of the 1000 temperature levels. Figure 3.1(d) shows the evolving path of 1000 samples collected at the last iteration of each temperature level.

Figure 3.1: Simulation study to compare SAA with SA.



The comparison indicates that simulated annealing tends to get trapped into local

Table 3.1: Comparison of SAA and simulated annealing.

	Average of Minimum Energy Values						
	20000	40000	60000	80000	100000	prop	cpu
SAA	-8.1145	-8.1198	-8.1214	-8.1223	-8.1229	92.0	0.17
	(3.0×10^{-4})	(1.5×10^{-4})	(1.0×10^{-4})	(7.5×10^{-5})	(5.9×10^{-5})		
SA(sr)	-5.9227	-5.9255	-5.9265	-5.9269	-5.9271	3.5	0.14
	(1.3×10^{-2})	(1.3×10^{-2})	(1.3×10^{-2})	(1.3×10^{-2})	(1.3×10^{-2})		
SA(geo)	-6.5534	-6.5598	-6.5611	-6.5617	-6.5620	30.7	0.13
	(3.3×10^{-2})	(3.3×10^{-2})	(3.3×10^{-2})	(3.3×10^{-2})	(3.3×10^{-2})		

energy minima while SAA does not. For this example, even though the starting temperature is very low, SAA can still transverse over the energy landscape and locate the two global minima very quickly.

Later, each of the above three algorithms was run 1000 times for this example. The numerical results are summarized in Table 3.4. In column $2-6$, the average (over 1000 runs) of minimum energy values found during the first 20000, 40000, 60000, 80000, and 100000 iterations are given and the standard deviations of the averages are given in the parentheses. In column 7, we give in percentage the proportion of the runs with minimum energy values less than -8.12. The three listed methods are SAA, simulated annealing with a square-root cooling schedule and simulated annealing with a geometric cooling schedule respectively. The comparison indicates that SAA is superior to simulated annealing for this example. Even with only 20000 iterations, SAA can produce much lower energy values than simulated annealing with 10^5 iterations.

4. IMPROVING NMR PROTEIN STRUCTURE DETERMINATION USING ADVANCED MONTE CARLO METHOD

4.1 Introduction

In this chapter, We applied SAMC algorithm for model selection on the peak picking problem, which is a very hot topic in protein structure determination. In the peak picking problem, nuclear magnetic resonance is applied to a protein and an NMR spectrum gives us the intensities for chemical shifts on nitrogen and hydrogen dimension. Each peak on the NMR spectrum corresponds to a nitrogen-hydrogen bond. And identification of those bond is the first step for any other structure calculation. We model the intensities on the spectrum as the distribution density and each peak can be treated as a component of the mixture distribution. So this is easily turned into a model selection problem that asks the question: how many components are included in the mixture model, while most of the other existing methods try to treat the spectrum as a surface and use machine learning technique to solve the problem. See, for example, Corne et al., (1992) and Alipanchi et al., (2009) for successful examples among others. The main part the existing method fail to solve is how to identify true peaks and from false peaks. And numerical results show that our algorithm works better in terms of identifying true peaks while being able to exclude false peaks with high intensity.

The rest of this chapter is structured as follows: in section 2, we introduce the model and bayesian setup for NMR spectrum data. In section 3, we describe in detail the general SAMC algorithm and SAMC for peak picking. In section 4, results are given for both simulation study and real NMR data which show the benefit of our algorithm.

4.2 Model for NMR Spectrum

For simplicity, in this section, we only describe the model in the case that NMR spectrum is in 2D space. Extensions to higher dimensions follow easily from our discussion.

Suppose the NMR spectrum consists of a total of $L \times W (= n)$ grid points and we use $g(i,j)$ to denote the intensity of grid point $(i,j),\ i = 1, \cdots, L,\ j = 1, \cdots, W$. Then we model $g(i,j)$ as a mixture of bivariate Gaussian densities. We have:

$$g(i,j) = \sum_{k=1}^{|M|} a_k \phi_k(i,j|\mu_{k,1}, \mu_{k,2}, \tau_{k,1}^2, \tau_{k,2}^2) + \epsilon_{ij},\ i = 1, \cdots, L,\ j = 1, \cdots, W, \quad (4.1)$$

where $\phi(\cdot)_k$ is the kth bivariate Gaussian density function, it has $(\mu_{k,1}, \mu_{k,2})'$ as its mean and $diag(\sigma_{k,1}^2, \sigma_{k,2}^2)$ as its covariance matrix. a_k is the volume (or amplitude) of the kth Gaussian density component. And ϵ_{ij} is the error term for grid point (i,j), which is assumed to be normally distributed with mean 0 and variance σ^2. We use M to denote the model and $|M|$ to denote the size of model M, which is the number of Gaussian components or the number of peaks. So essentially, we changed the peak picking problem into a variable selection problem. We want to figure out how many components are in the model and where are the centers, i.e. what are the values of $(\mu_{k,1}, \mu_{k,2})$, $k = 1, \cdots, |M|$.

By lining up all those n data points, the above expression can be written in matrix form as follows:

$$\mathbf{Y} = \mathbf{\Phi A} + \epsilon, \quad (4.2)$$

where

$$\mathbf{Y} = \begin{pmatrix} g(1,1) \\ \vdots \\ g(1,W) \\ \vdots \\ g(L,1) \\ \vdots \\ g(L,W) \end{pmatrix}, \quad \mathbf{\Phi} = \begin{pmatrix} \phi_1(1,1) & \cdots & \phi_m(1,1) \\ \vdots & & \\ \phi_1(1,W) & \cdots & \phi_m(1,W) \\ \vdots & & \\ \phi_1(L,1) & \cdots & \phi_m(L,1) \\ \vdots & & \\ \phi_1(L,W) & \cdots & \phi_m(L,W) \end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}, \quad \text{and } \epsilon = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1W} \\ \vdots \\ \epsilon_{L1} \\ \vdots \\ \epsilon_{LW} \end{pmatrix}.$$

Y is a vector of length n, representing the spectrum intensity for each grid point. Φ is a $n \times m$ matrix that carries the information of those m Gaussian density function on each grid point, with each column corresponding to one Gaussian density component. A is a vector of length m, consists of the volume for each component. And ϵ is a vector of length n, denoting the error term.

According to Raftery et. al. (1997) and Liang et al. (2013b), we use the following

prior distribution for the unknown parameters.

$$\begin{aligned}
&\mathbf{A} \sim \mathcal{N}(\mathbf{0}, \sigma^2 V), \\
&\mu_{i,1} \sim U(0, L), \quad \mu_{i,2} \sim U(0, W), \\
&\tau_{i,1}^2 \sim IG(\alpha, \beta), \quad \tau_{i,2}^2 \sim IG(\alpha, \beta), \\
&\frac{\nu}{\sigma^2} \sim \mathcal{X}_\nu^2.
\end{aligned}$$

Where $IG(\cdot, \cdot)$ denotes Inverse-Gamma distribution. $U(\cdot, \cdot)$ denotes uniform distribution. ν, V are hyperparameters to be chosen. Here we set $V = (\Phi'\Phi)^{-1}, \nu = 1$ as in Raftery et. al., and $\alpha = \beta = 0.05$ which forms a vague priors for $\tau_{i,1}$'s and $\tau_{i,1}$'s. Because positions of peak can not exceed the region for a given spectrum, so we put a uniform prior between 0 and the length/width on $\mu_{i,1}$'s and $\mu_{i,1}$'s. The problem is to determine m and the locations $(\mu_{11}, \mu_{12}), \cdots, (\mu_{m1}, \mu_{m2})$.

Furthermore, we assume the prior distribution of $(|M| = m)$ follows a Poisson distribution with mean λ. Here, λ is another hyperparamter to be chosen, and we want to set it to be a small number because we don't want to discover many false peaks. In this paper, we set λ to be 1 for all computations and the results seems to be good.

Let $\vartheta = (\vartheta_{\mathbf{1}}, \cdots, \vartheta_{\mathbf{n}})$, $\vartheta_{\mathbf{i}} = (\mu_{\mathbf{i1}}, \mu_{\mathbf{i2}}, \mathbf{log}(\tau_{\mathbf{i1}}^{\mathbf{2}}), \mathbf{log}(\tau_{\mathbf{i2}}^{\mathbf{2}}))$, then likelihood function is:

$$f(\mathbf{Y}|\vartheta, \mathbf{A}, \sigma^{\mathbf{2}}, \mathbf{M} = \mathbf{m}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\frac{\mathbf{n}}{\mathbf{2}}} |\sigma^{\mathbf{2}}\mathbf{I_m}|^{\frac{\mathbf{n}}{\mathbf{2}}}} \exp\left\{-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{Y} - \mathbf{\Phi A})^{\mathbf{T}}(\sigma^{\mathbf{2}}\mathbf{I_n})^{-\mathbf{1}}(\mathbf{Y} - \mathbf{\Phi A})\right\}.$$

Here, I_k means a k by k indentity matrix.

The prior density functions are:

$$
\begin{aligned}
&P(\mathbf{A}) = \frac{1}{(2\pi)^{\frac{m}{2}}|\sigma^2 V|^{\frac{m}{2}}} \exp\left\{-\frac{1}{2}\mathbf{A}^T(\sigma^2 V)^{-1}\mathbf{A}\right\},\\
&P(\mu_{i1}) = \frac{1}{L} I[0, L]\ , \quad P(\mu_{i2}) = \frac{1}{U} I[0, U],\\
&P(\tau_{i1}^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_{i1}^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\tau_{i1}^2}\right)\ , \quad P(\tau_{i2}^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_{i1}^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\tau_{i2}^2}\right),\\
&P(\sigma^2) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{-\nu}{\sigma^4}\right)\left(\frac{\nu}{\sigma^2}\right)^{\frac{\nu}{2}-1}\exp\left\{-\frac{\nu}{2\sigma^2}\right\},\\
&P(|M| = m) = \frac{1}{C}\frac{\lambda^m}{m!}e^{-\lambda}, \quad m \in \{0, 1, \cdots, M_{max}\}.
\end{aligned}
$$

Here, $C = \sum_{i=1}^{M_{max}} \frac{\lambda^m}{m!}e^{-\lambda}$. And M_{max} is the maximum number of components allowed in our model. So it is equal to the total number of local maxima on a spectrum.

Integrating out $\mathbf{A}$ and σ^2 gives us:

$$
\begin{aligned}
&f(\mathbf{Y}|\vartheta, |\mathbf{M}| = \mathbf{m})\\
&= \int\int P(\mathbf{Y}|\vartheta, \mathbf{A}, \sigma^{\mathbf{2}}, |\mathbf{M}| = \mathbf{m})\mathbf{P}(\mathbf{A})\mathbf{P}(\sigma^{\mathbf{2}})\mathbf{dA}\mathbf{d}\sigma^{\mathbf{2}}\\
&= \frac{\Gamma(\frac{\nu+n}{2})(\nu)^{\nu/2}}{\pi^{n/2}\Gamma(\frac{\nu}{2})|I_n + \Phi V \Phi^T|^{1/2}} \times \{\nu + \mathbf{Y}^T(I + \Phi V \Phi^T)^{-1}\mathbf{Y}\}^{-(\nu+n)/2}.
\end{aligned}
$$

Then the posterior distribution is:

$$
\begin{aligned}
&f(\vartheta, |\mathbf{M}| = \mathbf{m}|\mathbf{Y}) \propto \mathbf{f}(\mathbf{Y}|\vartheta, |\mathbf{M}| = \mathbf{m})\mathbf{P}(\vartheta||\mathbf{M}|)\mathbf{P}(|\mathbf{M}| = \mathbf{m})\\
&\propto \frac{\lambda^m}{m!}e^{-\lambda}\frac{1}{L^m W^m}\prod_{i=1}^{m}\{\frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_{i,1}^2)^{-\alpha-1}\exp(\frac{-\beta}{\tau_{i,1}^2})\}\prod_{i=1}^{m}\{\frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_{i,2}^2)^{-\alpha-1}\exp(\frac{-\beta}{\tau_{i,2}^2})\}P(\mathbf{Y}|\vartheta, \mathbf{M}).
\end{aligned}
$$

In order to rank the peaks according to their volumes and intensities, we also need to get estimates of a_k, $i = 1, \cdots, |M|$ for each component. Note that the posterior distribution of $\mathbf{A}$ given everything else is normal with mean $(\Phi^T\Phi + V^{-1})^{-1}\Phi^T\mathbf{Y}$

and covariance matrix $\sigma^2(\Phi^T\Phi + V^{-1})^{-1}$. So we can estimate A at each step by its expectation.

4.3 Algorithm

It's not known how may peaks exist for a certain NMR spectrum, but at the position around the peaks, the intensity should be large. So according to this observation, we propose following algorithm.

For a $L \times W$ grid NMR spectrum, select N poles as the 'peak candidates'. This can be done by selecting the first N largest poles according to their intensities, or selecting points that are local maxima of if we have the results from some other methods, we can set them to be the candidates as well. In our paper, we did the simulation study by using the poles with large intensities. We use $(P_{1,1}, P_{1,2}), (P_{2,1}, P_{2,2}), \cdots, (P_{N,1}, P_{N,2})$ to record the position of each pole. We use the so-called (annealing) stochastic approximation Monte Carlo (SAMC (Liang et al., 2007), ASAMC (Liang, 2007)) model selection method (Liang, 2009a) to estimate both the number and the positions of the peaks. It is an advanced MCMC sampling method which has self-adjusting mechanism and is immune to local trap problems. Because of that, it can be applied for situation where the dimension of parameter space is high. Also, it is similar to reversible jump markov chain Monte Carlo (RJMCM (Green, 1995)) in their ability to deal with dimension mismatch.

At each step, SAMC method randomly choose to change the dimension by either add one component (Birth Move), delete one component (Death Move) or change a component (Invariant Move). We use $\mathcal{P}_I^t$ to denote the peaks already included in the model at iteration t, and $\mathcal{P}_R^t$ to denote the remaining peak candidates that are not included in the current sample. So $\mathcal{P}_I^t \bigcup \mathcal{P}_R^t = \{(P_{1,1}, P_{1,2}), (P_{2,1}, P_{2,2}), \cdots, (P_{N,1}, P_{N,2})\}$. The birth Move creates a new component by randomly select from the remaining

peak candidates $\mathcal{P}_R^t$ and generate a proposal position based on the selected peak. The death move removes one component from the existing list of peaks $\mathcal{P}_I^t$. And the invariant move randomly picks one component and propose a new component to substitute the old one. That is, it randomly select one component in $\mathcal{P}_I^t$ and propose to add a random vector to that component. So the invariant move does not change the dimension.

In this chapter, we follow the general SAMC algorithm described in Introduction and set $\gamma_t = \frac{\delta t_0}{max(t_0,t)}$, $t_0 = 500$, $\delta = 0.1$ for all of our real data example. In the following subsections, we listed the proposals and acceptance probabilities for model selection problem using SAMC. We use M^* to denote the proposed model, $M_{(t)}$ to denote the current model in iteration t, ϑ^* to denote the proposed parameter, $\vartheta_{(\mathbf{t})}$ to denote the parameter at iteration t. Also, we use $J(\vartheta)$ to denote the region index that contains ϑ.

4.3.1 Dimension Invariant Move ($M^* = M_{(t)}$)

SAMC algorithm chooses to do one of the following with equal probability: dimension invariant move, birth move and death move. Suppose at iteration t, m components are in the sample. And further assume that at iteration t+1 SAMC chooses to do a dimension invariant move. Then we randomly select one component that are one of the m components from step t. Denote the selected component as ith component and write the corresponding samples from iteration t as $\vartheta_{\mathbf{i}}^{(\mathbf{t})} = (\mu_{i,1}^{(t)}, \mu_{i,2}^{(t)}, \log(\sigma_{i,1}^2)^{(t)}, \log(\sigma_{i,2}^2)^{(t)})$.

Then the invariant move proposes $\vartheta_{\mathbf{i}}^*$ based on the current value of $\vartheta_{\mathbf{i}}$.

$$\vartheta_{i,j}^* = \vartheta_{i,j} + un \times S \times R_j \ , \quad j = 1, 2, 3, 4, \tag{4.3}$$

where $\vartheta_{\mathbf{i}}^* = (\mu_{\mathbf{i,1}}^*, \mu_{\mathbf{i,2}}^*, \mathbf{log}(\sigma_{\mathbf{i,1}}^{\mathbf{2}})^*, \mathbf{log}(\sigma_{\mathbf{i,2}}^{\mathbf{2}})^*)$. It denotes the proposed sample of com-

ponent i for iteration t+1. And un is a random sample drawn from a standard normal distribution. S is called the step size. It determines the level of variation between two iterations. As t increases, we decrease S. In our simulation study, we set $S = 0.5$ at $t = 1$ and $S = 0.01$ at $t = 200,000$. And R_j is the range of the jth parameter. Then:

$$\alpha = min\left\{1, \frac{exp\{\theta_{J(\vartheta_{(\mathbf{t})})}\}P(\mathbf{Y}|\vartheta^*, |\mathbf{M_{(t)}}|)\mathbf{P}(\vartheta^*||\mathbf{M_{(t)}}|)\mathbf{T}(\vartheta^* \to \vartheta)}{exp\{\theta_{J(\vartheta^*)}\}P(\mathbf{Y}|\vartheta, |\mathbf{M_{(t)}}|)\mathbf{P}(\vartheta||\mathbf{M_{(t)}}|)\mathbf{T}(\vartheta \to \vartheta^*)}\right\} \tag{4.4}$$

4.3.2 Birth Move ($M_{(t)} \to M^*$)

Similarly, if the move for iteration t+1 is chosen to be a birth move and the m components are included in iteration t. Then for the birth move, we randomly choose a position from the 'peak candidates' that are not being included in the components right now. If say, the ith peak candidates were chosen, i.e. $\{P_{i,1}, P_{i,2}\}$, then we propose to move in the following way:

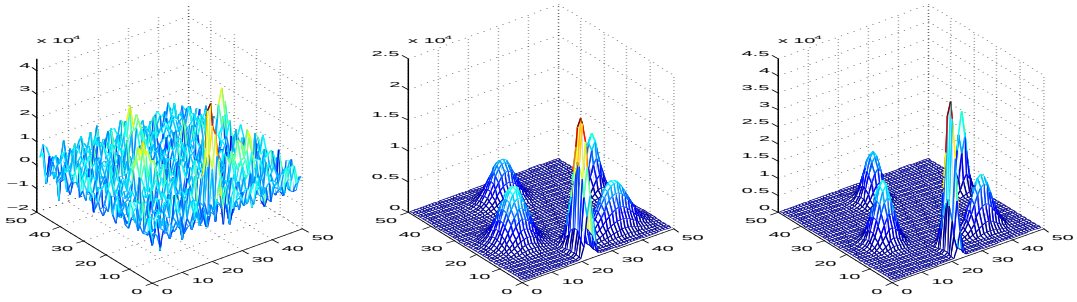
$$\begin{aligned}
\mu^*_{i,1} &= P_{i,1} + un_1 \times S \times R_1, \\
\mu^*_{i,2} &= P_{i,2} + un_2 \times S \times R_2, \\
log(\sigma^{*2}_{i,1}) &= U(log(L_3), log(U_3)), \\
log(\sigma^{*2}_{i,2}) &= U(log(L_4), log(U_4)),
\end{aligned}$$

where un_1 and un_2 are random samples drawn from standard normal distribution. The acceptance rate can be written as:

$$\alpha = min\left\{1, \frac{exp\{\theta_{J(\vartheta_{(\mathbf{t})})}\}}{exp\{\theta_{J(\vartheta^*)}\}} \frac{P(Y|\vartheta^*, |\mathbf{M}^*|)\mathbf{P}(\vartheta^*||\mathbf{M}^*|)\mathbf{P}(|\mathbf{M}^*|)\mathbf{Q}(|\mathbf{M}^*| \to |\mathbf{M_{(t)}}|)\mathbf{T}(\vartheta^* \to \vartheta)}{P(Y|\vartheta^*, |\mathbf{M_{(t)}}|)\mathbf{P}(\vartheta^*||\mathbf{M_{(t)}}|)\mathbf{P}(|\mathbf{M_{(t)}}|)\mathbf{Q}(|\mathbf{M_{(t)}}| \to |\mathbf{M}^*|)\mathbf{T}(\vartheta \to \vartheta^*)}\right\} \tag{4.5}$$

For a detailed explanation of θ_t and induction of the above formula, please refer to Liang et. al. (2007).

Figure 4.1: A simulated figure with 5 peaks.



left: a noisy image, middle: a recovered image, right: a pure image.

4.3.3 Death Move ($M_{(t)} \rightarrow M^*$)

For the death move, we randomly choose one existing component and delete it. So if there are m components at step t, then at step t+1, we randomly choose one of the m components and delete it. Then the proposed sample for step t+1 would include m-1 component. The equation for calculating acceptance rate α is the same as (4.5).

For all of the above move, generate u from a standard normal distribution and accept the proposed parameters if $\alpha > u$. Then set $\theta_t^* = \theta_t + \delta_{t+1}(\mathbf{e}_{t+1} - \pi)$, where $\mathbf{e_{t+1}} = (e_{1,t+1}, \cdots, e_{G,t+1})$, $(\pi) = (1/G, \cdots, 1/G)$, and G is the number of subregions defined in SAMC. Then at the end of the iterations, choose the model with greatest

posterior likelihood.

4.3.4 Annealing Stochastic Approximation Monte Carlo

Sometimes when the dimension is high or when the sampling space is too large that it takes an extremely long time to visit randomly through the whole energy space, then it is preferred to use a modified version of SAMC, i.e., the so-called annealing stochastic approximation Monte Carlo. The only difference is that at each iteration, we shrink the sampling space based on the energy function. So at each iteration, SAMC samples from

$$p_{\theta_t}(x) \sim \sum_{i=1}^{m} \frac{f(x)}{exp(\theta_{t,i})} I(x \in E_i). \tag{4.6}$$

While ASAMC samples from the follow distribution

$$p_{\boldsymbol{\theta}_t}(x) \sim \sum_{i=1}^{\kappa(U_{min}^{(t)}+\varkappa)} \frac{f(x)}{exp(\theta_{t,i})} I(x \in E_i), \tag{4.7}$$

where $U_{min}^{(t)}$ is the best value of $U(x)$ obtained by iteration t, $\varkappa > 0$ is a user defined parameter which determines the broadness of the sample space, and $\kappa(u)$ denotes the index of subregion based on the energy function, so if $u_{i-1} < u < u_i$, then $\kappa(u) = i$.

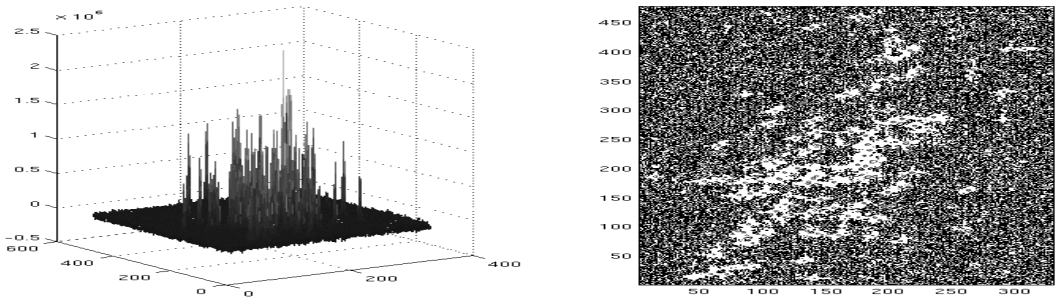
There is a trade off when choosing $\varkappa$. If we set $\varkappa$ to be a large number, then the convergence will be still, however, if we set $\varkappa$ to be too small, then ASAMC is prone to get trapped because that the shrunk region may be well separated. According the oscam's window (Madigan and Raftery, 1994), it is suffice to set $\kappa(u) = 20$, but we set $\kappa(u) = 1000$ just to be safe and avoid the algorithm being trapped.

4.4 Results

4.4.1 Simulation Study

In the simulation study, we generated an image of size 50×50 with 5 peaks. Extra noises were added to the image. The noise follows a normal distribution with mean 0 and standard deviation 4000, which makes the peaks quite hard to find using naked eyes. As shown in Figure 4.1, using the SAMC model selection method, the image was de-noised automatically and recovered the underlining structure quite well. From the left to the right are the image with noise added, the recovered image and the original pure image. In this case, although we managed to find all 5 peaks but the recovered image does not has as strong signals as the original image.

Figure 4.2: Illustration of the 2D NMR spectra data.



4.4.2 NMR Peak Picking

We utilized our method on real protein NMR data. In figure 4.2 we ploted the 3D NMR data and its contour plot. To check the accuracy of our algorithm, we applied SAMC to 6 proteins and compared its recall and precision percentages with other

Table 4.1: Performance on 6 proteins in percentage.

		PICKY			SAMC for model selection		
Protein	Length	recall	precision	average	recall	precision	average
TM1112	89	96	89	92.5	94	89	91.5
RP3384	64	94	86	90	93	91	92
ATC1776	101	78	82	80	83	84	83.5
COILIN	98	97	70	83.5	94	80	87
VRAR	72	87	93	90	93	98	95.5
HACS1	74	95	67	81	98	80	89
Average		91.2	81.2	86.2	92.5 (0.22)	87 (0.02)	89.8 (0.02)

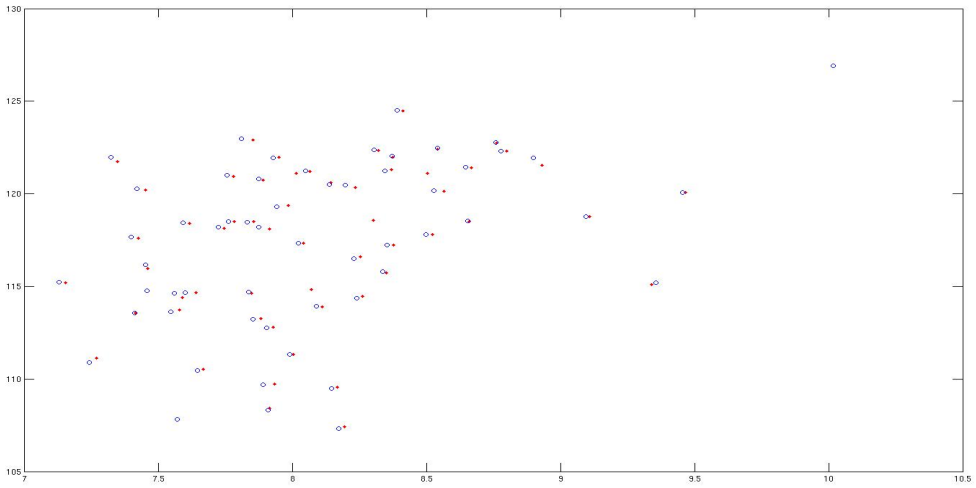
existing methods. We used 2D $^{15}N - HSQC$ spectrums for the experiment. For N dimension, we say the results is correct if the difference between our solution to the truth is less than 0.5. And for H dimension, a difference less than 0.05 is considered correct. And we say a peak that we picked is correct if both its N dimension and H dimension are with in the range when compared to the truth.

Assume the number of true peaks for a given spectrum is N_T, the number of peaks being picked is N_P and T_P of them are true peaks. Then the recall percentage is defined as N_P/N_T, which is the probability of identifying a peak when it is actually a true peak. The precision percentage is defined as N_P/T_P, which corresponds to the probability of a spot being a true peak when our algorithm said it is a peak.

Usually there is a trade-off between the recall and precision percentages. Using the same method, increasing one percentage will make the other one decrease. Because of this, we also calculated the average of precision and recall percentages and included it in the comparison. It shows in table 4.1 that our method gives more accurate solutions as compared to others.

In table 4.1, we give the comparison of using our method and PICKY (Alipanahi et. al., 2009). Column 2 gives the length for each protein, which is the true number

Figure 4.3: Result for protein VRAR.



of peaks. Column 3-5 correspond to the performance of PICKY, column 6-8 correspond to the performance of SAMC. And results show that our method gives better performance in terms of both recall accuracy and performance accuracy. We give the p-value for a paired t-test that compares our method and PICKY. And it showed that on average, our method performs better at a .05 level with such a small sample size.

In Figure 4.3, we showed the result for protein VRAR using our method. We used red dot to denote the true position of peaks and used circles to denote the peaks our algorithm found. And as shown in the picture, our method obtained a good recovery rate by producing both a good recall percentage and a good precision percentage.

5. CONCLUSION

In conclusion, in this dissertation we explored the benefit of using stochastic approximation in MCMC. We applied stochastic approximation in three different aspects.

- We applied the stochastic approximation method successfully to geostatistical data under the framework of Gaussian geostatistical model. It helps to alleviate the computational burden encountered when calculating maximum likelihood estimator. And the the same time, it keeps the nice asymptotic properties usual MLE owns.

- We applied stochastic approximation for Monte Carlo sampling method that improves the behavior of simulated annealing algorithm. And we showed theoretically that the proposed method will allow a cooling schedule much faster than the logarithmic rate, which is a requirement for simulated annealing algorithm to locate the global minimum almost surely.

- We applied stochastic approximation Monte Carlo method to the peak picking problem in protein structure determination. We tried our method on 6 proteins and calculate the precision and recall percentages. Results show that our method will perform better than the competing methods, especially about the precision, which shows that our method can identify the true peaks while including less false peaks in the list.

REFERENCES

Alipanahi, B., Gao, X., Karakov, E., Donaldson, L., and Li, M. (2009). PICKY: A Novel SVD-based NMR Spectra Peak Picking Method. *Bioinformatics*, 25, i268-i275.

Andrieu, C., Moulines, É, and Priouret, P. (2005), Stability of Stochastic Approximation Under Verifiable Conditions, *SIAM Journal of Control and Optimization*, **44**, 283-312.

Cerny, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm, *Journal of Optimization Theory and Applications*, **45**: 41â51.

Chen, H.F. (2002), *Stochastic Approximation and Its Applications*. Dordrecht: Kluwer Academic Publishers.

Corne, S., Jognson, A., and Fisher, J. (1992). An artificial neural network for classifying cross peaks in two dimemsional NMR spectra. *Journal of Magnetic Resonance*, 100, 256-266.

Du, J., Zhang, H. and Mandrekar, V.S. (2009), Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators, *Annals of Statistics*, **37**, 3330-3361.

Folland, G.B. (1990), Remainder Estimates in Taylor's Theorem, *The American Mathematical Monthly*, **97**, 233-235.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *J. Am. Statist. Ass.*, **102**, 321-331.

Furrer, R. (2006), *KriSp*: An R Package for Covariance Tapered Kriging of Large Datasets Using Sparse Matrix Techniques,

Furrer, R. Genton, M.G. and Nychka, D. (2006), Covariance Tapering for Interpolation of Large Spatial Datasets, *Journal of Computational and Graphical Statistics*, **15**, 502-523.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711-732.

Hall, P. and Heyde, C.C. (1980), *Martingale Limit Theory and Its Application*. New York: Academic Press.

Johns, C.J., Nychka, D., Kittel, T.G.F. and Daly, C. (2003), Infilling Sparse Records of Spatial Fields, *Journal of the American Statistical Association*, **98**, 796-806.

Kaufman, C.G., Schervish, M.J. and Nychka, D.W. (2008), Covariance Tapering for Likelihood-based Estimation in Large Spatial Data Sets, *Journal of the American Statistical Association*, **103**, 1545-1555.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983), Optimization by Simulated Annealing, *Science*, **220**, 671-680.

Lahiri, S.N. (1996), On Inconsistency of Estimators Based on Spatial Data under Infill Asymptotics, *Sankhyā: The Indian Journal of Statistics*, **58**, 403-417.

Liang, F. (2007), Annealing Stochastic Approximation Monte Carlo for Neural Network Training, *Machine Learning*, **68**, 201-233.

Liang, F. (2009a), On the Use of Stochastic Approximation Monte Carlo for Monte Carlo Integration, *Statistics & Probability Letters*, **79**, 581-587.

Liang, F. (2009b), Improving SAMC Using Smoothing Methods: Theory and Applications to Bayesian Model Selection Problems, *Annals of Statistics*, **37**, 2626-2654.

Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. (2013a). A Resampling-based Stochastic Approximation Method for Analysis of Large Geostatistical Data. *J. Amer. Statist. Assoc.*, **108**, 325-339.

Liang, F., Liu, C., and Carroll, R.J. (2007), Stochastic Approximation in Monte Carlo Computation, *Journal of the American Statistical Association*, **102**, 305-320.

Liang, F., Song, Q., and Yu, K. (2013b) Bayesian Subset Modeling for High Dimensional Generalized Linear Models. *J. Amer. Statist. Assoc.*, to appear.

Madigan, D.M. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *J. Amer. Statist. Assoc.* **89**, 1335-1346.

Matsuda, Y. and Yajima, Y. (2009), Fourier Analysis of Irregularly Spaced Data on $\mathbb{R}^d$, *Journal of the Royal Statistical Society, Series B*, **71**, 191-217.

Nummelin, E. (1984), *General Irreducible Markov Chains and Nonnegative Operators.* Cambridge: Cambridge University Press.

Pelletier, M. (1998), Weak Convergence Rates for Stochastic Approximation with Application to Multiple Targets and Simulated Annealing, *Annals of Applied Probability*, **8**, 10-44.

Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997) Bayesian Model Averaging for Linear Regression Models. *J. Amer. Statist. Assoc.*, 92, 437, 179-191

Ribeiro Jr., P.J. and Diggle, P.J.(2001), *geoR: A package for geostatistical analysis*. R-NEWS Vol.1, No 2. ISSN 1609-3631.

Robbins, H. and Monro, S. (1951), A Stochastic Approximation Method, *Annals of Mathiematical Statistics*, **22**,400-407.

Robert, C.P. and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd edition). Springer.

Roberts, G.O., and Tweedie, R.L. (1996), Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms, *Biometrika*, **83**, 95-110.

Rosenthal, J.S. (1995), Minorization Conditions and Convergence Rate for Markov Chain Monte Carlo, *Journal of the American Statistical Association*, **90**, 558-566.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC.

Rue, H. and Tjelmeland, H. (2002), Fitting Gaussian Markov Random Fields to Gaussian Field, *Scan. J. Statist.*, **29**, 31-49.

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer.

Stein, M.L. (2004), Equivalence of Gaussian Measures for Some Nonstationary Random Fields, *Journal of Statistical Planning and Inference*, **123**, 1-11.

Stein, M.L. Chi, Z. and Welty, L.J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc. B*, **66**, 275-296.

Tadić, V. (1997), On the Convergence of Stochastic Iterative Algorithms and Their Applications to Machine Learning, a short version of this paper was published in *Proceedings of the 36th Conference on Decision and Control*, San Diego, pp.2281-2286.

Zhang, H. (2004), Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics, *Journal of the American Statistical Association*, **99**, 250-261.

APPENDIX A

PROOFS FOR RSA

In appendix A, we give the supplimentary proofs for theorems ralated to resampling-based stochastic approximation method.

A.1 Conditions for Convergence of Algorithm 2.3.1

Theoretical properties of Algorithm 2.3.1 are studied under the following conditions:

(A_1) The function $h : \Theta \mapsto \mathbb{R}^d$ is continuous, and there exists a continuously differentiable function $v : \Theta \mapsto [0, \infty)$ such that:

(i) There exists $C_0 > 0$ such that

$$\mathcal{L} = \{\theta \in \Theta, \langle \nabla v(\theta), h(\theta) \rangle = 0\} \subset \{\theta \in \Theta, v(\theta) < C_0\}, \tag{A.1}$$

where $\langle x, y \rangle$ denotes the Euclidean inner product.

(ii) There exists $C_1 \in (C_0, \infty]$ such that $\mathcal{V}_{C_1}$ is a compact set, where $\mathcal{V}_C = \{\theta \in \Theta, v(\theta) \leq C\}$.

(iii) For any $\theta \in \Theta \setminus \mathcal{L}$, $\langle \nabla v(\theta), h(\theta) \rangle < 0$.

(iv) The closure of $v(\mathcal{L})$ has an empty interior.

(A_2) There exists a function $V : \mathcal{X} \to [1, \infty)$ such that for any compact subset $\mathcal{K} \subset \Theta$, there exists a constant c such that

(i) $\sup_{\theta \in \mathcal{K}} \|H(\theta, \cdot)\|_V \leq c$;

(ii) $\sup_{(\theta, \theta') \in \mathcal{K} \times \mathcal{K}} \ \|H(\theta, \cdot) - H(\theta', \cdot)\|_V \leq c\|\theta - \theta'\|$.

(A_3) The mean field function $h(\theta)$ is measurable and locally bounded. There exist a stable matrix F (i.e., all eigenvalues of F are with negative real parts), $\rho > 0$, and a constant c such that, for any $\theta_* \in \mathcal{L}$ (defined in (A.1)),

$$\|h(\theta) - F(\theta - \theta_*)\| \leq c\|\theta - \theta_*\|^2, \quad \forall \theta \in \{\theta : \|\theta - \theta_*\| \leq \rho\}.$$

(A_4) The sequences $\{a_t\}$ and $\{b_t\}$, which are defined to be $a(t)$ and $b(t)$ as functions of t and are exchangeable with $a(t)$ and $b(t)$, respectively, are non-increasing, positive, and satisfy the conditions:

$$\begin{aligned} \lim_{t\to\infty} a_t = 0, \quad \sum_{t=0}^{\infty} a_t = \infty, \quad \frac{a_{t+1} - a_t}{a_t} = O(a_{t+1}^{\tau_1}), \\ \lim_{t\to\infty} b_t = 0, \quad \sum_{t=1}^{\infty} \{a_t^{\tau_2} + (a_t/b_t)^{\tau_3} + a_t b_t^{\tau_4}\} < \infty, \end{aligned} \tag{A.2}$$

for some values of $\tau_1 \in (1, 2]$, $\tau_2 \in (1, 2]$, $\tau_3 \in [2, \infty)$ and $\tau_4 \in (0, 1]$.

Moreover, we assume that the function $a(t)$ is differentiable, with either (i) or (ii) holding:

(i) $a(t)$ varies regularly with exponent $(-\beta)$, $\frac{1}{2} < \beta < 1$; that is, for any $z > 0$, $a(zt)/a(t) \to z^{-\beta}$ as $t \to \infty$.

(ii) For $t \geq 1$, $a(t) = t_0/t$ with $t_0 > -1/(2\lambda_F)$, where λ_F denotes the largest real part of the eigenvalue of the matrix F (defined in condition A_3) with $\lambda_F < 0$.

Condition (A_4) can be applied to the usual gains $a_t = t_0/t^\beta$ and $b_t = t_0'/t^{\beta'}$ by choosing $\beta \in (\frac{1}{2}, 1]$, $\beta' \in (\frac{1}{2}, \beta - \frac{1}{\tau_3})$, $\tau_3 \in (2, \infty)$ and $\tau_4 = 1$. Following Pelletier

(1998), we deduce that

$$\left(\frac{a_t}{a_{t+1}}\right)^{1/2} = 1 + \frac{\beta}{2t} + o(\frac{1}{t}). \tag{A.3}$$

In terms of a_t, (A.3) can be rewritten as

$$\left(\frac{a_t}{a_{t+1}}\right)^{1/2} = 1 + \zeta a_t + o(a_t), \tag{A.4}$$

where $\zeta = 0$ for the case (i) of (A_4) and $\zeta = \frac{\beta}{2t_0}$ for the case (ii) of (A_4). Clearly, the matrix is $F + \zeta I$ is still stable.

A.2 Proof of Lemma 2.3.5.

In Algorithm 2.3.1, the sample X_t is generated at each iteration in an exact manner. Since the exact sampling procedure can be viewed as a special case of the Markovian sampling procedure, the convergence theorem established by Andrieu *et al.* (2005) for the varying truncation stochastic approximation MCMC algorithm can be applied to Algorithm 2.3.1. If we let P_θ denote the transition kernel corresponding to the exact sampling procedure, then it is irreducible, aperiodic, and admits $g_\theta(x)$ as the invariant distribution. In addition, it admits the whole sample space as a small set and satisfies the drift condition. The remaining part of the proof follows from from Theorem 5.4, Theorem 5.5 and Proposition 6.1 of Andrieu *et al.* (2005).

A.3 Proof of Theorem 2.3.3.

By Lemma 2.3.5, it suffices to show that Algorithm 2.2.1 satisfies the conditions (A_1), (A_2) and (A_4).

(A_1) As implied by (2.5), we have

$$h(\theta) = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \frac{\partial \log f_\theta(\boldsymbol{z}_i|\boldsymbol{s}_i)}{\partial \theta},$$

which is continuous in θ. Define

$$v(\theta) = \mathrm{KL}(f_\theta, \tilde{g}) = -\int\int \log\left(\frac{f_\theta(\boldsymbol{z}|\boldsymbol{s})}{\tilde{g}(\boldsymbol{z}|\boldsymbol{s})}\right) \tilde{g}(\boldsymbol{z}, \boldsymbol{s}) d\boldsymbol{z}\boldsymbol{s}, \tag{A.5}$$

which is continuously differentiable with respect to θ. By Jensen's inequality, we have $v(\theta) \geq 0$. In addition,

$$\langle \nabla v(\theta), h(\theta) \rangle = -\langle h(\theta), h(\theta) \rangle,$$

which implies that (A_1)-(iii) holds.

For any $\theta \in \mathcal{L}$, it corresponds to a local minimizer of $v(\theta)$. In addition, $v(\theta)$ is continuous. Hence, there exists a constant $C_0 \geq \sup_{\theta \in \mathcal{K}_0} v(\theta)$ such that (A_1)-(i) holds and $\mathcal{K}_0 \subset \mathcal{V}_{C_0}$. Note that at the true values of the parameters, $v(\theta) = 0$, so the set $\{\theta \in \Theta, v(\theta) < C_0\}$ always contains the true parameters for any $C_0 > 0$.

Since Θ is compact, we can set $C_1 = \sup_{\theta \in \Theta} v(\theta)$. Thus, ($A_1$)-(ii) is satisfied.

Since, for any $\theta \in \mathcal{L}$, it corresponds to a local minimizer of $v(\theta)$. It is obvious that the set $v(\mathcal{L})$ is nowhere dense. This verifies (A_1)-(iv).

(A_2) Set $V(x) \equiv 1$ for all $x \in \mathcal{X}$, where $\mathcal{X}$ denotes the sample space of X_t and it contains only $\binom{n}{m}$ elements. Since $H(\theta, x)$ is continuous in θ, for each $x_i \in \mathcal{X}$

and any compact set $\mathcal{K} \subset \Theta$, we let

$$H(x_i) = \max\{\sup_{\theta \in \mathcal{K}} \|H(\theta, x_i)\|, \sup_{(\theta,\theta') \in \mathcal{K} \times \mathcal{K}} \|\theta - \theta'\|^{-1} \|H(\theta, x_i) - H(\theta', x_i)\|\}.$$

Hence, (A_2) is satisfied by setting $c = \max_{x_i \in \mathcal{X}} H(x_i)$.

(A_4) This condition can be satisfied by choosing appropriate sequences $\{a_t\}$ and $\{b_t\}$.

Since $V(x) \equiv 1$, the condition $\sup_{x \in \mathcal{X}_0} V(x) < \infty$ is trivially satisfied. This means that the algorithm will converge for any starting sample $X_0 \in \mathcal{X}$.

A.4 Proof of Theorem 2.3.4.

By Lemma 2.3.6, it suffices to verify the conditions (A_1)–(A_4). Since the model has been assumed to be identifiable, the condition (A_3) is satisfied by choosing F to be the Hessian matrix of $v(\theta)$. The conditions (A_1), (A_2) and (A_4) can be verified as in Theorem 2.3.3.

APPENDIX B

PROOFS FOR SAA

In appendix B, we give the supplimentary proof for theorems ralated to simulated stochastic approximation annealing algorithm

B.1 Proof of Theorem 3.1

Lemma B.1.1. *Assume that $\mathbb{T}$ is compact and the condition (A_2) holds. Then the following results hold for the SAA algorithm:*

(B_1) *For any $\theta \in \Theta$ and $\tau \in \mathbb{T}$, the Markov kernel $P_{\theta,\tau}$ has a single stationary distribution $f_{\theta,\tau}$. In addition, $H : \Theta \times \mathcal{X} \to \Theta$ is measurable for all $\theta \in \Theta$ and $\tau \in \mathbb{T}$, $\int_{\mathcal{X}} \|H_\tau(\theta, x)\| f_{\theta,\tau}(x)dx < \infty$.*

(B_2) *For any $\theta \in \Theta$ and $\tau \in \mathbb{T}$, the Poisson equation $u_{\theta,\tau}(X) - P_{\theta,\tau}u_{\theta,\tau}(X) = H_\tau(\theta, X) - h_\tau(\theta)$ has a solution $u_{\theta,\tau}(X)$, where $P_{\theta,\tau}u_{\theta,\tau}(X) = \int_{\mathcal{X}} u_{\theta,\tau}(y)P_{\theta,\tau}(X, y)dy$. For any constant $\eta \in (0, 1)$ and any compact subset $\mathcal{K} \subset \Theta$, the following results hold:*

(i) $\sup_{\theta\in\mathcal{K},\tau\in\mathbb{T}} (\|u_{\theta,\tau}(\cdot)\| + \|P_{\theta,\tau}u_{\theta,\tau}(\cdot)\|) < \infty,$

(ii) $\sup_{\theta,\theta'\in\mathcal{K},\tau\in\mathbb{T}} \|\theta - \theta'\|^{-\eta} \{\|u_{\theta,\tau}(\cdot) - u_{\theta',\tau}(\cdot)\| + \|P_{\theta,\tau}u_{\theta,\tau}(\cdot) - P_{\theta',\tau}u_{\theta',\tau}(\cdot)\|\} < \infty.$

(iii) $\sup_{\theta\in\mathcal{K},\tau,\tau'\in\mathbb{T}} \|\tau - \tau'\|^{-\eta} \|P_{\theta,\tau}u_{\theta,\tau}(\cdot) - P_{\theta,\tau'}u_{\theta,\tau'}(\cdot)\| < \infty.$

Proof. Since $\mathbb{T}$ is compact and (A_2) holds, then it is easy to verify that the following condition holds for the SAA algorithm by choosing $\boldsymbol{C} = \mathcal{X}$, $V(x) \equiv 1$, $0 < \lambda < 1$, $b > 1$ and $\kappa > 1$:

There exist a function $V : \mathcal{X} \to [1, \infty)$*, a constant* $\alpha \geq 2$*, a set* $\boldsymbol{C} \subset \mathcal{X}$*,* $0 < \lambda < 1$*,* $b > 0$ *and* $\kappa > 0$ *such that for any compact subset* $\mathcal{K} \subset \Theta$*,*

$$\begin{aligned}\sup_{\theta \in \mathcal{K}, \tau \in \mathbb{T}} P_{\theta,\tau}^{l} V^{\alpha}(x) \leq & \lambda V^{\alpha}(x) + bI(x \in \boldsymbol{C}), \quad \forall X \in \mathcal{X}, \\ \sup_{\theta \in \mathcal{K}, \tau \in \mathbb{T}} P_{\theta,\tau} V^{\alpha}(x) \leq & \kappa V^{\alpha}(x), \qquad \forall x \in \mathcal{X},\end{aligned}$$

where $I(\cdot)$ *is the indicator function and* $P_{\theta,\tau} V^{\alpha}(X) = \int_{\mathcal{X}} P_{\theta,\tau}(X, y) V^{\alpha}(y) dy$*.*

As in Liang *et al.* (2007), we can verify that the following condition holds for the SAA algorithm: *There exists a constant* $c > 0$ *such that for all* $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$*,*

$$\begin{aligned}\|P_{\theta,\tau} g - P_{\theta',\tau} g\| &\leq c\|g\|\|\theta - \theta'|, \quad \forall g \in \mathcal{G}, \\ \|P_{\theta,\tau} g - P_{\theta,\tau'} g\| &\leq c\|g\||\tau - \tau'|, \quad \forall g \in \mathcal{G},\end{aligned}$$

where $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R}^d, \|g\| < \infty\}$.

In addition, it is easy to see that SAA satisfies the following conditions:

$$\begin{aligned}&\sup_{\theta \in \mathcal{K}, \tau \in \mathbb{T}} \|H_\tau(\theta, \cdot)\| \leq c, \\ &\sup_{(\theta,\theta') \in \mathcal{K} \times \mathcal{K}, \tau \in \mathbb{T}} \|\theta - \theta'\|^{-1} \|H_\tau(\theta, \cdot) - H_\tau(\theta', \cdot)\| \leq c. \\ &\sup_{\theta \in \mathcal{K}, (\tau,\tau') \in \mathbb{T} \times \mathbb{T}} |\tau - \tau'|^{-1} \|H_\tau(\theta, \cdot) - H_{\tau'}(\theta, \cdot)\| \leq c,\end{aligned} \tag{B.1}$$

where c denotes a constant. For the first equation, we can set $c = 2$. The other two equations hold because $H_\tau(\theta, X)$ is independent of τ and θ for a given value of X. Then, following from Proposition 6.1 of Andrieu *et al.* (2005), we have (B_1), (B_2)-(i) and (B_2)-(ii) hold. (B_2)-(iii) can be proved following the same line of the proof of (B_2)-(ii). □

Lemma B.1.2. *(Noise decomposition) For the SAA algorithm, there exist* $\mathbb{R}^{d_\theta}$*-valued*

random processes $\{\epsilon_t\}_{t\geq 0}$, $\{\epsilon_t^{'}\}_{t\geq 0}$ *and* $\{\epsilon_t^{''}\}_{t\geq 0}$ *such that*

$$\gamma_{t+1}\xi_{t+1} = \epsilon_{t+1} + \epsilon_{t+1}^{'} + \epsilon_{t+1}^{''} - \epsilon_t^{''} + \epsilon_t^{'''}, \quad t \geq 0, \tag{B.2}$$

where $\xi_{t+1} = H_{\tau_{t+1}}(\theta_t, X_{t+1}) - h_{\tau_{t+1}}(\theta_t)$ *is the observation noise.*

Proof. Apply Poisson equation to ξ_t and let $\epsilon_0 = \epsilon_0^{'} = 0$,

$$\begin{aligned}
\epsilon_{t+1} &= \gamma_{t+1}\big[u_{\theta_t,\tau_{t+1}}(x_{t+1}) - P_{\theta_t,\tau_{t+1}}u_{\theta_t,\tau_{t+1}}(x_t)\big],\\
\epsilon_{t+1}^{'} &= \gamma_{t+1}\big[P_{\theta_{t+1},\tau_{t+1}}u_{\theta_{t+1},\tau_{t+1}}(x_{t+1}) - P_{\theta_t,\tau_{t+1}}u_{\theta_t,\tau_{t+1}}(x_{t+1})\big]\\
&\qquad + (\gamma_{t+2} - \gamma_{t+1})P_{\theta_{t+1},\tau_{t+1}}u_{\theta_{t+1},\tau_{t+1}}(x_{t+1}),\\
\epsilon_{t+1}^{''} &= -\gamma_{t+2}P_{\theta_{t+1},\tau_{t+1}}u_{\theta_{t+1},\tau_{t+1}}(x_{t+1}),\\
\epsilon_t^{'''} &= \gamma_{t+1}\left(P_{\theta_t,\tau_{t+1}}u_{\theta_t,\tau_{t+1}}(x_t) - P_{\theta_t,\tau_t}u_{\theta_t,\tau_t}(x_t)\right).
\end{aligned}$$

It is easy to verify that (B.2) is satisfied. □

Proof of Theorem 3.1

Proof. The proof is completed in four steps by considering convergent subsequences of a sample path.

- **Step 1.** We show that there are constants $M > 0$, $K > 0$ such that for any $k \in [0, K]$ there exists a constant $t_k > 0$ such that for any $t > t_k$

$$\left\|\sum_{i=n_t}^{m}\gamma_{i+1}H_{\tau_{i+1}}(\theta_i, X_{i+1})\right\| \leq M, \quad \forall m : n_t \leq m \leq m(n_t, k), \tag{B.3}$$

  if $\{\theta_{n_t}\}$ is a convergent subsequence of $\{\theta_t\}$: as $t \to \infty$, $n_t \to \infty$ and $\theta_{n_t} \to \bar{\theta}$, where $m(t,k)$ is defined by $m(t,k) = \max\{m : \sum_{i=t}^{m}\gamma_i \leq k\}$, and M is independent of k and t.

Since $\|H_\tau(\theta, x)\| \leq 2$ holds for the SAA algorithm, we have

$$\left\| \sum_{i=n_t}^{m} \gamma_{i+1} H_{\tau_{i+1}}(\theta_i, X_{i+1}) \right\| \leq 2k \leq 2K, \quad \forall m : n_t \leq m \leq m(n_t, k),$$

which implies (B.3) holds for $M = 2K$.

- **Step 2.** We show that for all t large enough, there exists a constant c_1 such that

$$\|\theta_{m+1} - \theta_{n_t}\| \leq c_1 k, \quad \forall m : n_t \leq m \leq m(n_t, k), \quad \forall k \in [0, K], \tag{B.4}$$

if K is small enough.

If the number of truncations in SAA is finite, then for large enough t there is no truncation and thus

$$\|\theta_{m+1} - \theta_{n_t}\| \leq \| \sum_{i=n_t}^{m} \gamma_{i+1} H_{\tau_{i+1}}(\theta_i, X_{i+1}) \| \leq 2k,$$

because $\|H_\tau(\theta, x)\|$ is bounded by 2. Hence, it suffices to prove (B.4) for the case $\sigma_t \to \infty$ as $t \to \infty$.

It follows from (B.3) that for any $k \in [0, K]$,

$$\left\| \theta_{n_t} + \sum_{i=n_t}^{m} \gamma_{i+1} H_{\tau_{i+1}}(\theta_i, X_{i+1}) \right\| \leq M + \|\bar{\theta}\| + 1 \leq M_{\sigma_t}, \quad \forall m : n_t \leq m \leq m(n_t, k),$$

if t is large enough. That is, there are no truncations for $n_t \leq m \leq m(n_t, k)$, and thus

$$\theta_{m+1} = \theta_m + \gamma_{m+1} H_{\tau_{m+1}}(\theta_m, X_{m+1}), \quad \|\theta_{m+1}\| \leq M + 1 + \|\bar{\theta}\|. \tag{B.5}$$

Since $H_\tau(\theta, X)$ is bounded over $\Theta \times \mathbb{T}$, there exists a constant c_1 such that

$$\|\theta_{m+1} - \theta_{n_t}\| = \left\| \sum_{i=n_t}^{m} \gamma_{i+1} H_{\tau_{i+1}}(\theta_i, X_{i+1}) \right\| \leq c_1 k \tag{B.6}$$

for large enough t and and small enough K. This concludes Step 2.

- **Step 3.** We show the assertion: For any interval $[\delta_1, \delta_2]$ with $\delta_1 < \delta_2$ and $d([\delta_1, \delta_2], v(\mathcal{L})) > 0$, the sequence $\{v_{\tau_{t+1}}(\theta_t)\}$ cannot cross $[\delta_1, \delta_2]$ infinitely many times with $\{\|\theta_{n_t}\|\}$ bounded, where "crossing $[\delta_1, \delta_2]$ by $v_{\tau_{n_t+1}}(\theta_{n_t}), \ldots, v_{\tau_{m_t+1}}(\theta_{m_t})$" means that $v_{\tau_{n_t+1}}(\theta_{n_t}) \leq \delta_1$, $v_{\tau_{m_t+1}}(\theta_{m_t}) \geq \delta_2$, and $\delta_1 < v_{\tau_{i+1}}(\theta_i) < \delta_2$ for $n_t < i < m_t$.

  Assume the converse: there are infinitely many crossings by the sequence $\{v_{\tau_{t+1}}(\theta_t)\}$ and $\{\|\theta_{n_t}\|\}$ is bounded. By the boundedness of $\{\|\theta_{n_t}\|\}$, without loss of generality, we may assume $\theta_{n_t} \to \bar{\theta}$ as $t \to \infty$. Therefore, by the continuity of $v_\tau(\theta)$,

$$v_{\tau_{n_t+1}}(\theta_{n_t}) \to \delta_1 = v_{\tau_*}(\bar{\theta}) \quad \text{and} \quad d(\bar{\theta}, \mathcal{L}) \stackrel{\Delta}{=} \delta > 0, \tag{B.7}$$

  as $t \to \infty$. While, from (B.4), one can see that if k is sufficiently small, then

$$d(\theta_m, \mathcal{L}) \geq \frac{\delta}{2}, \quad \forall m : n_t \leq m \leq m(n_t, k), \tag{B.8}$$

  for sufficiently large t.

By (B.5) and (B.4), for large t we have

$$
\begin{aligned}
& v_{\tau_{m(n_t,k)+2}}(\theta_{m(n_t,k)+1}) - v_{\tau_{n_t+1}}(\theta_{n_t}) \\
= & \sum_{i=n_t}^{m(n_t,k)} [v_{\tau_{i+2}}(\theta_{i+1}) - v_{\tau_{i+1}}(\theta_{i+1}) + v_{\tau_{i+1}}(\theta_{i+1}) - v_{\tau_{i+1}}(\theta_i)] \\
= & \sum_{i=n_t}^{m(n_t,k)} \gamma_{i+1} H_{\tau_{i+1}}(\theta_i)^T \nabla_\theta v_{\tau_{i+1}}(\theta_i) + \sum_{i=n_t}^{m(n_t,k)} |\tau_{i+2} - \tau_{i+1}| \nabla_\tau v_{\tau_{i+1}}(\theta_{i+1}) + o(k) \\
= & \sum_{i=n_t}^{m(n_t,k)} \gamma_{i+1} h_{\tau_{i+1}}(\theta_i)^T \nabla_\theta v_{\tau_{i+1}}(\theta_i) + \sum_{i=n_t}^{m(n_t,k)} \gamma_{i+1} \xi_{i+1}^T \nabla_\theta v_{\tau_{i+1}}(\theta_i) \\
& + \sum_{i=n_t}^{m(n_t,k)} |\tau_{i+2} - \tau_{i+1}| \nabla_\tau v_{\tau_{i+1}}(\theta_{i+1}) + o(k).
\end{aligned}
$$

It follows from Step 2 that for all t large enough, there exists a constant c_1 such that

$$
\|\theta_m\| \le c_1 K + \|\bar{\theta}\| + 1 \triangleq \mathcal{M}, \quad \forall m : n_t \le m \le m(n_t,k)+1, \ \forall k \in [0,K], \tag{B.9}
$$

if K is small enough.

Consider the decomposition of $\gamma_{t+1}\xi_{t+1}$ given in Lemma B.1.2. Since

$$
E(u_{\theta_t,\tau_{t+1}}(x_{t+1})|\mathcal{F}_t) = P_{\theta_t,\tau_{t+1}} u_{\theta_t,\tau_{t+1}}(x_t),
$$

$\{\epsilon_{t+1} I(\|\theta_t\| \le \mathcal{M})\}$ forms a martingale difference sequence, where $\{\mathcal{F}_t\}_{t\ge 0}$ is a family of σ-algebras of $\mathcal{F}$ satisfying $\sigma\{\epsilon_t, \epsilon_t', \epsilon_t'', \epsilon_t'''\} \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, $t \ge 0$ and $\sigma\{\theta_0\} \subseteq \mathcal{F}_0$. Further, it follows from (B_2)-(i) that $\sum_{t=0}^{\infty} \|\epsilon_{t+1}\|^2 I(\|\theta_t\| \le \mathcal{M}) < \infty$. Then, by the martingale convergence theorem (Hall and Heyde, 1980; Theorem 2.15), $\sum_{t=0}^{\infty} \epsilon_{t+1} I(\|\theta_t\| \le \mathcal{M})$ converges almost surely. This, together

with (B.9), implies that for any k,

$$\lim_{t\to\infty} \| \sum_{i=n_t}^{m(n_t,k)} \epsilon_{i+1} \| = 0. \tag{B.10}$$

For the term ϵ'_{t+1} in (B.2), it follows from (B.9) and (B_2)-(ii) that

$$\|\epsilon^{'}_{m+1}\| \le c_2\gamma_{m+1}\|\theta_{m+1}-\theta_m\|^{\eta} + c_3\gamma_{m+1}^{1+\iota} \le c_4\gamma_{m+1}^{1+\eta}, \quad \forall m : n_t \le m \le m(n_t,k),$$

for some constants $c_2 > 0$, $c_3 > 0$, $c_4 > 0$, and $1 > \eta > \max\{\iota', \iota^{''}\}$ (ι, ι' and $\iota^{''}$ are defined in (A_3)). Therefore,

$$\lim_{t\to\infty} \| \sum_{i=n_t}^{m(n_t,k)} \epsilon'_{i+1} \| = o(k), \tag{B.11}$$

For the term $\epsilon^{''}_{t+1}$ in (B.2), it follows from (B.9) and (B_2)-(i) that for any k,

$$\lim_{t\to\infty} \| \sum_{i=n_t}^{m(n_t,k)} (\epsilon^{''}_{i+1} - \epsilon^{''}_{i}) \| = \lim_{t\to\infty} \|\epsilon^{''}_{m(n_t,k)+1} - \epsilon^{''}_{n_t}\| = 0. \tag{B.12}$$

For the term $\epsilon^{'''}_t$, it follows from (B.9) and (B_2)-(iii) that

$$\|\epsilon^{'''}_m\| \le c_5\gamma_{m+1}(\tau_m - \tau_{m+1})^{\eta}, \quad \forall m : n_t \le m \le m(n_t,k),$$

for some constants $c_5 > 0$ and $0 < \eta < 1$. Therefore, by (A_3),

$$\lim_{t\to\infty} \| \sum_{i=n_t}^{m(n_t,k)} \epsilon^{'''}_{i} \| = o(k). \tag{B.13}$$

It follows from (B.10)–(B.13) and the boundedness of $\nabla_\theta v_\tau(\theta)$ that

$$\lim_{t\to\infty}\left\|\sum_{i=n_t}^{m(n_t,k)}\gamma_{i+1}\xi_{i+1}^T\nabla_\theta v_{\tau_{i+1}}(\theta_i)\right\| = o(k). \tag{B.14}$$

Following from (A_3)-(ii) and the boundedness of $\nabla_\tau v_\tau(\theta)$ over the compact set $\mathbb{T}$, we have

$$\lim_{t\to\infty}\left\|\sum_{i=n_t}^{m(n_t,k)}|\tau_i-\tau_{i+1}|\nabla_\tau v_{\tau_{i+1}}(\theta_{i+1})\right\| = o(k). \tag{B.15}$$

Then, by (B.8) and condition (A_1), there exist $\alpha>0$ and k such that

$$v_{\tau_{m(n_t,k)+2}}(\theta_{m(n_t,k)+1}) - v_{\tau_{n_t+1}}(\theta_{n_t}) \le -\alpha k,$$

for sufficiently large t. Further, by (B.7), we derive

$$\limsup_{t\to\infty} v_{\tau_{m(n_t,k)+2}}(\theta_{m(n_t,k)+1}) \le \delta_1 - \alpha k. \tag{B.16}$$

However, by (B.4) we have

$$\lim_{k\to 0}\max_{n_t\le m\le m(n_t,k)}|v_{\tau_{m+2}}(\theta_{m+1}) - v_{\tau_{n_t+1}}(\theta_{n_t})| = 0,$$

which implies $v_{\tau_{m(n_t,k)+2}}(\theta_{m(n_t,k)+1}) \in [\delta_1,\delta_2)$. This contradicts (B.16).

- **Step 4.** We now show that the number of truncations is bounded.

  Since $v(\mathcal{L})$ is nowhere dense, a nonempty interval $[\delta_1,\delta_2]$ exists such that $[\delta_1,\delta_2]\subset(\sup_{\tau\in\mathbb{T}} v_\tau(\tilde\theta_0),\ \inf_{\|\theta\|=c_0,\tau\in\mathbb{T}} v_\tau(\theta))$ and $d([\delta_1,\delta_2],v(\mathcal{L}))>0$. If $\sigma_t\to\infty$, then θ_t, starting from $\tilde\theta_0$, will cross the sphere $\{\theta:\|\theta\|=c_0\}$ infinitely many times. Consequently, $v_{\tau_{t+1}}(\theta_t)$ will cross $[\delta_1,\delta_2]$ infinitely often with $\{\theta_{n_t}\}$

bounded. In step 3, we have shown this process is impossible. Therefore, starting from some t_0, the SAA algorithm will have no truncations and $\{\theta_t\}$ is bounded.

□

B.2 Proof of Theorem 3.2

Using Lemma B.1.1, we can prove the following Lemma, which is an extension of Theorem 4.1 and Lemma 2.2 of Tadić (1997).

Lemma B.2.1. *Assume that $\mathbb{T}$ is compact and the conditions (A_1)–(A_3) hold. Then the following results hold:*

(C_1) *The series $\sum_{t=1}^{\infty}\|\epsilon'_{t+1}\|$, $\sum_{t=1}^{\infty}\|\epsilon''_{t+1}\|^2$, $\sum_{t=1}^{\infty}\|\epsilon_{t+1}\|^2$ and $\sum_{t=1}^{\infty}\|\epsilon'''_t\|$ all converge a.s. and*

$$E(\epsilon_{t+1}|\mathcal{F}_t)=0, \quad a.s., \quad n\geq 0, \tag{B.17}$$

where $\{\mathcal{F}_t\}_{t\geq 0}$ is a family of σ-algebras of $\mathcal{F}$ satisfying $\sigma\{\epsilon_t,\epsilon'_t,\epsilon''_t,\epsilon'''_t\}\subseteq\mathcal{F}_t\subseteq\mathcal{F}_{t+1}$, $t\geq 0$ and $\sigma\{\theta_0\}\subseteq\mathcal{F}_0$.

(C_2) *Let $R_t=R'_t+R''_t$, $t\geq 0$, where $R'_t=\gamma_{t+1}\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\xi_{t+1}$, and*

$$R''_t=\int_0^1\left[\nabla_\theta v_{\tau_{t+1}}(\theta_t+s(\theta_{t+1}-\theta_t))-\nabla_\theta v_{\tau_{t+1}}(\theta_t)\right]^T(\theta_{t+1}-\theta_t)ds.$$

Then $\sum_{t=1}^{\infty}\gamma_{t+1}\xi_{t+1}$ and $\sum_{t=1}^{\infty}R_t$ converge a.s..

Proof. By Theorem 3.1, the number of truncations in SAA is finite. Hence, for simplicity, this lemma can be proved by assuming that the number of truncation is 0 and $\{\theta_t\}$ remains in a compact set.

(C_1) Since

$$E(u_{\theta_t,\tau_{t+1}}(x_{t+1})|\mathcal{F}_t)=P_{\theta_t,\tau_{t+1}}u_{\theta_t,\tau_{t+1}}(x_t),$$

which concludes (B.17). The conditions (B_2) and (A_3) imply that there exist constants c_1, c_2, c_3, c_4, c_5, $c_6 \in \mathbb{R}^+$ and $\eta \in (0,1)$ such that

$$\begin{aligned}
&\|\epsilon_{t+1}\|^2 \leq 2c_1\gamma_{t+1}^2, \\
&\|\epsilon_{t+1}^{'}\| \leq c_2\gamma_{t+1}\|\theta_{t+1}-\theta_t\|^{\eta} + c_3\gamma_{t+1}^{1+\iota} \leq c_4\gamma_{t+1}^{1+\eta}, \\
&\|\epsilon_{t+1}^{''}\|^2 \leq c_5\gamma_{t+1}^2, \\
&\|\epsilon_{t}^{'''}\| \leq c_6\gamma_{t+1}(\tau_t - \tau_{t+1})^{\eta},
\end{aligned} \tag{B.18}$$

where ι is defined in (A_3). Setting $1 > \eta \geq \max\{\iota', \iota''\}$ (ι' and ι'' are defined in A_3), it follows from (A_3) that the series $\sum_{t=0}^{\infty}\|\epsilon_{t+1}\|^2$, $\sum_{t=0}^{\infty}\|\epsilon_{t+1}^{'}\|$, $\sum_{t=0}^{\infty}\|\epsilon_{t}^{''}\|^2$ and $\sum_{t=0}^{\infty}\|\epsilon_{t}^{'''}\|$ all converge almost surely.

(C_2) Let $M = \sup_t\{\|h_{\tau_{t+1}}(\theta_t)\|, \|\nabla_\theta v_{\tau_{t+1}}(\theta_t)\|\}$. It follows from (A_1) that $M < \infty$. Since $\sigma\{\theta_t\} \subset \mathcal{F}_t$, the condition (C_1) implies that $E(\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}|\mathcal{F}_t) = 0$. In addition, we have

$$\sum_{t=0}^{\infty} E\big(|\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}|\big)^2 \leq M^2\sum_{t=0}^{\infty} E\big(\|\epsilon_{t+1}\|^2\big) < \infty.$$

It follows from the martingale convergence theorem (Hall and Heyde, 1980; Theorem 2.15) that both $\sum_{t=0}^{\infty}\epsilon_{t+1}$ and $\sum_{t=0}^{\infty}\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}$ converge almost surely. Since

$$\begin{aligned}
&\sum_{t=0}^{\infty}|\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}^{'}| \leq M\sum_{t=1}^{\infty}\|\epsilon_t^{'}\|, \\
&\sum_{t=1}^{\infty}\gamma_t^2\|\xi_t\|^2 \leq 5\sum_{t=1}^{\infty}\|\epsilon_t\|^2 + 5\sum_{t=1}^{\infty}\|\epsilon_t^{'}\|^2 + 10\sum_{t=0}^{\infty}\|\epsilon_t^{''}\|^2 + 5\sum_{t=0}^{\infty}\|\epsilon_t^{'''}\|^2
\end{aligned}$$

it follows from (C_1) that $\sum_{t=0}^{\infty}|\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}^{'}|$, $\sum_{t=1}^{\infty}|\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\epsilon_t^{'''}|$ and $\sum_{t=1}^{\infty}\gamma_t^2\|\xi_t\|^2$

all converge.

Following from (A_1), there exists a constant c such that

$$\|R_t^{''}\| \leq c\|\theta_{t+1}-\theta_t\|^2 = c\|\gamma_{t+1}h_{\tau_{t+1}}(\theta_t)+\gamma_{t+1}\xi_{t+1}\|^2 \leq 2c\big(M^2\gamma_{t+1}^2+\gamma_{t+1}^2\|\xi_{t+1}\|^2\big),$$

which implies

$$\sum_{t=1}^{\infty}|R_t^{''}| \leq 2cM^2\sum_{t=1}^{\infty}\gamma_{t+1}^2 + 2c\sum_{t=1}^{\infty}\gamma_{t+1}^2\|\xi_{t+1}\|^2 < \infty,$$

i.e., $\sum_{t=1}^{\infty} R_t^{''}$ converges. In addition, following from (A_1), there exists a constant c' such that

$$\begin{aligned}
&\left|\big(\nabla_\theta v_{\tau_{t+1}}(\theta_t) - \nabla_\theta v_{\tau_t}(\theta_{t-1})\big)^T \epsilon_t^{''}\right| \\
=&\left|\big(\nabla_\theta v_{\tau_{t+1}}(\theta_t) - \nabla_\theta v_{\tau_t}(\theta_t) + \nabla_\theta v_{\tau_t}(\theta_t) - \nabla_\theta v_{\tau_t}(\theta_{t-1})\big)^T \epsilon_t^{''}\right| \\
\leq&\left[c\|\theta_t-\theta_{t-1}\| + c'|\tau_t-\tau_{t+1}|\right]\|\epsilon_t^{''}\| \\
\leq&\left[c\gamma_t\|h_{\tau_t}(\theta_{t-1})\| + c\gamma_t\|\xi_t\| + c'|\tau_t-\tau_{t+1}|\right]\|\epsilon_t^{''}\|.
\end{aligned}$$

Consequently, by Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sum_{t=1}^{\infty}\left|\big(\nabla_\theta v_{\tau_{t+1}}(\theta_t) - \nabla_\theta v_{\tau_t}(\theta_{t-1})\big)^T \epsilon_t^{''}\right| \\
\leq&\left(3c^2M^2\sum_{t=1}^{\infty}\gamma_t^2 + 3c^2\sum_{t=1}^{\infty}\gamma_t^2\|\xi_t\|^2 + 3c'^2\sum_{t=1}^{\infty}|\tau_t-\tau_{t+1}|^2\right)^{1/2}\left(\sum_{t=1}^{\infty}\|\epsilon_t^{''}\|^2\right)^{1/2} \\
<&\infty.
\end{aligned}$$

where the last inequality follows from the condition $\tau_t - \tau_{t+1} = o(\gamma_t)$ given in

(A_3)-(ii). Since

$$\begin{aligned}
\sum_{t=0}^{n}\gamma_{t+1}\xi_{t+1} &= \sum_{t=0}^{n}\epsilon_{t+1}+\sum_{t=0}^{n}\epsilon_{t+1}^{'}+\epsilon_{n}^{''}-\epsilon_{1}^{''}+\sum_{t=1}^{n}\epsilon_{t}^{'''},\\
\sum_{t=0}^{n}R_{t+1}^{'} =& \sum_{t=0}^{n}\nabla_{\theta}^{T}v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}+\sum_{t=0}^{n}\nabla_{\theta}^{T}v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}^{'}+\nabla_{\theta}^{T}v_{\tau_{t+1}}(\theta_t)\epsilon_{t+1}^{''}\\
&-\sum_{t=1}^{n}\left(\nabla_{\theta}v_{\tau_{t+1}}(\theta_t)-\nabla_{\theta}v_{\tau_t}(\theta_{t-1})\right)^{T}\epsilon_{t}^{''}-\nabla_{\theta}^{T}v_{\tau_1}(\theta_0)\epsilon_{0}^{''}\\
&+\sum_{t=0}^{n}\nabla_{\theta}^{T}v_{\tau_{t+1}}(\theta_t)\epsilon_{t}^{'''},
\end{aligned}$$

and $\epsilon_n^{''}$ converges to zero by (C_1), it is obvious that $\sum_{t=1}^{\infty}\gamma_t\xi_t$ and $\sum_{t=1}^{\infty}R_t = \sum_{t=1}^{\infty}R_t^{'}+\sum_{t=1}^{\infty}R_t^{''}$ converge almost surely.

The proof for Lemma B.2.1 is completed. □

Proof of Theorem 3.2

Proof. Let $M' = \sup_t\{\|h_{\tau_{t+1}}(\theta_t)\|, \|v_{\tau_{t+1}}(\theta_t)\|\}$ and $\mathcal{V}_\varepsilon = \{\theta : d(\theta, \mathcal{L}_{\tau_*}) \le \varepsilon\}$. It follows from (A_1) that $M' < \infty$. It follows from Taylor's expansion formula (Folland, 1990) that

$$\begin{aligned}
v_{\tau_{t+1}}(\theta_{t+1}) &= v_{\tau_{t+1}}(\theta_t)+\gamma_{t+1}\nabla_{\theta}^{T}v_{\tau_{t+1}}(\theta_t)[h_{\tau_{t+1}}(\theta_t)+\xi_{t+1}]\\
&\quad+\int_0^1[\nabla_{\theta}v_{\tau_{t+1}}(\theta_t+s(\theta_{t+1}-\theta_t))-\nabla_{\theta}v_{\tau_{t+1}}(\theta_t)]^{T}(\theta_{t+1}-\theta_t)ds\\
&= v_{\tau_{t+1}}(\theta_t)+\gamma_{t+1}\dot{v}_{\tau_{t+1}}(\theta_t)+R_{t+1}^{'}+R_{t+1}^{''},
\end{aligned}$$

where $\dot{v}_{\tau_{t+1}}(\theta_t) = \nabla_\theta v_{\tau_{t+1}}(\theta_t)h_{\tau_{t+1}}(\theta_t)$, $R_{t+1}^{'} = \gamma_{t+1}\nabla_\theta^T v_{\tau_{t+1}}(\theta_t)\xi_{t+1}$ and $R_{t+1}^{''} = \int_0^1[\nabla_\theta v_{\tau_{t+1}}(\theta_t+s(\theta_{t+1}-\theta_t))-\nabla_\theta v_{\tau_{t+1}}(\theta_t)]^T(\theta_{t+1}-\theta_t)ds$ are as defined in (C_2). There-

fore,

$$\begin{aligned}\sum_{i=0}^{t}\gamma_i\dot{v}_{\tau_{i+1}}(\theta_{i+1}) =& v_{\tau_{t+1}}(\theta_{t+1}) - v_{\tau_1}(\theta_0) + \sum_{i=1}^{t}\left(v_{\tau_i}(\theta_i) - v_{\tau_{i+1}}(\theta_i)\right) - \sum_{i=0}^{t} R_{i+1} \\ \geq& -2M' - L(\tau_1 - \tau_{t+1}) - \sum_{i=0}^{t} R_{i+1},\end{aligned}$$

where $R_{i+1} = R'_{i+1} + R''_{i+1}$, and L is the Lipschitz constant of $v_\tau(\theta)$ (with respect to τ), i.e.,

$$\sup_{\theta\in\mathcal{K},(\tau,\tau')\in\mathbb{T}\times\mathbb{T}} |v_\tau(\theta) - v_{\tau'}(\theta)| \leq L|\tau - \tau'|.$$

Since $\sum_{i=0}^{t} R_{i+1}$ converges (owing to Lemma B.2.1), $\sum_{i=0}^{\infty}\gamma_{i+1}\dot{v}_{\tau_{i+1}}(\theta_{i+1})$ also converges.

Furthermore, for $t \geq 0$, we have

$$v_{\tau_t}(\theta_t) = v_{\tau_1}(\theta_0) + \sum_{i=0}^{t-1}\gamma_{i+1}\dot{v}_{\tau_{i+1}}(\theta_i) + \sum_{i=1}^{t-1}\left(v_{\tau_{i+1}}(\theta_i) - v_{\tau_i}(\theta_i)\right) + \sum_{i=0}^{t-1} R_{i+1}.$$

Suppose that $\underline{\lim}_{t\to\infty} d(\theta_t, \mathcal{L}_{\tau_*}) > 0$. Then there exists $\varepsilon > 0$ and n_0 such that $d(\theta_t, \mathcal{L}_{\tau_*}) \geq \varepsilon$, $t \geq n_0$. Since $\sum_{t=1}^{\infty}\gamma_t = \infty$ and $p = \sup\{\dot{v}_{\tau_*}(\theta) : \theta \in \mathcal{V}_\varepsilon^c\} < 0$, there exists a constant K' such that

$$\begin{aligned}\sum_{t=n_0}^{\infty}\gamma_{t+1}\dot{v}_{\tau_{t+1}}(\theta_t) =& \sum_{t=n_0}^{\infty}\gamma_{t+1}\dot{v}_{\tau_*}(\theta_t) + \sum_{t=n_0}^{\infty}\gamma_{t+1}\left(\dot{v}_{\tau_{t+1}}(\theta_t) - \dot{v}_{\tau_*}(\theta_t)\right) \\ \leq& p\sum_{t=1}^{\infty}\gamma_{t+1} + c\sum_{t=1}^{\infty}\gamma_{t+1}|\tau_{t+1} - \tau_*| = -\infty,\end{aligned}$$

where the last equality holds due to the condition (A_3)-(ii). This contradicts with the convergence of $\sum_{i=0}^{\infty}\gamma_{i+1}\dot{v}_{\tau_{i+1}}(\theta_{i+1})$. Hence, $\underline{\lim}_{t\to\infty} d(\theta_t, \mathcal{L}_{\tau_*}) = 0$.

Suppose that $\overline{\lim}_{t\to\infty} d(\theta_t, \mathcal{L}_{\tau_*}) > 0$. Then, there exists a constant $\varepsilon > 0$ such that $\overline{\lim}_{t\to\infty} d(\theta_t, \mathcal{L}_{\tau_*}) \geq 2\varepsilon$. Let $t_0 = \inf\{t \geq 0 : d(\theta_t, \mathcal{L}_{\tau_*}) \geq 2\varepsilon\}$, while $t'_k = \inf\{t \geq$

$t_k : d(\theta_t, \mathcal{L}_{\tau_*}) \leq \varepsilon\}$ and $t_{k+1} = \inf\{t \geq t'_k : d(\theta_t, \mathcal{L}_{\tau_*}) \geq 2\varepsilon\}$, $k \geq 0$. Obviously, $t_k < t_{k'} < t_{k+1}$, $k \geq 0$, and

$$d(\theta_{t_k}, \mathcal{L}_{\tau_*}) \geq 2\varepsilon,\ d(\theta_{t'_k}, \mathcal{L}_{\tau_*}) \leq \varepsilon,\ \text{and } d(\theta_t, \mathcal{L}_{\tau_*}) \geq \varepsilon,\ t_k \leq t < t'_k,\ k \geq 0.$$

By the definition of $p = \sup\{\dot{v}_{\tau_*}(\theta) : \theta \in \mathcal{V}_\varepsilon^c\}$, we have

$$p\sum_{k=0}^{\infty}\sum_{i=t_k}^{t'_k-1}\gamma_{i+1} + \sum_{k=0}^{\infty}\sum_{i=t_k}^{t'_k-1}\gamma_{i+1}\left(\dot{v}_{\tau_{i+1}}(\theta_i) - \dot{v}_{\tau_*}(\theta_i)\right)$$
$$\geq \sum_{k=0}^{\infty}\sum_{i=t_k}^{t'_k-1}\gamma_{i+1}\dot{v}_{\tau_{i+1}}(\theta_i) \geq \sum_{t=0}^{\infty}\gamma_{t+1}\dot{v}_{\tau_{t+1}}(\theta_t) > -\infty,$$

where the second to the last inequality follows from the condition (A_1) that $\dot{v}_\tau(\theta) \leq 0$ for all $\theta \in \Theta$ and $\tau \in \mathbb{T}$, and the last inequality holds because $\sum_{t=0}^{\infty}\gamma_{t+1}\dot{v}_{\tau_{t+1}}(\theta_t)$ converges.

Since, by (A_1) and (A_3)-(iii), there exists a constant K' such that

$$\left|\sum_{k=0}^{\infty}\sum_{i=t_k}^{t'_k-1}\gamma_{i+1}\left(\dot{v}_{\tau_{i+1}}(\theta_i) - \dot{v}_{\tau_*}(\theta_i)\right)\right| \leq K'\sum_{t=1}^{\infty}\gamma_{t+1}|\tau_{t+1} - \tau_*| < \infty,$$

we conclude that $\sum_{k=0}^{\infty}\sum_{i=t_k}^{t'_k-1}\gamma_{i+1} < \infty$, and consequently, $\lim_{k\to\infty}\sum_{i=t_k}^{t'_k-1}\gamma_{i+1} = 0$. Since $\sum_{t=1}^{\infty}\gamma_t\xi_t$ converges (owing to Lemma B.2.1), as $k \to \infty$,

$$\varepsilon \leq \|\theta_{t'_k} - \theta_{t_k}\| \leq M'\sum_{i=t_k}^{t'_k-1}\gamma_{i+1} + \left\|\sum_{i=t_k}^{t'_k-1}\gamma_{i+1}\xi_{i+1}\right\| \longrightarrow 0,$$

recalling the definition of M'. This contradicts with our assumption $\varepsilon > 0$. Hence, $\overline{\lim}_{t\to\infty} d(\theta_t, \mathcal{L}_{\tau_*}) > 0$ does not hold. Therefore, $\lim_{t\to\infty} d(\theta_t, \mathcal{L}_{\tau_*}) = 0$ almost surely.

□

B.3 Proof of Theorem 3.3

Proof. By Poisson equation,

$$u_{\theta,\tau}(x) - P_{\theta,\tau}u_{\theta,\tau}(x) = g(x) - f_{\theta,\tau}(g),$$

where $P_{\theta,\tau}$ denotes the joint Markov transition kernel as defined previously, x denotes a sample generated by $P_{\theta,\tau}$, $P_{\theta,\tau}u_{\theta,\tau}(x) = \int_{\mathcal{X}} u_{\theta,\tau}(x')P_{\theta,\tau}(x,x')dx'$, $f_{\theta,\tau}(g) = \int g(x)f_{\theta,\tau}(x)dx$, and $f_{\theta,\tau}(x)$ denotes the stationary distribution of $P_{\theta,\tau}$. Consider the decomposition

$$\gamma_{t+1}[g(X_{t+1}) - f_{\theta_t,\tau_{t+1}}(g)] = \epsilon_{t+1} + \epsilon'_{t+1} + \epsilon''_{t+1} - \epsilon''_t + \epsilon'''_t,$$

where ϵ_{t+1}, ϵ'_{t+1}, ϵ''_{t+1} and ϵ'''_t are defined as in Lemma B.1.2, and $\epsilon''_0 = 0$. As a result, we have

$$\sum_{t=0}^{n} \gamma_{t+1}[g(X_{t+1}) - f_{\theta_t,\tau_{t+1}}(g)] = \sum_{t=0}^{n} \epsilon_{t+1} + \sum_{t=0}^{n} \epsilon'_{t+1} + \sum_{t=0}^{n} \epsilon'''_{t+1} + \epsilon''_{n+1} - \epsilon''_0.$$

Since g is bounded, we can show, as in Lemma B.2.1, that $\{\epsilon_{t+1}\}$ forms a martingale difference sequence and $\sum_{t=0}^{\infty} \epsilon_{t+1}$ converges almost surely. Similarly, it follows from (B.18) that

$$\| \sum_{t=0}^{\infty} \epsilon'_{t+1} + \sum_{t=0}^{\infty} \epsilon'''_{t+1} \| < \infty.$$

Since $\|\epsilon''_{n+1} - \epsilon''_0\|$ is also upper bounded, we have

$$\sum_{t=0}^{n} \gamma_{t+1}[g(X_{t+1}) - f_{\theta_t,\tau_{t+1}}(g)] < \infty, \quad a.s. \tag{B.19}$$

Applying Kronecker's Lemma to equation (B.19), we obtain

$$\frac{1}{n}\sum_{t=1}^{n}\left[g(X_{t+1}) - \int_{\mathcal{X}} g(x) f_{\theta_t,\tau_{t+1}}(x)dx\right] \to 0, \quad a.s. \tag{B.20}$$

By Theorem 3.1, which implies that X_{t+1} will converge in distribution to a random variable distributed according to $f_{\theta_*,\tau_*}(x)$, and the boundedness of $g(x)$, we have

$$\int_{\mathcal{X}} g(x) f_{\theta_t,\tau_{t+1}}(x)dx \to \int_{\mathcal{X}} g(x) f_{\theta_*,\tau_*}(x)dx, \quad \text{as } t \to \infty. \tag{B.21}$$

Then, the proof is concluded by combining equations (B.20) and (B.21). □