Chapter 4

Development of a Bayesian mobility spectrum

4.1 Introduction

The mobility spectrum problem is approached by the Bayesian formalism with maximum entropy principle. It has been applied to many physics problems; for example, the determination of the dynamic properties of a solvated electron from equilibrium path integral Monte Carlo data [Gallicchio and Berne (1996)], the analytic continuation of imaginary-time quantum Monte Carlo data [Jarrell and Gubernatis (1996)], the determination of depth profile from RBS data [Toussaint et al. (1999)], the determination of crystallite-size distribution from x-ray line profiles [Armstrong et al. (2001)], and the QCD spectral analysis [Nakahara (2001)].

This chapter describes details of developing the Bayesian and maximum entropy method of mobility spectrum. Firstly, the formulation of the problem and the variable notation are defined. In Section 4.2, the information theory and maximum entropy principle are briefly introduced. Next, Bayes' theorem is introduced to the data analysis in Section 4.3. In Section 4.4, Bayes' theorem with maximum entropy principle is applied, and the problem is related to probability functions. The calculation procedure and the demonstration are also included. Section 4.5 describes the error analysis of mobility calculation. The usage numerical algorithm, Markov chain Monte Carlo (MCMC), are separately introduced in the last section.

As same as the maximum entropy method, a set of partial conductivity $\{s_i\}$ and magnetoconductivity tensor components $\sigma_{xx}(B_j)$ and $\sigma_{xy}(B_j)$ at M different magnetic fields are normalized with the conductivity at zero magnetic field

$$p_i = \frac{s_i}{\sigma_0} \tag{4.1}$$

and

$$\sigma_j = \begin{cases} \frac{\sigma_{xx}(B_j)}{\sigma_0} & \text{for } 1 \leq j \leq M \\ \\ \frac{\sigma_{xy}(B_j)}{\sigma_0} & \text{for } M+1 \leq j \leq 2M \end{cases}, \tag{4.2}$$

where $\{p_i\}$ is a set of probabilities with i =1, 2, ..., N (a number of mobility points) and j =1, 2, ..., $2M$ (a number of data points). Eqs. (3.5) and (3.6) can be written in a matrix form in terms of parameters defined above as

$$\sum_{i=1}^{N} K_{ji} \cdot p_i = \sigma_j, \tag{4.3}$$

where

$$K_{ji} = \begin{cases} \frac{1}{1+\mu_i^2 B_j^2} & \text{for } 1 \leq j \leq M \\ \\ \frac{\mu_i B_j}{1+\mu_i^2 B_j^2} & \text{for } M+1 \leq j \leq 2M \end{cases}. \tag{4.4}$$

It is noted that, in this formulation, hole mobility is positive and electron mobility is negative.

4.2 Information Entropy and Maximum entropy principle

Suppose that an event x with probability p has been observed. In an information theory, the information can be measured by defining the information in terms of the probability p. The information is defined to satisfy the following properties. They are:

1) Information is a non-negative quantity : $I(p) \geq 0$.

2) If an event has a probability of 1, there is no information to get from the observation : $I(p) = 0$.

3) The information of two independent events is the sum of each information : $I(p_1 \times p_2) = I(p_1) + I(p_2)$.

4) The information measured is a monotonic and continuous function of the probability. From these required properties, the information function is derived as [see Carter (2002)]

$$I(p) = \log\left(\frac{1}{p}\right). \tag{4.5}$$

Suppose that the variable X provides the set of events $x_1, x_2, ..., x_N$ with probabilities $p_1, p_2, ..., p_N$, respectively. Each event is assumed to be independent from the others. The information of this multiple-outcome event is found to be a summation of weighted average among the information of each independent event

$$I(\{p_i\}) = \sum_{i=1}^{N} p_i \times \log\left(\frac{1}{p_i}\right), \tag{4.6}$$

which is called an information entropy introduced by Shannon (1948). The entropy is the quantity used to measure the degree of ignorance (uncertainty) of any probability distribution. Generally, the entropy H is defined for any continuous probability distribution $P(x)$ over the range of interest,

$$\begin{aligned} H(P) &= \int P(x) \log\left(\frac{1}{P(x)}\right) dx \\ &= -\int P(x) \log P(x)\, dx. \end{aligned} \tag{4.7}$$

This function is concave, and the global maximum occurs when the probability distribution is uniform.

The information entropy is used as a tool in the inverse problem, namely Maximum Entropy principle [Jaynes (1957) and Agmon et al. (1979)]. The solution

of a problem is considered probability distribution. From the maximum entropy view point, the uniform distribution indicates that all events happen equally and this is preferable if there is no measurement. After data are taken, the distribution must be modified to yield the result that fit to data. In general, the fitting method can not produce a unique distribution due to limitation of data points. As a result, there are many possible distributions that agree with given data. In principle, the distribution which is maximally noncommittal to unmeasured data is the most feasible one that is highest in entropy. Jaynes (1957) proposed that the distribution with maximum entropy is the least bias on the given data or any constraints, and it is the most likely solution among all of the possible distributions.

In summary, the method using the maximum entropy principle yields a unique solution distribution by producing a good fitting to the data.

4.3 Bayesian theory

Bayesian theory plays an important role on modern statistics. It explains how the existing belief should be modified in the light of new data. On the other hand, it provides an alternative approach for inferring parameters from their observed data via a statistical model, and the result is summarized by the probability distribution. Interestingly, it allows adding the measurement noise into the model and the uncertainty in the result is computed explicitly based on the statistical data analysis.

Let X and Y are sequential events. Bayesian formulation is derived from probabilistic rules, the sum rule

$$P(X) + P(\overline{X}) = 1 \tag{4.8}$$

and the product rule

$$P(X,Y) = P(X|Y)\,P(Y) = P(Y|X)\,P(X)\,. \tag{4.9}$$

The sum rule states that the probability that X is true ($P(X)$) and the probability that X is false ($P(\overline{X})$) sum to unity. The product rule states that the probability that both X and Y are true ($P(X,Y)$) is equal to the multiplication between the probability that X is true given that Y is true ($P(X|Y)$) and the probability that Y is true ($P(Y)$), and vice versa. These rules led to Bayes' theorem (see Sivia (1996) and Lee (1997) for more details)

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y)}. \tag{4.10}$$

The probability $P(X)$, which is called the prior probability of X, represents the existing knowledge of X before the observed data Y is available. The probability $P(Y|X)$ is called the likelihood function which relates to probability model of data measurement. The probability $P(X|Y)$ is called the posterior probability which yields the modified knowledge of X in the light of data Y. The probability $P(Y)$ is called the evidence and is considered normalization constant.

However, Bayes' theorem provides only the relation between the probability terms. It does not specify how to define the model of each probability distribution. For the likelihood function, there are many choices of statistical models available; for example, a normal distribution, Poisson distribution, $t-$distribution and Lorentzian distribution. To assign the appropriate model of likelihood function and prior distribution, the characteristic of the concerned problem must be considered. The interested likelihood function and prior distribution are presented at the end of this section. Moreover, the determination of the posterior probability distribution is also outside the Bayes' theorem. There are many methods, including both analytical and

numerical methods, to calculate the posterior probability or the required expectation value. In this thesis, the Markov chain Monte Carlo (MCMC) sampling method is chosen.

To demonstrate the usage of Bayes' theorem, a simple coin-tossing experiment is performed [Sivia (1996)]. The question is "Is this a fair coin?". Let h denotes the head bias-weighting, ranging from 0 to 1. $h = 0.5$ represents a fair coin. To make the conclusion about the bias of this coin, Bayes' theorem gives

$$\begin{aligned} P(h|\{d\}) &\propto P(\{d\}|h)\,P(h) \\ &\propto \prod_{i=1}^{N} P(d_i|h)\,P(h), \end{aligned} \tag{4.11}$$

where d is the result of any tossing (head or tail) and N represents the number of tossings. Each tossing is independent then the likelihood can be written as the product of the likelihood of all tossings. The prior probability $P(h)$ represents the prior knowledge about the coin. If the information about bias is not presented, the probability distribution of head-bias is assumed to be uniform. The likelihood function $P(d|h)$ relates to the measurement of data that there are two independent outcomes, head or tail, and it is given by binomial distribution for R heads in N tosses

$$P(d|h) \propto h^R (1-h)^{N-R}. \tag{4.12}$$

The product of prior probability and likelihood function yields the posterior that represents the inference. For large number of tosses, the results are shown in Fig. 4.1. Fig. 4.1 (a) shows posterior distribution for none of data that indicates the lack of information state. If the number of tosses increases, the width of posterior distribution becomes narrow indicating that the knowledge about bias-weighting of this coin is clearer. Finally, we observe that the coin is head bias-weighting about 0.25 in Fig. 4.1 (f). In addition, the width of the posterior distribution can be used

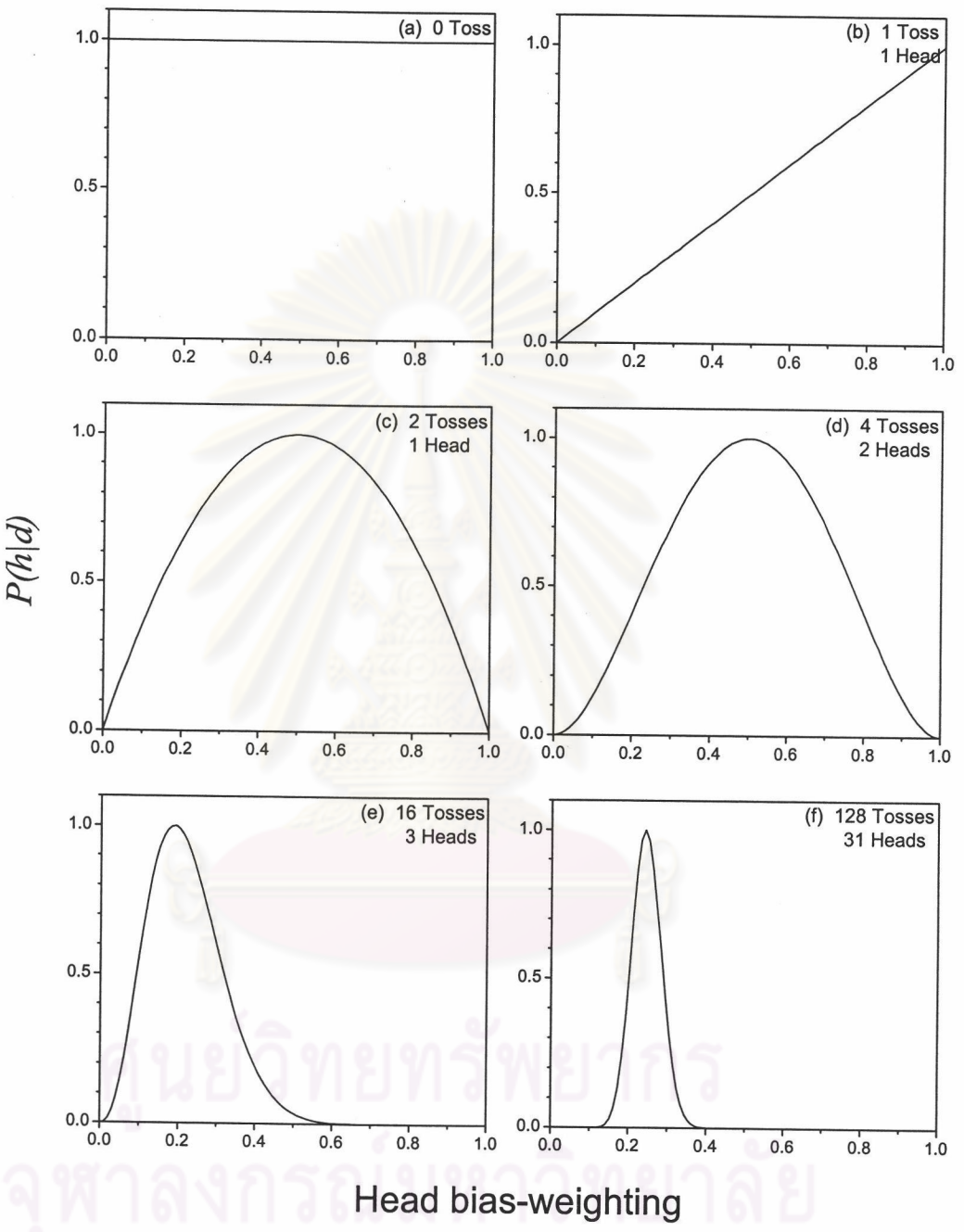

Figure 4.1: The posterior probability distribution of bias-weighting for head of a coin (after Sivia (1996)).

to represent the uncertainty in a solution; the narrower width the more confidence on inference.

4.3.1 Gaussian likelihood

In an experiment, there are noises associated with the measured data. Considering only the random noise σ, a single measurement of data d gives

$$d = f(x) \pm \sigma, \tag{4.13}$$

where noise is assumed to be a normal distribution (Gaussian noise). A function $f(x)$ defines the theoretical relation between the variable x and the observed data d. The Gaussian distribution is a theoretical model which is usually used to describe a random noise in measurement. Considering the probability of measuring a single variable y from the distribution which has a mean μ and a finite variance σ^2. The Gaussian distribution is defined as

$$G(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{4.14}$$

For the case of M variables $\mathbf{y} = \{y_k | k = 1, 2, \ldots, M\}$ with the mean vector $\boldsymbol{\mu} = \{\mu_k | k = 1, 2, \ldots, M\}$ and the constant covariance matrix $\mathbf{C}$, the multivariate Gaussian distribution is generally given by

$$G(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{M}{2}} Det(\mathbf{C})} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\intercal} \mathbf{C}^{-1} (\mathbf{y}-\boldsymbol{\mu})\right), \tag{4.15}$$

where $Det(\mathbf{C})$ is the determinant of $\mathbf{C}$. The diagonal components of covariance matrix $\mathbf{C}$ are the variance σ_k^2 of each datum index k, and the off-diagonal components are the covariance σ_{ij}^2 between a pair of data.

In the problem of inferring the variables $\mathbf{x} = \{x_i | i = 1, 2, ..., N\}$ by measuring data $\mathbf{d} = \{d_k | k = 1, 2, \ldots, M\}$ imposed with random noise, the likelihood

function can be assigned to the Gaussian distribution, and it is called the Gaussian likelihood. The Gaussian likelihood for multivariable case is presented as

$$P(\mathbf{d}|\mathbf{x},\mathbf{C}) = \frac{1}{(2\pi)^{\frac{M}{2}} Det(\mathbf{C})} \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{f}(\mathbf{x}))^{\intercal}\mathbf{C}^{-1}(\mathbf{d}-\mathbf{f}(\mathbf{x}))\right), \quad (4.16)$$

where $\mathbf{f}(\mathbf{x})$ is the known model of the problem and the covariance matrix $\mathbf{C}$ is already known. For simplification, the measured data are assumed to be independent to each other so that only the diagonal components of $\mathbf{C}$ exist. The likelihood can be simply written as

$$P(\mathbf{d}|\mathbf{x},\mathbf{C}) = \frac{1}{(2\pi)^{\frac{M}{2}} \prod_{k=1}^{M}\sigma_k^2} \exp\left(-\sum_{k=1}^{M}\frac{(d_k - f_k(\mathbf{x}))^2}{2\sigma_k^2}\right). \quad (4.17)$$

A summation term in the bracket of exponent can be considered as the chi-square misfit of measured data. Maximizing the Gaussian likelihood is equivalent to the method of chi-square reduction. In general, the technique of maximizing the likelihood function is called the method of maximum likelihood.

4.3.2 Entropic prior

In many physics problems, especially the spectral analysis, the interested spectrum $A(x)$ is a positive and additive distribution (PAD) which has two properties [Sivia (1996)].

1) Positive: All components of the distribution are non-negative, $A(x) \geq 0$ for all x.

2) Additive: The summation of all components has a physical meaning.

For example, the power spectrum of light, the electron density in crystal, and mobility spectrum are positive and additive distributions. The properties of PAD relates to the probability concept. Since the considered spectrum is PAD, the maximum entropy principle can be applied to the prior distribution. Because

the probability is defined positive and sum up to unity, Skilling (1989) suggested that the most appropriate prior probability for PAD is the entropic prior

$$P\left(A\left(x\right)|\alpha, m\left(x\right)\right) = \frac{1}{Z_s} \exp\left(\alpha H\right). \tag{4.18}$$

H is the Shannon-Jaynes entropy or the relative entropy which can be expressed generally as

$$H\left(A\left(x\right)\right) = \int \left[A\left(x\right) - m\left(x\right) - A\left(x\right) \log\left(\frac{A\left(x\right)}{m\left(x\right)}\right)\right] dx, \tag{4.19}$$

where $A(x)$ and $m(x)$ are not normalized and the integral is taken over the range of interest. The function $m(x)$ is a default model and α is a real positive constant. Z_s is a normalization constant. The default model $m(x)$ is the PAD that defines the initial shape of $A(x)$. It is consistent to the background knowledge about the problem when the observed data is not considered. It is usually assigned to a uniform distribution if the prior information of the spectrum $A\left(x\right)$ is unknown. The parameter α is adjustable. It corresponds to the width of entropic prior probability in such a way that the larger α the sharper the entropic prior. The entropy function H is a concave function. It measures the distance of $A\left(x\right)$ in relative to $m\left(x\right)$. H is globally maximum at the value of zero when $A\left(x\right)$ and $m\left(x\right)$ are equal. H is negative when $A\left(x\right)$ departs from $m\left(x\right)$, and the magnitude of H is qualitatively larger when $A\left(x\right)$ is further apart from $m\left(x\right)$.

When $A(x)$ and $m(x)$ are normalized to unity, the entropy in Eq. (4.19) is reduced to

$$H = -\int A\left(x\right) \log\left(\frac{A\left(x\right)}{m\left(x\right)}\right) dx. \tag{4.20}$$

This reduced form is slightly different from Eq. (4.7) by the default model $m\left(x\right)$, and the values are only different by some constant. The advantage of Eq. (4.20) over Eq. (4.7) is to allow adding the background information into the prior probability.

In the absence of measured data, the posterior is considered proportional to the prior distribution. The most probable spectrum $A(x)$ that maximizes the posterior probability can be obtained by maximizing the entropy according to the principle of maximum entropy.

4.4 Bayesian mobility spectrum

In order to apply Bayes' theorem to mobility spectrum analysis, one has to relate the probability distribution terms to the problem terms. In the present case, the posterior probability $P(\{s_i\} \mid \{\sigma_j\}, \{var(\sigma_j)\})$ represents the probability of mobility spectrum $\{s_i\}$ from the measured data $\{\sigma_j\}$ (σ_{xx} and σ_{xy}) and variances $\{var(\sigma_j)\}$. It is connected to the likelihood function $P(\{\sigma_j\} \mid \{s_i\}, \{var(\sigma_j)\})$, the prior probability $P(\{s_i\})$, and the evidence $P(\{\sigma_j\})$.

The likelihood function provides all the statistical information of the measured data given by a certain mobility spectrum and noise in the measurement. In the case of Gaussian noise, the likelihood function is presented as Gaussian distribution of misfit where each data, taken at different magnetic field strengths, is assumed independent. This gives

$$P(\sigma_j \mid \{s_i\}, var(\sigma_j)) = \frac{1}{\prod_{j=1}^{2M} (2\pi \cdot var(\sigma_j))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\chi^2\right), \tag{4.21}$$

where

$$j = \begin{cases} 1, 2, \ldots, M & \text{for } \sigma_{xx} \\ M+1, M+2, \ldots, 2M & \text{for } \sigma_{xy} \end{cases},$$

which gives

$$\chi^2 = \sum_{j=1}^{2M} \frac{\left(\sigma_j - \sum_{i=1}^{N} K_{ji} \cdot p_i\right)^2}{var(\sigma_j)}. \tag{4.22}$$

M and N are already defined in Section 4.1.

The prior probability distribution $P(\{p_i\})$ contains the knowledge about mobility spectrum $\{s_i\}$ before the measured data is available. Since the mobility spectrum is positive and additive, the appropriate prior probability is an entropic prior (Eq. (4.18)),

$$P(\{s_i\} \mid \{m_i\}, \alpha) = \frac{1}{Z_s} \exp(\alpha H). \tag{4.23}$$

H is the entropy function in Eq. (4.20), which is rewritten in a discrete form as

$$H = -\sum_{i=1}^{N} p_i \log\left(\frac{p_i}{m_i}\right), \tag{4.24}$$

where

$$\sum_i p_i = \sum_i m_i = 1. \tag{4.25}$$

The default model $\{m_i\}$ is defined to be a uniform distribution over the interested mobility range. This is to assume that all carriers have the same conductivity if the data is not available. The normalization term Z_s is dependent on the parameter α and it is approximately given by [Jarrell and Gubernatis (1996)]

$$Z_s(\alpha) \approx \left(\frac{2\pi}{\alpha}\right)^{\frac{N}{2}}. \tag{4.26}$$

The evidence $P(\{\sigma_{xx}, \sigma_{xy}\})$, the denominator in Bayes' formula, is a normalized constant that is neglected in the calculation process.

Combining all probability terms, the posterior probability distribution can be summarized as

$$\begin{aligned} P(\{s_i\} \mid \{\sigma_j\}, \{var(\sigma_j)\}, \{m\}, \alpha) &\propto \exp\left(-\frac{1}{2}\chi^2 + \alpha H\right) \\ &\propto \exp(Q), \end{aligned} \tag{4.27}$$

where $Q = -\frac{1}{2}\chi^2 + \alpha H$ and the normalization terms are neglected.

To determine the most probable mobility spectrum $\{p_i\}_\alpha$ for a given α, the posterior probability is maximized until Q is globally maximum. The problem of

maximizing Q can be considered as the regularization technique [Press et al. (1986)]. The technique is to fit the spectrum to the measured data by minimizing χ^2 where H is a constraint, which is kept as maximum as possible. In addition, the entropy is used to stabilize the convergence of the optimization procedure. At this stage, the parameter α can be considered as the Lagrange multiplier that controls the weight of the entropy H and the χ^2 function. The small value of α makes the minimizing of χ^2 important. On the other hand, the large value of α enhances the maximizing of the entropy H. There are several algorithms available to perform the task. In this thesis, we follows the applications in Barradus et al. (1999). In their approach, the MCMC sampling algorithm was used. It has been applied successfully in the thin film depth profile reconstruction problem by Barradas et al. (1999). The set of spectrum $\{p_i\}$ will be generated randomly to be a Markov chain. By the Markov process, the probability distribution of the spectrum will converge to the posterior distribution. The most probable spectrum $\{p_i\}_\alpha$ occurs when a maximum value of Q is found. In addition, the Markov chain provides a set of sampled spectra according to the posterior probability distribution. The obtained probability distribution gives all the statistical information about the solution spectrum such as the most probable spectrum (more details are described in Section 4.5), expectation values, variances, and confident interval.

Bayesian with Maximum entropy procedure using MCMC is performed according to these following steps.

Step 1) Calculating the most probable spectrum for a given α.

At the first step, a large value of α is set allowing the entropy term to play an important role in the optimizing process. This step aims to calculate the extreme smooth spectrum. A procedure starts at the global maximum of H, $\{p_i\} = \{m_i\}$. Then MCMC method is used to sample a set of spectrum $\{p_i\}$ from the posterior

probability distribution. When the Markov chain reaches its equilibrium state, the desired spectrum can be expected.

Step 2) Finding a suitable α.

The value of α has to be slowly stepwise decreased toward zero during the calculation [Gallicchio and Berne (1996), Jarrell and Gubernatis (1996), and Nakahara (2001)]. As a result, the χ^2 term becomes dominant and the spectrum tends to reduce misfit to the data. The feature of the spectrum then becomes clearer (sharper peak). For a fixed α, χ^2 slowly converges to a certain value, ideally $2M$, then the parameter α is selectively stopped. At this stopped α, the sampling still continues and the members of Markov chain are recorded as a set of solution spectrum. It is recommend that Markov chain should be done after the equilibrium is reached.

Step 3) Calculating the mobility, carrier concentration and error bound.

By the statistical method, the mean spectrum and its covariance matrix are expected from the set of solution spectra in Step 2). The resultant spectrum is expected with vertical error bars which are the standard deviation of each partial conductivity. Using Eqs. (2.34) and (2.35), Hall mobility and Hall concentration of each carrier species are calculated. The uncertainties of the calculated values are also obtained by the method of error propagation described in Section 4.6.

The Bayesian technique was performed on a synthetic data in order to demonstrate a mechanism of the developing procedure. The synthetic data of Hall coefficient and resistivity ($\mu_1 = 2,000\ \mathrm{cm^2V^{-1}s^{-1}}$, $n_1 = 1\times10^{11}\ \mathrm{cm^{-2}}$ and $\mu_2 = 6,000\ \mathrm{cm^2V^{-1}s^{-1}}$, $n_2 = 1 \times 10^{11}\ \mathrm{cm^{-2}}$) are generated for 100 magnetic field points uniformly distributed from 0.1 to 10 Tesla. The mobility of each carrier species was assigned the mobility spectrum of normal distribution with standard deviation of 250 and 500 $\mathrm{cm^2V^{-1}s^{-1}}$ respectively. The 0.1 % Gaussian noise was added to Hall coefficient and resistivity data.

The mobility range was defined from 100 to 10,000 $cm^2V^{-1}s^{-1}$ for holes and from -100 to -10,000 $cm^2V^{-1}s^{-1}$ for electrons with 50 mobility points spaced equally over the mobility range. The default model $\{m_i\}$ was chosen to be the uniform distribution. Fig. 4.2 shows the formation of mobility spectra at different iterations for $\alpha = 900$. The problem is quite well-defined because the number of data points (200) is greater than the number of parameters (50). The lowest value of mobility in the system corresponds to $\mu_{\min} B_{\max} = 2$ which makes the problem be sufficient for the analysis. In Fig. 4.3, the corresponding χ^2 in Eq. (4.22), and entropy H in Eq. (4.24) versus the iterations are shown. Initially, the mobility spectrum is uniform. χ^2 shows a large value of misfit about 10^8, and H starts at zero. After 1,000 iterations, the spectrum has a high degree of pointwise oscillation. The χ^2 decreases rapidly and H becomes a negative value. For a few thousands of iterations, the sharp peaks near the mobility 2,000 and 6,000 $cm^2V^{-1}s^{-1}$ are formed. Both χ^2 and H decreases extremely. During the first few 10,000 iterations, the feature of spectrum consists of several sharp peaks because the value of χ^2 is very large compared to αH. From 10,000 to 200,000 iterations, H increases slowly; however, χ^2 still decreases to produce a good fit. At 3,000,000 iterations, the required spectrum completely forms with satisfying smoothness, and peaks locate around the mobilities of 2,000 and 6,000 $cm^2V^{-1}s^{-1}$ respectively. Even though two peaks are not resolved completely, the result has validated the calculation procedure.

Fig. 4.4 shows the mobility spectrum from the same synthetic data but different α. It demonstrates that the smooth mobility spectrum requires a reasonable large α.

According to Step 2), α is manually decreased by a factor 0.9 for every 200,000 iterations. The evaluation of mobility spectrum is in Fig. 4.5. In Fig. 4.6, the corresponding χ^2 (solid square) and H (hollow circle) versus the number of iterations are shown. The spectrum in Fig 4.2 (f) is used as the initial state for this

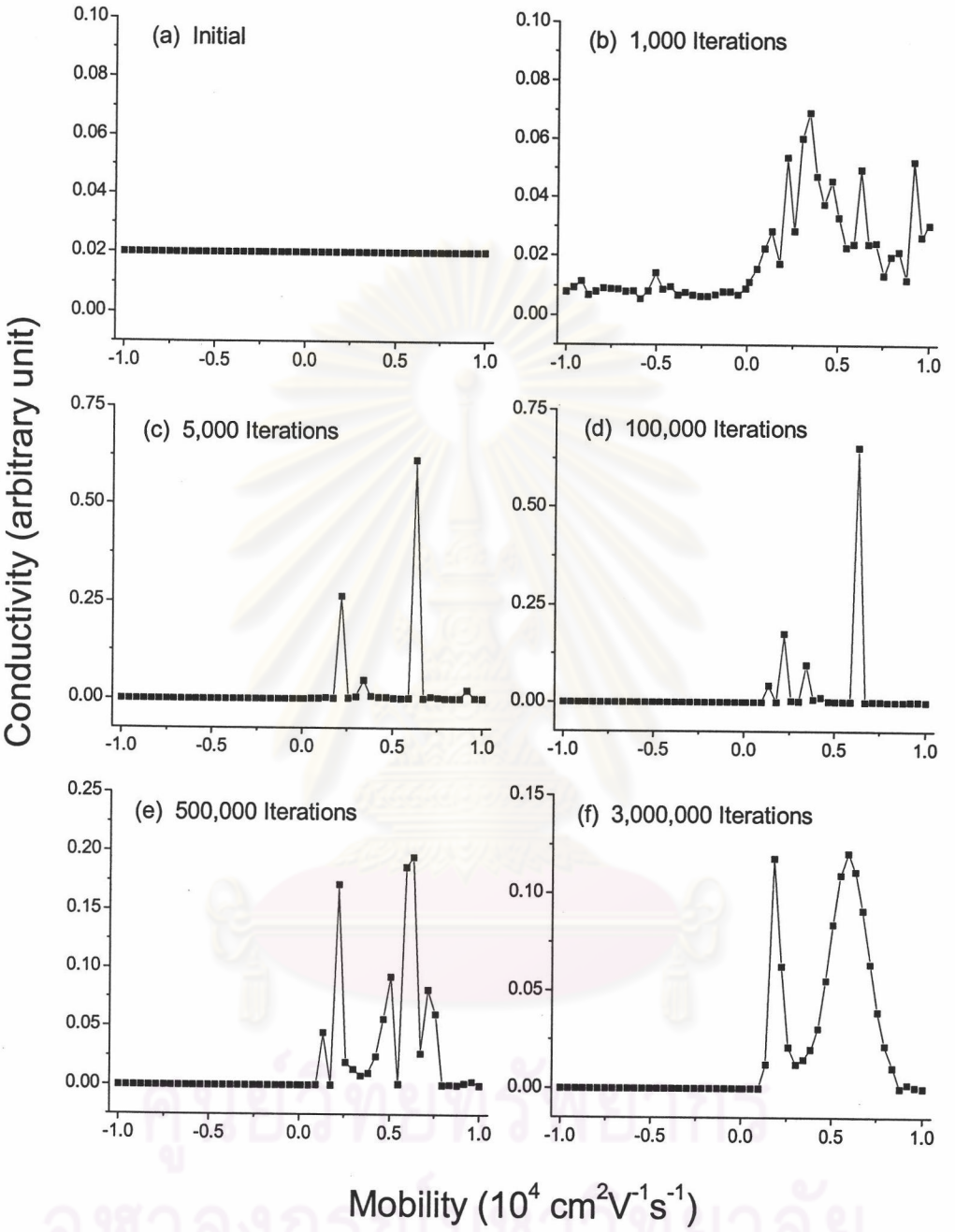

Figure 4.2: Bayesian mobility spectra at different iterations for synthetic data of two hole species (n_1=1×$10^{11}$$cm^{-2}$, μ_1=2,000 $cm^2V^{-1}s^{-1}$, n_2=1×$10^{11}$$cm^{-2}$, μ_2=6,000 $cm^2V^{-1}s^{-1}$ with standard deviations of 250 and 500 $cm^2V^{-1}s^{-1}$, respectively) with noise 0.1%. α is 900.

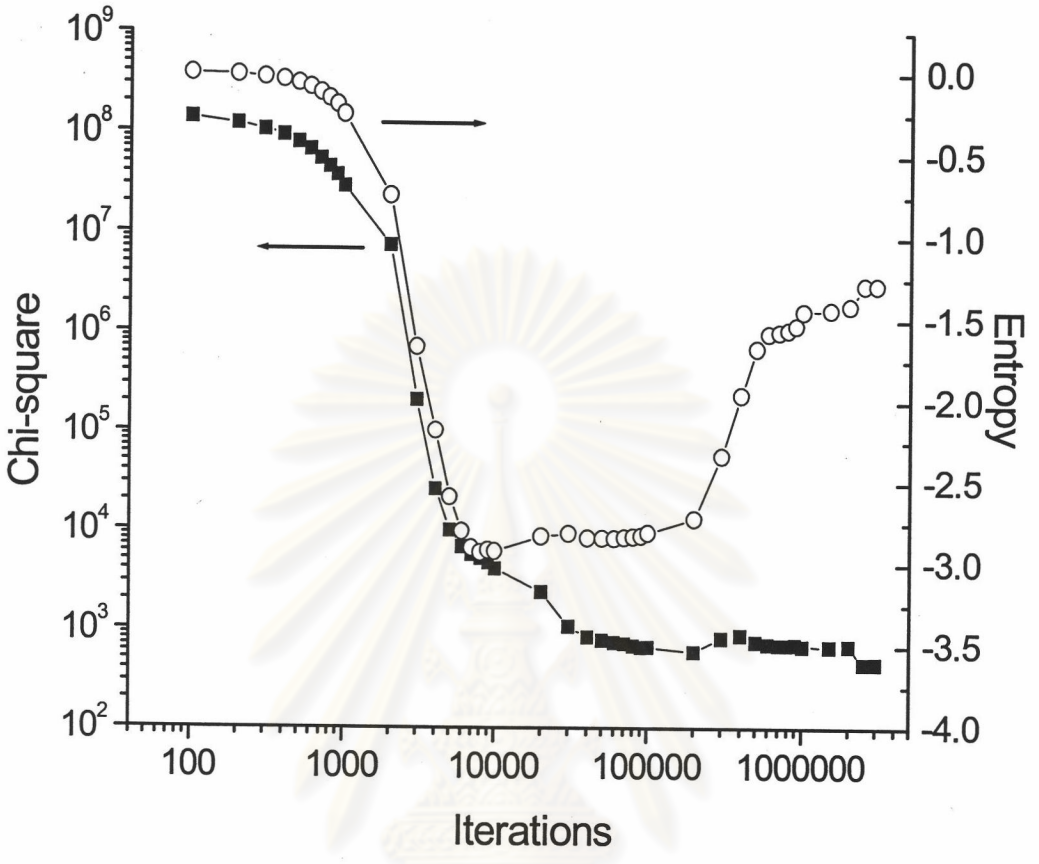

Figure 4.3: The Chi-square (solid square) and entropy (hollow circle) versus the number of iterations.

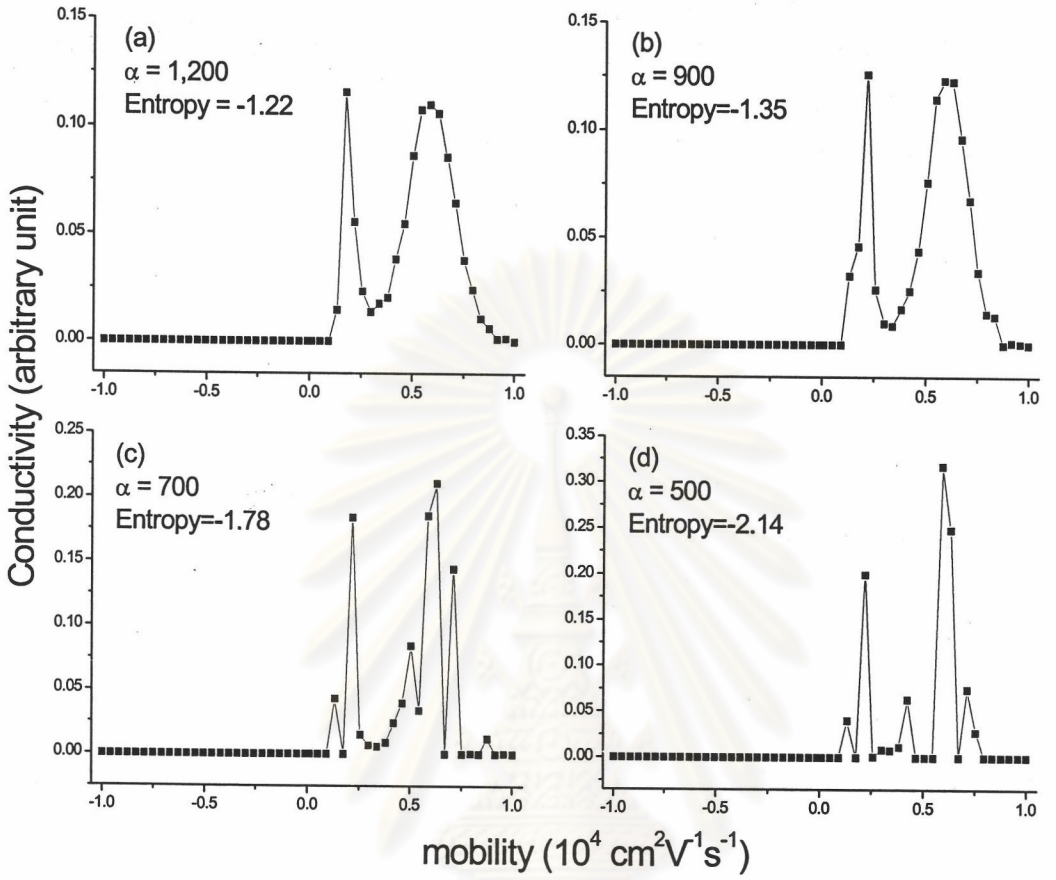

Figure 4.4: Bayesian mobility spectra of the test synthetic data for different α's at 2,000,000 iterations.

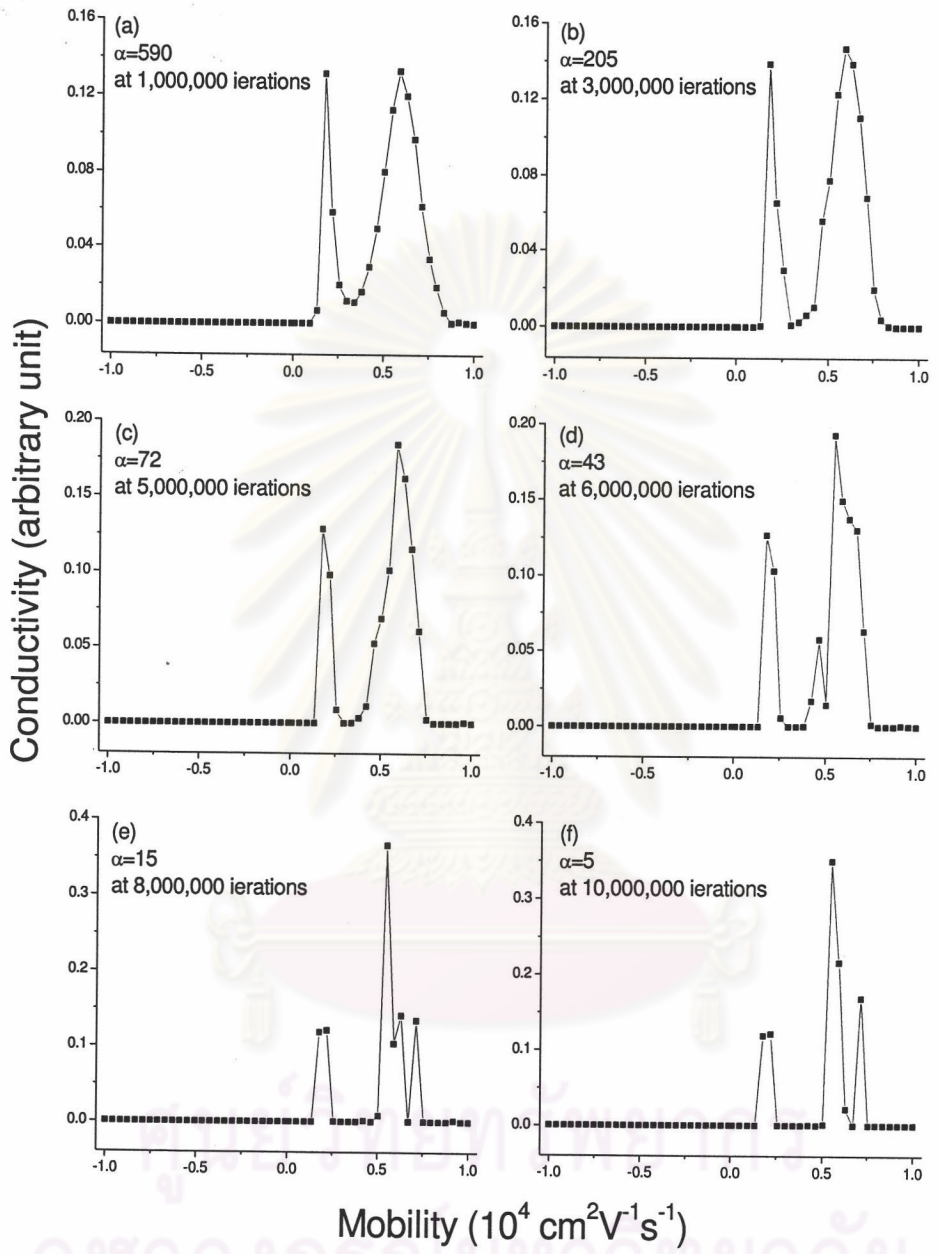

Figure 4.5: Bayesian mobility spectra of synthetic data (n_1=1×10^{11} cm^{-2}, μ_1=2,000 $cm^2V^{-1}s^{-1}$, n_2=1×10^{11} cm^{-2}, μ_2=6,000 $cm^2V^{-1}s^{-1}$ with standard deviations of 250 and 500 $cm^2V^{-1}s^{-1}$, respectively, and with 0.1% random noise) at different iterations (continued from spectrum in Fig. 4.2 (f)). α is decreased from 900 by a factor of 0.9 every 200,000 iterations.

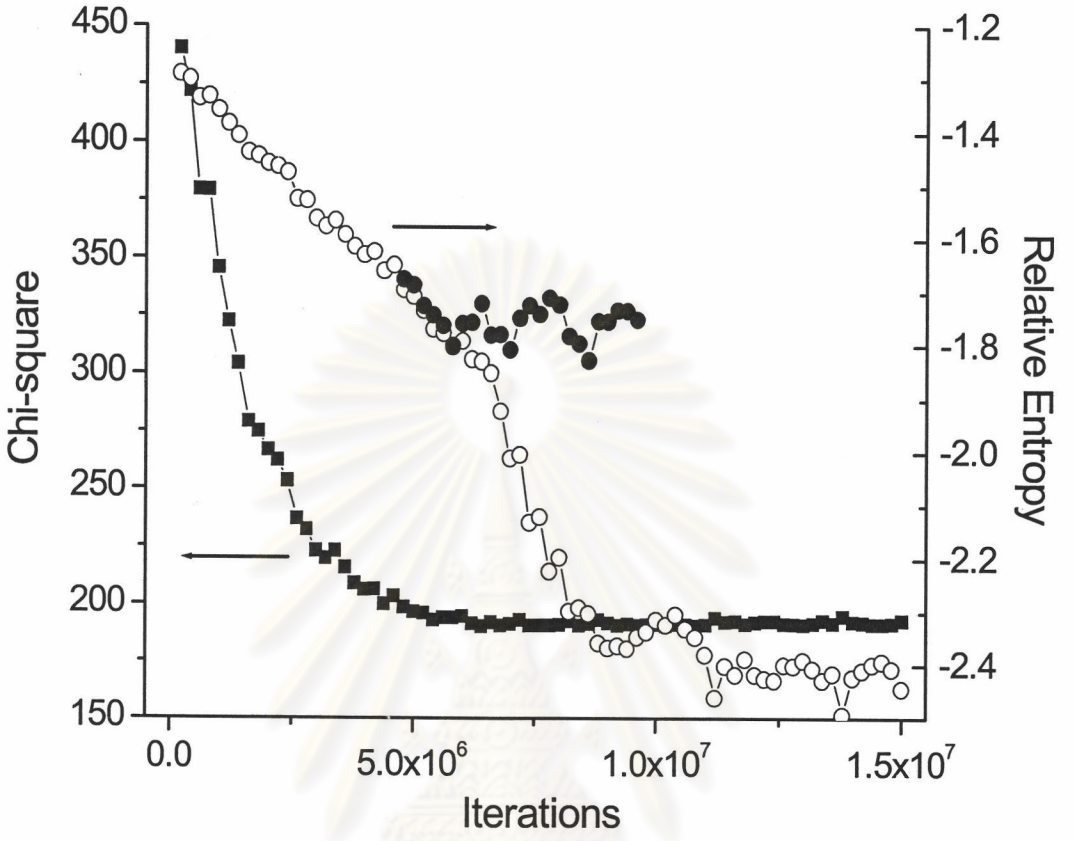

Figure 4.6: Chi-square (solid square) and entropy (hollow circle) versus iterations where α is decreased from 900 by a factor of 0.9 every 200,000 iterations. Solid circle represents the entropy where α is stopped at 80 at 5,000,000 iterations.

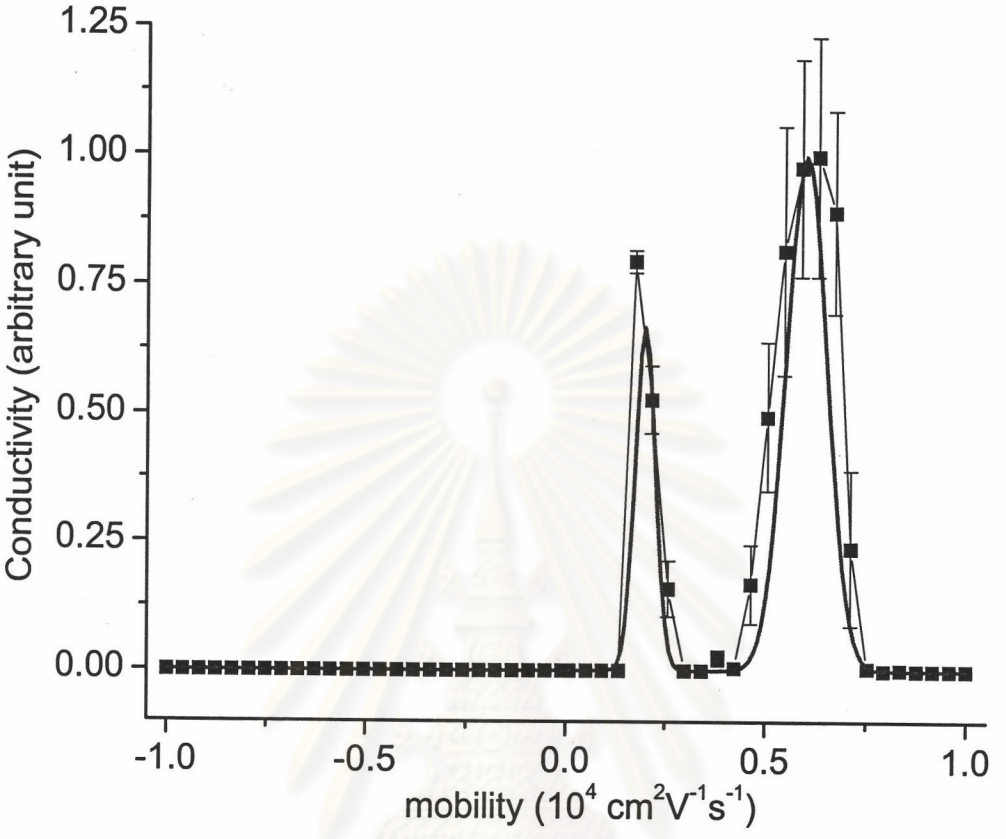

Figure 4.7: The resultant mobility spectrum (line-solid square) with error bars obtained from full operation of Bayesian method where α is 80. Two hole peaks located at mobility about 2,000 and 6,000 $cm^2V^{-1}s^{-1}$. The true mobility spectrum (thick line) is shown with equal maximum partial conductivity in arbitrary unit. The Bayesian peaks are little broader than the true peaks and its error bars do not perfectly cover the true spectrum.

step. Decreasing α results in the reduction of both χ^2 and H. At 5,000,000 iterations, χ^2 approximately reaches its saturated theoretical value of 195, $\left(\frac{\chi^2}{2M} \approx 0.97\right)$. Over 5,000,000 iterations, χ^2 fluctuates around the saturated value but H continues to decrease. This causes the split peaks. Therefore, we will choose to stop when χ^2 saturates. At this point, the solution is regarded in equilibrium and the Markov chain will be collected. The solid circle in Fig. 4.6 represents H at $\alpha = 80$ showing the equilibrium state. The final mobility spectrum with error bar is shown in Fig. 4.7. The Bayesian peaks are little broader than the true peaks, and its error bar do not perfectly cover the true spectrum.

4.5 Mobility Calculation and Error Analysis

Each peak represents the distinct species of carrier. Summing the partial conductivities over the mobility range of each peak, the conductivity of that species is obtained :

$$s = \sum_{i \in peak} s_i. \tag{4.28}$$

Hall mobility and Hall concentration of each species can be calculated by considering all partial conductivity points constituting a particular peak (see Eqs. (2.34) and (2.35)) which gives

$$\mu_{Hall} = \frac{\sum_{i \in peak} \mu_i s_i}{\sum_{i \in peak} s_i} \tag{4.29}$$

and

$$n_{Hall} = \sum_{i \in peak} \frac{s_i}{e\mu_i}. \tag{4.30}$$

The corresponding uncertainties are obtained by using an error propagation equation.

To estimate the uncertainty in determination of a variable $x = f(u, v, \ldots)$, the error propagation equation is given as

$$\sigma_x^2 \approx \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2 + \cdots + 2\sigma_{uv}^2 \left(\frac{\partial x}{\partial u}\right)\left(\frac{\partial x}{\partial v}\right) + \cdots, \tag{4.31}$$

where σ_x^2 is the variance of variable x, σ_u^2 and σ_v^2 are the known variances of variable u and v respectively, and σ_{uv}^2 is covariance between u and v [Bevington and Robinson (1992)]. Using Eq. (4.31) to estimate the uncertainty of conductivity, mobility and concentration given by Eqs. (4.28) to (4.30), the variances of those parameters are

$$\sigma_s^2 = \sum_i \sigma_{s_i}^2 + 2\sum_{i \neq j} \sigma_{s_i,s_j}^2, \tag{4.32}$$

$$\frac{\sigma_{\mu_H}^2}{\mu_H} = \frac{\sum_i (\mu_i)^2 \sigma_{s_i}^2 + 2\sum_{i \neq j} (\mu_i \mu_j) \sigma_{s_i,s_j}^2}{\left(\sum_i s_i \mu_i\right)^2} + \frac{\sum_i \sigma_{s_i}^2 + 2\sum_{i \neq j} \sigma_{s_i,s_j}^2}{\left(\sum_i s_i\right)^2}, \tag{4.33}$$

and

$$\sigma_{n_{Hall}}^2 = \sum_i \left(\frac{1}{e^2 \mu_i^2}\right) \sigma_{s_i}^2 + 2\sum_{i \neq j} \left(\frac{1}{e^2 \mu_i \mu_j}\right) \sigma_{s_i,s_j}^2, \tag{4.34}$$

for conductivity , Hall mobility, and Hall concentration respectively.

4.6 Markov chain Monte Carlo

Markov chain Monte Carlo is a numerical technique that shares the same principle with Monte Carlo integral. It is the method for sampling an object from any distribution. By constructing a Markov chain in a sample space, the distribution of a chain tends to converge to its assigned stationary distribution. Then the Markov chain reaches its equilibrium, further state of chain will represent a random sample which forms desired target distribution.

The Markov chain is a sequence of random variables, $\{X_t | t = 0, 1, 2, \ldots\}$, that sampling from a transition probability distribution $P(X_{t+1}|X_t)$. The next state X_{t+1} depends only on the current state X_t. The Markov chain will converge to a

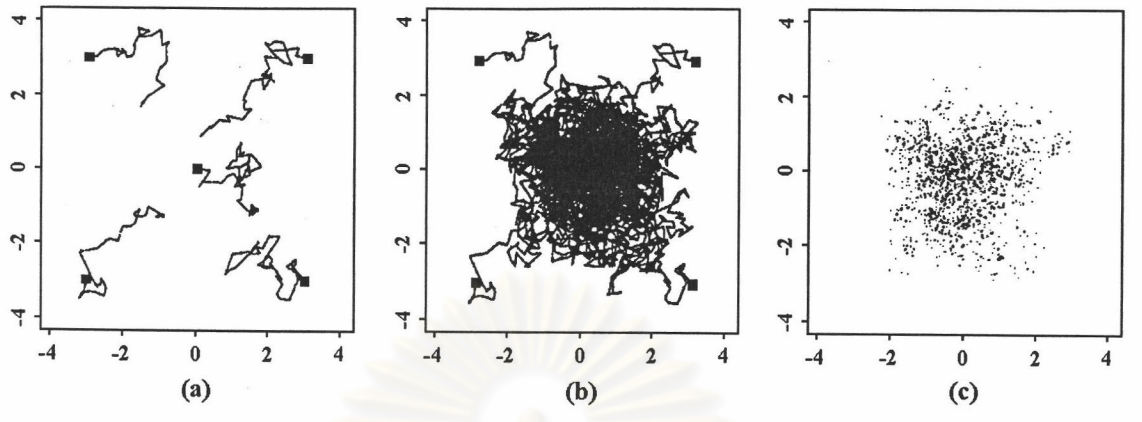

Figure 4.8: Five independent sequences of a Markov chain simulation of the bivariate unit normal distribution. Solid squares indicate different starting points (a) First 50 iterations, the sequence are still far from convergence. (b) After 1000 iterations, the sequences are nearer to convergence. (c) Second half of the sequence states (after Gelman et al. (1995)).

unique stationary distribution that does not depend on choices of the starting point X_0. Fig. 4.8 shows an example with the target distribution of a bivariate unit normal distribution $N(0,1)$. In Fig 4.8 (a), five independent Markov chains are constructed as random walks in two dimensions. After 1,000 iterations, Markov chains in Fig. 4.8 (b) are likely to display a convergence to a bivariate stationary distribution as their target distribution. The empirical bivariate unit normal distribution is illustrated by the last 500 states of those chains in Fig. 4.8 (c). In general, the Markov chain Monte Carlo method is used in many applications such as estimating the expectation value from any distribution function, multi-dimensional integration and evaluating the feature of Bayesian posterior distribution. For example; the expectation $E[f(X)]$, where X has any distribution, is calculated from the output of Markov chain by

$$E[f(X)] = \langle f(X) \rangle = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t). \tag{4.35}$$

Mean and variance are estimated by

$$\overline{X} = \frac{1}{n-m} \sum_{t=m+1}^{n} X_t \tag{4.36}$$

and

$$\sigma^2 = \frac{1}{n-m-1}\sum_{t=m+1}^{n}\left(X_t - \overline{X}\right)^2, \tag{4.37}$$

where m is the number of iterations during the burn-in period. More details can be found in Gelman et al. (1995) and Gilks et al. (1996).

The Metropolis algorithm

There are many algorithms to construct a Markov chain. Metropolis algorithm is the specific one proposed by Metropolis et al. (1953) and has the following sequence:

1) The algorithm starts with any initial point $X_t = 0$.

2) At time t, the candidate point X^* was sampled from a proposal distribution $q(\cdot|X_t)$ which depends only on the current point X_t. The proposal distribution must be symmetric, that is $q(X_{t+1}|X_t) = q(X_t|X_{t+1})$ for all t, such as a multivariate distribution with mean X_t and some constant covariance matrix. In addition, the random walk is the most popular tool assigning the proposal distribution function .

3) The candidate point X^* is accepted with probability

$$P_{\text{accept}}(X_t, X^*) = \min\left(1, \frac{\pi(X^*)}{\pi(X_t)}\right) \tag{4.38}$$

where $\pi(X)$ is the interested distribution. $X_{t+1} = X^*$ is then set if the candidate is accepted; otherwise, $X_{t+1} = X_t$. These step is repeated until convergence is approximated. Hastings (1970) generalized the method for an arbitrary proposal including an asymmetric distribution and the accepted probability is modified to be

$$P_{\text{accept}}(X_t, X^*) = \min\left(1, \frac{\pi(X^*)\, q(X_t|X^*)}{\pi(X_t)\, q(X^*|X_t)}\right), \tag{4.39}$$

which is called Metropolis-Hastings algorithm.

An appropriated proposal distribution is necessary because it contributes a rapid convergence to the stationary distribution. There are many techniques suggested to improve the efficiency of the used algorithm. However, this topic goes beyond the basic idea of Markov chain simulation. The convergence monitoring method may be done empirically by constructing several chains with difference starting points. If all chains merge together then Markov chains converge to their equilibrium states.