On Systems with Limited Communication

A thesis presented

by

Jian Zou

to

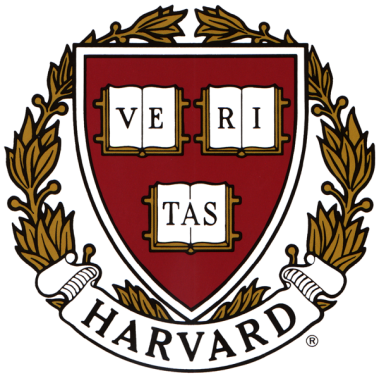
The Division of Engineering and Applied Sciences
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in the subject of

Engineering Sciences

Harvard University
Cambridge, Massachusetts

May 2004



To My Family.

Abstract

Communication constraints on channels connecting different components of a control system are traditionally ignored. As a result of advances in communication, computation and network, many current control systems are distributed, asynchronous and networked. The traditional assumption which ignores communication constraints doesn't apply to those systems. This thesis considers the problem in general and focuses the impact of communication constraints on state estimation.

Under a set of assumptions which includes ignoring time delay caused by limited communication, systems of limited communication are formulated as two classes of mathematical models. The state estimation problems for these classes of models are presented. Quantized measurement sequential Monte Carlo(QMSMC) method, Quantized measurement Kalman filter and Quantized measurement conditional sampling method are proposed as candidates for state estimator. Their properties, respective advantages and disadvantages are discussed. In particular, QMSMC's asymptotical optimality is proven under a few further assumptions regarding system model and quantizers.

These methods are applied to models of realistic complexity and their effectiveness are demonstrated through simulation-based comparison with existing other methods.

Optimization of quantization for Gauss-Markov process is also considered. Specifically, optimal quantizer in least squared estimation error sense for second order Gauss-Markov process is numerically studied with respect to Quantized measurement Kalman filter.

Acknowledgements

About four years ago, I arrived on campus with enthusiasm for new intellectual journey and admire for Harvard. As I prepare to leave Harvard now, I am looking forward to my future endeavors with more confidence. Four years of graduate student life gave me the chance to advance further intellectually, inspired a deeper respect for Harvard as a great institution.

Many individuals contributed to my experiences at Harvard. I owe the most to my advisor, Professor Roger Brockett. Roger was largely responsible for bringing me to Harvard. I was deeply grateful with his insistence on clear thinking, rigorous mathematical education and his support during the years.

I wish also to thank the rest of my committee, Professor Aleksandar Kavcic, Professor Navin Khaneja, Professor Garrett Stanley, for their suggestions and feedbacks concerning my research. Navin, in particular, provided many valuable advices on my research and served as an inspiring role model of being dedicated and successful for me.

My time at Harvard was greatly enhanced by my best friend, officemate Michael. His humor, energy, good manner, generous sharing of his knowledge and patient tutoring on English is greatly appreciated. Special thanks goes to Haidong Yuan. Our discussions about many concrete mathematical problems prove to be productive and fun. Other current or previous officemates and labmates, Ben, Jason, Ali, Sean, Randy, Mark, Manuela provided help in English, good accompany and hearty support during my life at Harvard. Thank you, folks!

Many other friends of mine at Harvard helped me have a broader experience as a graduate student. Interacting and learning from them was not only entertaining but also an invaluable experience for me in understanding and respecting diversity in opinions, cultures and lifestyles.

Finally, I wish to extend my thanks to my family, Yue, Yi, Yuhua and Wenbo. Their encouragement for efforts and tolerance of possible failures irreplaceably helped me endure many of the ups and downs in my life and reminds me patience, good will and tolerance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Information theoretical issues are traditionally decoupled from consideration of decision and control problems. The decoupling is achieved by ignoring the capacity constraints on communication channels connecting different components of a system. Standard assumptions in decision and control theory suppose all communication channels within a system are of infinite channel capacities. This assumption implies that all data can be transmitted over those channels with infinite precision and zero time delay. Decoupling information theoretical issues from the decision and control problems in a system greatly simplifies the analysis and generally works well for classical applications. Since in many classical systems, exclusive communication channels with sufficient capacity are dedicated for the data transmission among different components of a system. The constraints imposed on the channels and the effects resulted from them are often negligible.

Advances in communication, computation and networks greatly expanded the range and complexity of a control system. Many newly emerged control systems are distributed, asynchronous and networked. These systems pose challenges to the traditional assumption which ignores the communication constraints on channels among different components of a system. Integrating communication constraints into estimation and control of a system becomes an inevitable task on our way to achieve deeper understanding of distributed, asynchronous and networked control systems.

We present a few example systems where communication constraints are too significant to be ignored.

1.1.1 Control of elements in Micro-Electro-Mechanical Systems (MEMS)

MEMS is the integration of mechanical elements, sensors, actuators, and electronics on a common silicon substrate through microfabrication technology. MEMS promises to revolutionize nearly every product category by bringing together silicon-based microelectronics with micromachining technology. Current and future generations of technology will package as many as 10^4 to 10^6 units (actuators or sensors) on a single MEMS chip. It is practically impossible to build an exclusive communication channel for each unit which connects it with its corresponding controller. A common communication channel will be shared by many units and their controllers. The channel between one unit and its corresponding controller will only have limited capacity. The communication constraints on this channel will be more severe as the level of integration on one MEMS chip increases, assuming the capacity of the common channel doesn't increase as fast as the level of integration does.

1.1.2 System with wireless communication

As a result of rapid growth in wireless communication, we are facing an increasing number of control systems where different components of the system are connected through wireless digital communication channels. The necessity of choosing wireless communication arises naturally in situations where some components of the system are required to be mobile. In most cases, the mobile component is the plant in the control system equipped with certain computational ability. We restrict our discussion to only those cases. The plant will be referred to as remote plant in the following.

Based on the assumption regarding remote plants' computational capacities, control systems with wireless communication can be approximately organized into two categories. The following discussion explains the information transmission required over the wireless channel for both categories.

The first category assumes that the remote plant doesn't possess any computational ability. In other words, the remote plant only consists of a plant, sensors and actuators. A control law has to be computed at a distant controller. Information about the state observations and control commands needs to be transmitted over the wireless channel between the remote plant and the controller. Control of a simplified version of an unmanned vehicle falls into this category. Remotely controlled car model as a toy is one example many of us have experienced.

The second category assumes the remote plant has enough computational ability to

compute control law. However, in many cases, the remote plant lacks all information required for the control law. This happens in coordination of several remote plants. Consider formation control for unmanned aerial vehicle (UAV). To achieve coordination, in most cases, individual vehicles should execute control law dependent on information about fellow vehicles. When the individual vehicle lacks the ability to gather all the information required, the information has to be transmitted over the wireless channels among vehicles and the ground controller.

Controlling a single vehicle in an environment for which the vehicle does not have sufficient information also falls into this category. The lack of information may result from a lack of sensing capability onboard the vehicle or lack of sufficient onboard storage capability for environment information. In order to compute a control law governing the movement of this vehicle, two options can be pursued. The first is to effectively represent the environment information and transmit it to the vehicle. The second is to transmit state information of the vehicle to the ground controller with access to the environment information and then the ground controller transmits control commands back to the vehicle.

As discussed above, information flow needs to be carried by wireless communication channels within a control system. A mobile remote plant often has limited payload hence limited power in its transmitter. Sometimes, the information transmission is performed over a long geographical range in a hostile environment. All these factors will result in limited, sometimes severely limited, capacity of the wireless channels. How to most effectively use the scarce communication resources to achieve our goal becomes an increasingly important problem as the gap between the capacity available and capacity required diminishes.

1.1.3 Biological Systems

In neurobiological systems, the controllers(such as the brain) and plants (such as muscle) are separated and connected via a neural system which in nature is a communication channel of limited capacity. Animals have exhibited an amazing ability to control the plant with limited communication capacity of their neural systems. Though distant from the examples discussed above, study of those systems maybe provide inspiring ideas for most effective usage of communication resources to achieve certain control goal.

1.2 Previous work

Classical system and control theory has established a powerful framework to analyze systems without communication constraints (referred to as classical systems). Incorporating communication constraints in analysis of a control system will cause difficulties in addition to the existing challenges.

To systematically investigate the effect of the communication constraints on systems is of theoretical significance and practical impact. Establishment of a theoretical framework analogous to the one for classical system and control theory will be of great value and requires sustained efforts. Many important questions should be answered in the framework, such as effective state estimator in least squared error sense, controllability and stability of the system (the definitions of controllability and stability should be carefully reconsidered), optimal control with respect to a joint function of communication and traditional costs and robust control for systems with limited communication.

Since communication constraints are intrinsically nonlinear and discontinuous, the possible framework would not be as elegant and beautiful as the one for classical systems. However, with certain approximation, a meaningful framework is still possible.

Inspired by Wong and Brockett (1997) in their seminal work, the subject has been attracting more and more attention in the literature and has been studied by several researchers from very different perspectives. Vastly different assumptions on the plant model, protocols used in the communication channel, exact forms of communication constraints and measurement models result in significantly different mathematical models and hence different approaches to the problem. Much of the work is still preliminary in the sense that it largely focuses on a very restricted class of systems and is based on many additional assumptions which is not necessarily reflecting the common practice in technology. To the author's best knowledge, none of these results has been tested against systems with realistic complexity and given significant gains over the existing methods.

As an attempt to serve as a building block for the general framework, this thesis focuses on seeking an effective state estimator in least the mean-squared-error sense. The results presented are obtained under a variety of assumptions which we view as reasonable. Their effectiveness and practicality are demonstrated for systems with realistic complexity.

1.3 Main Issues and Major Contributions

Generally speaking, the impact of limited communication on state estimation is twofold. First, it restrict the amount of information which can be transmitted from the plant to the state estimator. When the system makes analog measurements of its state variable, that requires the analog measurements to be quantized into a codebook of certain size in order to be transmitted reliably over the communication channel of limited capacity. As an irreversible process, quantization will cause inevitable information loss. Second, it causes time delay in estimation because of the lag between the time instant when measurements of the system are made and the one when the corresponding codewords are received at state estimator. The more severe the communication constraints, the more impact the two effects will have on the system.

Latency (time delay) has been a longtime topic in control and system theory. Several references have studied latency in great details. This thesis ignores latency for the sake of simplicity. Instead, it focuses on the impact of quantization on state estimation. More specifically, we only consider filtering problem in state estimation for a general class of systems with analog measurements. Filtering here means estimation at time index k of state using measurements up until time index k.

Traditionally, quantization is treated as additive noise. An overview of this perspective can be found in Gersho and Gray (1992). Specific assumptions as to the statistical nature of the quantization noise are made when they are treated as additive noise. The most common assumptions assume the quantization noise process are white, stationary and uncorrelated with the process being quantized. The purpose of those assumptions is to make it possible to handle the quantization analytically as a classical additive noise. Those assumptions eliminate the need to deal with the nonlinearity of a quantizer at the expense of oversimplifying the effect of quantization. For many high resolution applications, the model is valuable for obtaining satisfactory simple approximations to quantization. However, they would be grossly inaccurate in many circumstances where the resolution of quantization is low. Severe communication constraints would naturally result in low resolution quantization. Seeking other alternatives to deal with quantization and compare them against these common assumptions for the purpose of filtering becomes one of the central themes of this thesis.

Except for Gauss Markov models, optimal filters in the least mean-square-error sense for all other systems are nonlinear. Because of the quantization in the measurement model,

the optimal filtering problem for systems with communication constraints is inevitably nonlinear and defies a closed-form solution except in trivial cases. Since a closed-form optimal filter for those systems is analytically intractable, we have to seek an effective sub-optimal numerical approximation of the optimal filter for the classes of systems and quantizers under consideration.

Suppose we have found an effective sub-optimal numerical approximation of the optimal filter, then, the next step would be to try to optimize quantizer with respect to that specific suboptimal filter. The resulting "optimal" quantizer may be different from the optimal quantizer with respect to the optimal filter. However, those filters are the ones within our reach of analysis and of practical importance.

Essentially, we deal with optimization with respect to both the quantizer and the filter. We are reducing the set of filters in the consideration to the set of effective suboptimal filters and then optimize the quantizer with respect to that set of filters. The word "effective" and "optimize" are both interpreted with respect to mean-square-error in state estimation.

The thesis' contributions follow this logic. First, after modelling a system with limited communication, we propose a few suboptimal filters in mean-square-error sense for a few classes of models. More specifically, we study Quantized Measurement Sequential Monte Carlo (QMSMC) method, Quantized Measurement Kalman Filter (QMKF) or Extended Kalman Filter (QMEKF) and Quantized Measurement Conditional Sampling(QMCP) algorithm. We prove that under a few further assumptions regarding the underlying system and the quantizer in the measurement model, the QMSMC is asymptotically optimal in mean-square-error sense for a large group of systems and quantizers. In the proof, we establish a more clear and rigorous framework than the one of classical sequential Monte Carlo method available in Doucet et al. (2001).

We compare the relative advantages and disadvantages of these suboptimal filters based on simulation of systems with realistic complexity. In particular, we apply QMSMC to the navigation model of MIT instrumented X-60 SE Cell helicopter and QMCS method to the model of Harvard Robotic Lab rotary light weighted double pendulum.

Second, we study optimization of quantizer with respect to those suboptimal filters. As a preliminary, we first establish properties of the optimal quantizer for the standard normal distribution. Then, we numerically study the optimal quantizer for second order Gauss-Markov systems with different damping ratios with respect to QMKF and present the conclusion regarding the effect of damping ratio on the optimal quantizer.

1.4 Organization of rest of the thesis

Chapter 2 serves as the modelling part of the thesis. After making a set of assumptions, the systems with limited communication under consideration are identified with a few classes of models.

In chapter 3, we propose QMSMC and QMKF for one specific class of models and prove QMSMC's asymptotical optimality under further assumptions on the model and quantizer in measurement.

In chapter 4, we apply QMSMC to the problem of filtering from quantized measurement in the navigation model of MIT instrumented X-60 SE Cell helicopter and demonstrates its practicality by comparing it against QMEKF.

In chapter 5, we present QMKF and QMCS for another class of models under slightly different assumptions regarding the noise in measurement models. Empirical comparison of those two methods are presented for a few Gauss-Markov Models.

In chapter 6, we apply QMCS to the state estimation problem of Harvard Robotic Lab rotary light weighted double pendulum and show its improved performance compared with existing methods.

In chapter 7, we first consider the properties of the optimal quantizer for the standard normal distribution. Then, we numerically examine the impact of damping ratio in second order Gauss-Markov systems on the optimal quantizer with respect to QMKF.

Chapter 8 concludes the thesis and discusses future directions of research in this area.

Chapter 2

Mathematical Model of Systems with Limited Communication

2.1 Overview

Systems with limited communication is a broad concept. Every system in practice has limited communication. As discussed in chapter 1, we focus on the set of systems in which the communication constraints are so severe that ignoring communication constraints would fail to capture the nature of those systems.

Systems within this set can be significantly different from one another in terms of models of underlying physical objects or processes, type of communication channels within the system, protocols used in those channels, form of communication constraints and information to be transmitted. This thesis intends to address only a group of models whose configurations are similar to that of a system consisting of a simplified UAV and its ground controller. In order to mathematically model systems of this type, it is necessary to make reasonable assumptions which will simplify the analysis while preserving the essential impact of communication constraints on the system.

2.2 System Specifications and Assumptions

Consider the block diagram of systems with limited communication in figure 2.1.

The function of each component in the diagram is specificed as follows. "Plant" includes the physical object or process being controlled, sensors measuring its state and actuators. "y" denotes analog state measurements of the physical object or process. Analog to Digital

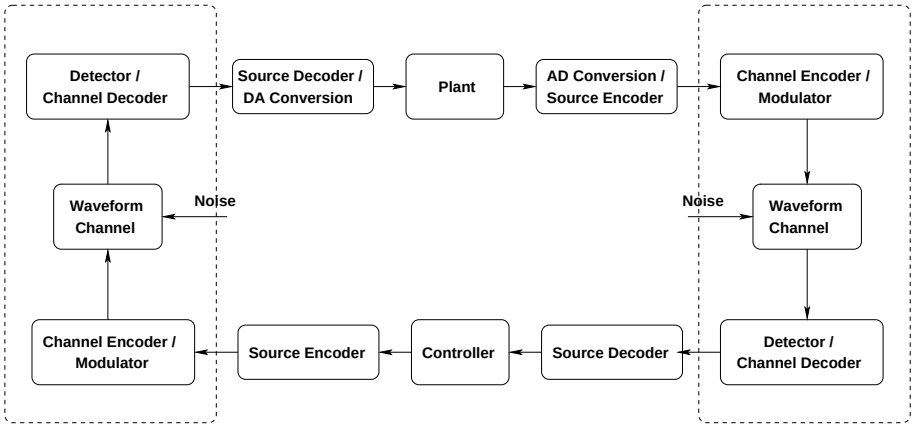


Figure 2.1: Block Diagram of Systems with Communication Constraints

(AD) Conversion in "AD Conversion / Source Encoder" consists of sampling, quantization and encoding consecutively as shown in figure 2.2. In the sampling operation, sample values of measurement y at uniformly spaced discrete time instances are retained. The samples values are mapped by quantizer q into a set of finite size and then coded into binary codewords. "Source Encoder", "Channel Encoder/ Modulation", "Waveform Channel", "Noise", "Detector /Channel decoder", "Source Decoder" are of the same functions as in figure 1.3 in Haykin (1988). "Controller" maps received codewords of measurement y to digital control command u according to a certain control law. "DA Conversion" in "Source Decoder / DA Conversion" maps the reconstructed source codewords of control command u to an analog signal so that control can be applied to the physical object or process in the plant.

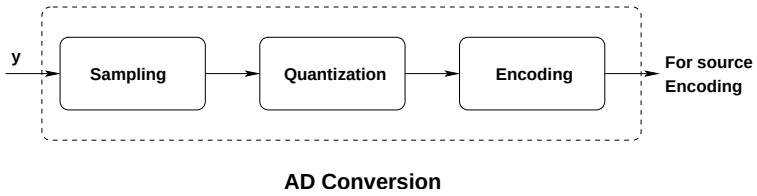


Figure 2.2: Analog-to-digital Conversion

Assumption 2.2.1. *We only consider systems with limited communication which can be modelled into figure 2.1 with the function of each element in the diagram specified as above.*

Definition 2.2.1. *Define the plant-to-controller channel to be "uplink" and controller-to-plant channel "downlink".*

Assumption 2.2.2 (Model of Plant). *The analog control signal applied to the plant remains constant within the time interval between two consecutive sampling time instances (referring to as time indices) in AD Conversion.*

The physical process or object in plant can be modelled as time invariant discrete time system,

$$x_{k+1} = f(x_k, u_k, w_k)$$

where $x_k \in \mathcal{R}^{d_x}$ is the state vector of the physical process or object at time index k, $u_k \in \mathcal{R}^{d_u}$ is the constant control during time index k and $k+1$ and $w_k \in \mathcal{R}^{d_w}$ is system noise at time index k.

The sample value of the measurement process y at time index k, y_k, can be modelled as

$$y_k = h(x_k, v_k)$$

where $h : \mathcal{R}^{d_x} \times \mathcal{R}^{d_v} \to \mathcal{R}^{d_y}$ is time invariant measurement function, $x_k \in \mathcal{R}^{d_x}$ is the state vector at time index k, $v_k \in \mathcal{R}^{d_v}$ is measurement noise at time index k and $y_k \in \mathcal{R}^{d_y}$ is the measurement at time index k.

Assumption 2.2.3 (Uplink). *The sampling rate in the sampling operation within "AD Conversion / Source Encoder" is f_s Hertz. f_s is a constant. The quantization, encoding and source encoder within "AD Conversion / Source Encoder" are time invariant and generate a binary source codeword with fixed-length l_u for each sample value y_k. Encoding in AD Conversion is a one-to-one map.*

The combination of the components of the uplink enclosed in the dashed rectangle is able to provide error free transmission for source codewords at data rate r_u bits per second.

The following relation holds true.

$$r_u = l_u f_s$$

Uplink is used in the following way: At time index k, source codeword representing y_k is generated instantaneously after the sample value y_k is obtained. It is then transmitted error-free through the components of uplink enclosed in the dashed rectangle and received at the source decoder at time index $k+1$. The source decoding is instantaneous.

Remark 2.2.1. *For a given r_u, there exists a tradeoff between sampling rate f_s and the length of a source codeword, l_u. The longer each source codeword is, the more information about each sample value will be transmitted to and received by the "Controller". However, since the overall source codeword data rate in uplink is limited from above, longer source codewords generally result in lower sampling rate. Lower sampling rate often has adverse effects on the performance of a control system. This tradeoff by itself is a research topic and should be carefully studied. We don't attempt the address this issue in this thesis.*

Assumption 2.2.4 (Controller). *The first operation for incoming source codewords in "Controller" is decoding corresponding to encoding in "AD Conversion". This operation is instantaneous. At each time index k, a causal control law maps the reconstructed quantized measurements and other information it has received up to time index k to a digital control command. The execution of the control law in the controller is instantaneous.*

Assumption 2.2.5 (Downlink). *The downlink source encoder is time invariant and generates source codeword with fixed length l_d for each control command.*

The combination of the components of the downlink enclosed in the dashed rectangle is able to provide error free transmission for source codewords at data rate r_d bits per second.

The following relation holds:

$$r_d = l_d f_s$$

The downlink is used in the following way. When a digital control command is available from controller at time index k, source codeword for the command is generated instantaneously. It is then transmitted error-free through the components of downlink enclosed in the dashed rectangle and received at the source decoder at time index $k+1$. The source decoding and DA conversion are both instantaneous. DA conversion is time invariant. The analog control signal is then applied to the physical object or process in the plant and holds constant from time index $k+1$ to $k+2$.

Proposition 2.2.1 (Mathematical Model of the System). *Based on Assumption* 2.2.2 *to* 2.2.5*, the system can be modelled as the following system of equations.*

$$\begin{aligned} x_{k+1} &= f(x_k, u_k, w_k) \\ y_k &= h(x_k, v_k) \\ z_k &= q(y_k) = q(h(x_k, v_k)) \\ \alpha_k &= F_k(E, z_{k-1}, z_{k-2}, \cdots, z_1) \\ u_k &= K(\alpha_{k-1}) \end{aligned} \tag{2.1}$$

where $q : \mathcal{R}^d \to \mathbf{Z}_{2^{l_u}} = \{1, 2, \cdots, 2^{l_u}\}$ *is the quantizer in AD Conversion,* z_k *is the quantized measurement at time* k*,* $F_k : S_E \times \mathbf{Z}_{2^{l_u}}^{k-1} \to \mathbf{Z}_{2^{l_d}} = \{1, 2, \cdots, 2^{l_d}\}$ *represents the control law at time* k*, which maps codewords* $z_{k-2}, z_{k-1}, \cdots, z_1$ *and some additional information* E *to set of digital control commands,* α_k *is the digital control command at time* k*.* $K : \mathbf{Z}_{2^{l_d}} \to \mathcal{R}^{d_u}$ *denotes the DA conversion.* u_k *is the analog control at time* k*.*

Proof. The first two equations come from Assumptions 2.2.2. As in Haykin (1988), source encoder is one-to-one map. From Assumption 2.2.3, Encoding in AD Conversion is also one-to-one map. Thus, the size of the codebook of q is the same as the cardinality of the range space of source encoder in uplink, 2^{l_u}. Choose the range space to be $\mathbf{Z}_{2^{l_u}} = \{1, 2, \cdots, 2^{l_u}\}$. The quantization of the sample value at k, y_k, can be written as

$$z_k = q(y_k) = q(h(x_k, v_k))$$

with $q : \mathcal{R}^{d_y} \to \mathbf{Z}_{2^{l_u}}$.

From Assumption 2.2.3 to 2.2.4, we see that at time k, to compute the digital control command α_k, controller has access to quantized measurement up to time $k-1$. $z_{k-1}, z_{k-2}, \cdots, z_1$. Consider some other outside information needed for control law, α_k can be written as follows:

$$\alpha_k = F_k(E, z_{k-1}, z_{k-2}, \cdots, z_1)$$

where E denotes the other information used in F_k. S_E denotes the set of all possible E. So, $F_k : S_E \times \mathbf{Z}_{2^{l_u}}^{k-1} \to \mathbf{Z}_{2^{l_d}} = \{1, 2, \cdots, 2^{l_d}\}$. Since the source encoder in the downlink is one-to-one map and the binary source codeword length in downlink is l_d, C_k's range space, which is the domain of source encoder in downlink, also has size 2^{l_d}. Choose the range space to be $\mathbf{Z}_{2^{l_d}} = \{1, 2, \cdots, 2^{l_d}\}$. From assumption 2.2.5, at time k, the DA Conversion in "Source Decoder / DA Conversion" has input α_{k-1}. So, we have

$$u_k = K(\alpha_{k-1})$$

with $K : \mathbf{Z}_{2^{l_d}} \to \mathcal{R}^{d_u}$.

■

As stated in Chapter 1, we ignore the time delay caused by limited communication. We consider the following filtering problem.

Problem 2.2.1. *Consider system*

$$\begin{aligned} x_{k+1} &= f(x_k, u_k, w_k) \\ y_k &= h(x_k, v_k) \\ z_k &= q(y_k) = q(h(x_k, v_k)) \end{aligned} \tag{2.2}$$

where x_k, u_k, w_k, v_k, f, h, q, z_k are defined as before. How to obtain the optimal filter in mean-squared-error sense, i.e.

$$E[x_k | z_k, z_{k-1}, \cdots, z_0]$$

$E[x_k|z_k, z_{k-1}, z_{k-2}]$ denotes the conditional mean of x_k conditioned on a sequence of quantized measurements $\{z_k, z_{k-1}, \cdots, z_0\}$

Further assumptions can be made about $f(x_k, u_k, w_k)$ and $h(x_k, v_k)$. Different assumptions would result in different approaches to the problem. Different assumptions as to $h(x_k, v_k)$ become the central difference between problems studied in chapter 3 and chapter 5. In chapter 3, we focus on the cases where v_k is present in $h(x_k, v_k)$. More specifically, we assume v_k enters the measurement model as an additive noise, i.e. $h(x_k, v_k) = C_k x_k + v_k$ where C_k is a matrix of proper dimensions. In chapter 5, we discuss the cases where v_k is absent in $h(x_k, v_k)$. $h(x_k, v_k)$ becomes a deterministic function of x_k. More specially, $h(x_k, v_k) = Cx_k$ where C is some constant matrix of proper dimensions.

Chapter 3

State Estimation from Quantized Measurement I : Noisy Measurement

3.1 Problem Statement

Chapter 2 concluded with mathematical models for a class of systems with limited communication and formulation of the filtering problem for those systems. This chapter focuses on a subset of that class of systems where the noise in measurement is additive before quantization. More specifically, we consider the following filtering problem.

Problem 3.1.1. *Consider system*

$$\begin{aligned} x_{k+1} &= f(x_k, u_k, w_k) \\ y_k &= C_k x_k + v_k \\ z_k &= q(y_k) = q(C_k x_k + v_k) \end{aligned} \tag{3.1}$$

where x_k, u_k, w_k, v_k, f, h, q, z_k are defined as in chapter 2. How to obtain the optimal filter in mean-squared-error sense, i.e.

$$E[x_k | z_k, z_{k-1}, \cdots, z_0]$$

$E[x_k|z_k, z_{k-1}, z_{k-2}]$ denotes the conditional mean of x_k conditioned on a sequence of quantized measurements $\{z_k, z_{k-1}, \cdots, z_0\}$

This filtering problem falls into the category of nonlinear filtering. This problem is unique among other nonlinear filtering problems in the sense that it has specific structure in its measurement model, which is the presence of quantization.

3.2 Kalman Filter and Particle Filter

Kalman filter is a computationally efficient recursive least-squares state estimator for a linear state space model when the system and observation noises in the model are white and uncorrelated with each other.

Consistent efforts in seeking efficient least-squares state estimator for general nonlinear state space model has been made after Kalman filter theory was established. Conceptually, at time index k, sequential state estimation for a discrete time state space model consists of two steps, Projection and Bayesian Update. In the projection step, the distribution of state vector at time $k-1$, x_{k-1}, is propagated through the system evolution equation to obtain a prior distribution of x_k. Bayesian Update uses well-known Bayesian formula to update the prior distribution of x_k based on measurement y_k at time k which is not independent with x_k to obtain a posterior distribution of x_k. The mean of the a posterior distribution of x_k will be the least-squares estimator for x_k given the distribution of x_{k-1} and measurement y_k. However, this simple framework itself does not provide a computationally efficient way for realization. In fact, an efficient realization has eluded the nonlinear filtering theory for a long time. Several numerical attempts, such as adaptive grid generation in state space, suffer from various serious drawbacks. One of those troubles comes from when the dimension of the system increases, the nodes in the grid used increases exponentially. It causes not only exponential increase in computational load but also in significant error accumulation which is very difficult to quantify.

Several ad hoc methods inspired by Kalman filter have been studied and used extensively in the past. Extended Kalman filter attracted the most intention**?**. Assuming the following nonlinear state space model:

$$\begin{aligned} x_{k+1} &= f_k(x_k) + g_k(x_k)w_k \\ y_k &= h_k(x_k) + v_k \end{aligned} \tag{3.2}$$

where x_k is the state vector, y_k is the measurement, f_k is the system function, g_k is the noise coefficient function, w_k is system noise, h_k is the measurement function and v_k is the measurement noise.

Extended Kalman filter linearizes the nonlinear state space model around state estimate in previous step and apply Kalman filter to that linearized model in an ad hoc way. A variety of other nonlinear filtering methods are derived based on extended Kalman filter, such as second order extended Kalman filter, iterated extended Kalman filter and Gaussian sum filter. They are easy to implement and have been demonstrated to be successful for certain applications. Many qualitative analysis and judgments have been reached as to the performance of these methods. Essentially, they remain ad hoc and don't provide a theoretically solid approach for general nonlinear filtering problem. They don't consider all the information of the system and noises and, thereby, often lead to poor results.

Last three decades witnessed extraordinary increases in computational power. Approximately summarized by Moore's Law in Moore (1965), the phenomenal advances in computational power makes computation-intensive Monte Carlo method increasingly more appealing for applications. Study of the limit behaviors of Monte Carlo method as computation goes to infinity becomes a subject of not only theoretical interests but also practical impact. As the application of Monte Carlo method to sequential state estimation problem, sequential Monte Carlo(SMC) method (also known as Particle filter) was first introduced by Gordon et al. (1993) as a sample based method for probability distribution propagation and Bayesian update in nonlinear filtering problems. Being successfully applied to numerous application areas as evidenced in Doucet et al. (2001), **? ?**, it has promise to be the next milestone in the history of sequential state estimation after Kalman filter.

As all other Monte Carlo methods, central idea in SMC is about generating and managing a group of samples. Instead of propagating certain statistics of target distribution as in EKF, it tries to manage a group of samples of state vector in such a way that they closely and accurately follow the probability distribution propagation and Bayiesan update in sequential state estimation. Many theoretical studies on properties of SMC have been carried out. However, one of the major reasons that SMC has yet claimed the same fame as that of Kalman filter lies behind the lack of deep understanding of behaviors of SMC. More specifically, the conditions under which SMC will have more desired properties than the ones which have been revealed now and how to quantitatively predict its performance for a given system are still open problems.

Based on the overview of nonlinear filtering problem above, we present two basic approaches to the filtering problem 3.1.1. The first one is to treat the quantization as an additive noise. After defining an one-to-one map which maps range space of quantizer q to a set of real vectors, we define the quantization noise process and make certain assump-

tions as to the statistical properties of this artificially made up additive noise. Kalman filter and extended Kalman filter can be easily modified to incorporate this new item in the measurement model.

The second one is to apply SMC. Since measurement model of classical SMC as in Crisan (2001) is incompatible with the measurement model in Problem 3.1.1, we propose QMSMC as the modified SMC for this specific problem. It will become obvious that applying QMSMC to filtering problem 3.1.1 is not only possible but natural. We will prove the asymptotical optimality of QMSMC which is one of the most desired properties of this algorithm for this problem. We will compare the effectivenss of these two approaches in next chapter by extensive simulation.

3.3 Quantized Measurement Kalman Filter (Extend Kalman Filter)

Recall the system 3.1

$$x_{k+1} = f_k(x_k) + g_k(x_k)w_k$$
$$y_k = q(h_k(x_k) + v_k)$$

3.3.1 Additive Quantization Noise Model

$$y_k = q(h_k(x_k) + v_k)$$

In order to be able to treat quantization as additive noise, we first define inverse mapping, quantization noise function and quantization noise sequence.

Definition 3.3.1 (Inverse mapping and quantization function). *Consider quantizer* $y = q(x)$ *where* $q : \mathcal{R}^d \to \mathbf{Z_M} = \{0, 1, 2, \ldots, \mathbf{M} - 1\}$. *Define an inverse mapping for* q *to be a bijection between* $\mathbf{Z_M}$ *and* R.

$$i : \mathbf{Z_M} \to R$$

with $R = \{r_0, r_1, r_2, \cdots, r_{\mathbf{M}-1}\}$, $r_i \in \mathcal{R}^d$ *for* $i = 0, 1, 2, \cdots, \mathbf{M} - 1$.

Quantization function h *is defined as*

$$h \stackrel{\Delta}{=} i \circ q$$

Definition 3.3.2 (quantization noise function). *For quantization function $h : \mathcal{R}^d \to \mathcal{R}^d$, quantization noise function $n(x)$ is defined to be*

$$n(x) = h(x) - x$$

Definition 3.3.1 can be illustrated as in figure 3.1.

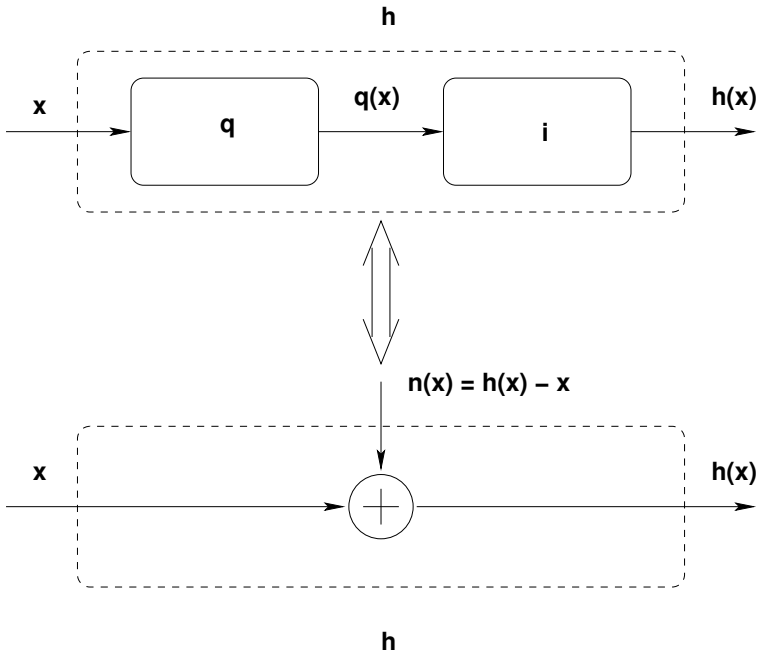


Figure 3.1: Additive noise model of a quantizer

Definition 3.3.3 (quantization noise sequence). *Given quantization noise function $n(x)$ for a quantizer q. For a random sequence $\{x_i\}$, $x_i \in \mathcal{R}$, $i = 1, 2, \cdots, n$, the sequence $\{n(x_i)\}$ is called $\{x_i\}$'s quantization noise sequence.*

Assumption 3.3.1. *In this thesis, if variable in quantizer x has a known distribution, inverse mapping i is chosen to be mapping index to the conditional mean of x conditioned on that x lies in the quantization region with that index.*

Assumption 3.3.2. *In system 3.1, we assume that with quantization noise sequence $n(h_c(x_k)+ v_k)$ is white and uncorrelated with underlying processs $\{h_c(x_k)\}$ and $\{v_k\}$ and system noise process $\{w_k\}$.*

Remark 3.3.1. *Assumption 3.3.1 is almost never true. To which extent assumption 3.3.1 is true will partially affect to which extent Quantized measurement Kalman filter will be Least-squares estimator or Quantized measurement Extended Kalman filter will be effective for nonlinear filtering.*

System 3.1 can be rewritten as

$$\begin{aligned} x_{k+1} &= f_k(x_k) + g_k(x_k)w_k \\ y_k &= h_k(x_k) + v_k + n(h_k(x_k) + v_k) \end{aligned} \tag{3.3}$$

Under assumption 3.3.1, we regard $v_k + n(h_k(x_k) + v_k)$ as additional measurement noise which is white and uncorrelated with $h_k(x_k)$ and v_k. Viewed in this way, system 3.3 is essentially the same as system 3.2. No significant difficulties exist in applying Kalman filter for linear state space model and extended Kalman filter for nonlinear state space model.

3.3.2 Quantized Measurement Kalman Filter

Quantized measurement Kalman Filter is concerned with the following linear state space model:

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k w_k + G_k u_k \\ y_k &= q(C_k x_k + v_k) \\ z_k &= C_k x_k + v_k + n(C_k x_k + v_k) \end{aligned} \tag{3.4}$$

where A_k, B_k and C_k are matrices with proper dimensions. System noise w_k and observation noise v_k are both white gaussian process with known covariance and mean for each time index k.

From Assumption 3.3.1, to apply Kalman filter, we need to know the variance of $n(C_k x_k + v_k)$ at each time step. Variance of $n(C_k x_k + v_k)$ depends on quantizer q, inverse mapping i and the distribution of $C_k x_k + v_k$ at time index k. Suppose the distribution of $o_k = C_k x_k + v_k$ is continuous and has probability density function f.

Suppose quantizer q is given, then, to minimize variance of $n(C_k x_k + v_k)$

Assume the quantizer to be time variant. At each time index, the conditional variance and mean can be represented as.

QMKF is stated as below: Prediction, Update.

Algorithm 3.3.1. *1.* **Prediction :**

$$\hat{x}_{k/k-1} = A_{k-1}\hat{x}_{k-1/k-1} + G_{k-1}u_{k-1}$$

$$\Sigma_{k/k-1} = A_{k-1}\Sigma_{k-1/k-1}A_{k-1}^T + B_{k-1}Q_{k-1}B_{k-1}^T$$

2. **Update :**

$$K_k = \Sigma_{k/k-1}C_k^T(C_k\Sigma_{k/k-1}C_k^T + R_k + R_q(y_k))^{-1}$$

$$\Sigma_{k/k} = \Sigma_{k/k-1} - K_kC_k\Sigma_{k/k-1}$$

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k(z_k - C_k\hat{x}_{k/k-1})$$

where $R_q(y_k)$ is the conditional variance of $n(C_kx_k+v_k)$ conditioned on $q(C_kx_k+v_k) = y_k$.

This algorithm is different from Kalman filter only in the step involving observation noise. The variance of quantization noise is chosen to be the conditional variance conditioned on the quantization index y_k. This generally will reduce the variance of the quantization and result in reduced $\Sigma_{k/k}$. However, whether the reduce $\Sigma_{k/k}$ will result in reduced estimation error is unclear since the effect of assumption 3.3.1 on validity of assumption 3.3.2 is not studied.

3.3.3 Quantized Measurement Extended Kalman Filter

Quantized measurement Kalman Filter is concerned with the following linear state space model:

$$\begin{aligned} x_{k+1} &= f_k(x_k) + g_k(x_k)w_k + r_k(u_k) \\ y_k &= q(h_k(x_k) + v_k) \\ z_k &= C_kx_k + v_k + n(C_kx_k + v_k) \end{aligned} \tag{3.5}$$

where f_k, g_k, h_k and u_k are vectors with proper dimensions. System noise w_k and observation noise v_k are both white gaussian process with known covariance and mean for each time index k.

From Assumption 3.3.1, to apply Kalman filter, we need to know the variance of $n(C_kx_k + v_k)$ at each time step. Variance of $n(C_kx_k + v_k)$ depends on quantizer q, inverse mapping i and the distribution of $C_kx_k + v_k$ at time index k. Suppose the distribution of $o_k = C_kx_k + v_k$ is continuous and has probability density function f.

Suppose quantizer q is given, then, to minimize variance of $n(C_k x_k + v_k)$

Assume the quantizer to be time variant. At each time index, the conditional variance and mean can be represented as.

QMKF is stated as below: Prediction, Update.

Algorithm 3.3.2. *1.* **Prediction :**

$$\hat{x}_{k/k-1} = A_{k-1}\hat{x}_{k-1/k-1} + G_{k-1}u_{k-1}$$

$$\Sigma_{k/k-1} = A_{k-1}\Sigma_{k-1/k-1}A_{k-1}^T + B_{k-1}Q_{k-1}B_{k-1}^T$$

2. **Update :**

$$K_k = \Sigma_{k/k-1}C_k^T(C_k\Sigma_{k/k-1}C_k^T + R_k + R_q(y_k))^{-1}$$

$$\Sigma_{k/k} = \Sigma_{k/k-1} - K_k C_k \Sigma_{k/k-1}$$

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k(z_k - C_k\hat{x}_{k/k-1})$$

where $R_q(y_k)$ is the conditional variance of $n(C_k x_k + v_k)$ conditioned on $q(C_k x_k + v_k) = y_k$.

3.4 Sequential Monte Carlo Method

For this class of models, except for trivial ones, optimal filter in LMSE sense doesn't admit a closed-form expression. We have to seek effective approximation.

Attracting significant attention in the literature in recent years Gordon et al. (1993), Liu and Chen (1998), Crisan (2001), Crisan and Doucet (2002), classical Sequential Monte Carlo(SMC) methods have shown their versatility in solving a wide class of nonlinear filtering problem through theoretical justification and numerous successful applicationsDoucet et al. (2001). classical Sequential Monte Carlo(SMC) consider the system as follows:

$$\begin{aligned} x_{k+1} &= f(k, x_k, u_k, w_k) \\ y_k &= h(k, x_k) + v_k \end{aligned} \tag{3.6}$$

Loosely interpreting, SMC methods recursively maintain a group of samples which approximately can be viewed as a group of identical, independent samples from the posterior distribution of state given all the measurements up to date. It aims to numerically

approximate the optimal filter which often is analytically intractable with manageable computational load. For certain models, it has been shown that classical SMC methods are asymptotically optimal in the number of particles usedCrisan (2001).

System model in (3.6) is more general than system model in Problem 2.2.1. This difference is insignificant from the derivation of the proof of convergence of classical SMC. Observation model in (3.6) is different from that in problem 2.2.1. The quantization part can be viewed as an additive noise term which will transform Problem 2.2.1 into system 3.6. However, from the theoretical foundation of SMC, this treatment is unnecessary and unnatural. A revised SMC, named QMSMC will be presented in the following.

3.4.1 Problem Formulation

Notation

N denotes the set of natural numbers. Z^+ denotes $0 \cup N$. $Z_{\mathbf{M}}$ denotes $\{1, 2, \cdots \mathbf{M}\}$. $\mathcal{R}$ denotes the set of real numbers. $\mathcal{R}^+$ denotes the set $\{x | x \in \mathcal{R}, x \geq 0\}$. $\mathcal{R}^d$ denotes d-dimensional Euclidean Space. $\mathcal{B}(\mathcal{R}^d)$ denotes Borel σ-algebra on $\mathcal{R}^d$. $\mu : \mathcal{B}(\mathcal{R}^d) \mapsto \mathcal{R}^+$ denotes Borel measure on $\mathcal{R}^d$(we only consider Borel measure in this paper). $\mathcal{M}_F(\mathcal{R}^d)$ denotes the set of all finite Borel measures on $\mathcal{R}^d$. $\mathcal{P}(\mathcal{R}^d)$ denotes the set of all probability measures on $\mathcal{R}^d$. $C_b(\mathcal{R}^d)$ denotes the set of bounded continuous functions on $\mathcal{R}^d$. μf denotes $\int_{\mathcal{R}^d} f(x)\mu(dx)$ for $\mu \in \mathcal{M}_F(\mathcal{R}^d)$ and $f \in C_b(\mathcal{R}^d)$. δ_x denotes probability measure degenerate at point x.

Formulation

We consider discrete-time state-space model with quantized measurement which can be modelled as below.

$$\begin{aligned} x_k &= f(x_{k-1}, u_{k-1}, w_{k-1}) \\ y_k &= q(C_k x_k + v_k) \end{aligned} \tag{3.7}$$

In (1), $k \in Z^+$ for x_k. $k \in N$ for y_k. f denotes the time invariant system function. $\{w_k\}_{k=0}^{\infty}$ are system noises and $\{v_k\}_{k=1}^{\infty}$ measurement noises. $\{u_k\}_{k=1}^{\infty}$ is the control sequence. $\{C_k\}_{k=1}^{\infty}$ are $d_o \times d_x$ measurement matrices where d_x denotes the dimension of x_k and d_o denotes the dimension of v_k. q denotes the quantizer in measurement. $q : \mathcal{R}^{d_o} \mapsto Z_{\mathbf{M}}$, in which $\mathbf{M}$ denotes the size of the codebook of quantizer q. From the definition of resolution of a quantizer, $\mathbf{M} = d_o \cdot 2^r$ where r is the resolution of quantizer q.

We assume the following in the rest of the paper.

Assumption 3.4.1.

1. *$f : \mathcal{R}^{d_x} \times \mathcal{R}^{d_u} \times \mathcal{R}^{d_w} \to \mathcal{R}^{d_x}$ is measurable. d_u denotes the dimension of u_k and d_w the dimension of w_k.*

2. *Quantizer q satisfies the following constraints.*

   (a) *Cells $Q_i = \{x | x \in \mathcal{R}^{d_o}, q(x) = i \in Z_{\mathbf{M}}\}$ for $i = 1, \cdots, \mathbf{M}$ are disjoint subsets of $\mathcal{R}^{d_o}$. $\bigcup_{i=1}^{\mathbf{M}} Q_i = \mathcal{R}^{d_o}$.*

   (b) *$Q_i = \{x | x \in \mathcal{R}^{d_o}, F_{i1}x < b_{i1}, F_{i2}x \leq b_{i2}\}$ where matrices F_{i1}, F_{i2} and vectors b_{i1}, b_{i2} are of proper and finite dimensions. Each of F_{i1} and F_{i2}'s rows has at least one nonzero element.*

   (c) *$m^{d_o}(Q_i) > 0$ where m^{d_o} denotes the d_o dimensional lebesgue measure.*

3. *System noises $\{w_k\}_{k=0}^{\infty}$ and measurement noises $\{v_k\}_{k=1}^{\infty}$ are white stochastic processes, which are independent of each other and are of known distributions.*

4. *Joint probability density function of v_k, $p(v_k)$, is positive and continuous on $\mathcal{R}^{d_o}$ for all $k \geq 1$.*

5. *Independent samples from the distributions of x_0 and $\{w_k\}_{k=0}^{\infty}$ can be obtained.*

In Assumption 1, the constraint regarding the quantizer is satisfied by a large class of quantizers on $\mathcal{R}^{d_o}$, such as polytopal vector quantizersGersho and Gray (1992). From the constraint, it is clear that Q_i's boundary, $\partial Q_i = \bar{Q}_i - Q_i^o$, consists of a finite union of subsets of $d_o - 1$ dimensional hyperplanes. $\bar{Q}_i$ and Q_i^o denote closure and interior of Q_i respectively.

Problem 3.4.1 (Filtering using quantized measurement).

For system specified in (1)*, given the distribution of x_0, a sequence $\{u_k\}_{k=0}^{K-1}$, a sequence of measurements $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K) \in (Z_{\mathbf{M}})^K$ and $\forall h \in C_b(\mathcal{R}^{d_x})$, optimal filtering is to obtain*

$$E[h(x_K) | \{y_i\}_{i=1}^{K} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)]$$

i.e. the conditional expectation of $h(x_K)$ with respect to the event $\{y_i\}_{i=1}^{K} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)$ which is the optimal estimation of $h(x_K)$ in LMSE sense.

3.4.2 QMSMC Method

We first define linear combination of measures.

Definition 3.4.1 (Linear Combination of Measures). *For* $\mu_1, \mu_2 \in \mathcal{M}_F(\mathcal{R}^d)$, $\alpha, \beta \in \mathcal{R}^+$, $\forall E \in \mathcal{B}(\mathcal{R}^d)$

$$(\alpha\mu_1 + \beta\mu_2)(E) = \alpha\mu_1(E) + \beta\mu_2(E)$$

From the definition of finite measure, it is obvious that $(\alpha\mu_1 + \beta\mu_2) \in \mathcal{M}_F(\mathcal{R}^d)$. Furthermore, if $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{R}^d)$ and $\alpha + \beta = 1$, $(\alpha\mu_1 + \beta\mu_2) \in \mathcal{P}(\mathcal{R}^d)$.

We present the QMSMC method in a way more mathematical than necessary for implementation. The purpose of this is to make the proof of its asymptotical optimality easier to state.

The QMSMC method for Problem 1 proceeds as follows:

1. **Initiation :**

   Draw a set of n independent samples from the distribution of x_0, denoted by $\{\mathbf{x}_0^i\}_{i=1}^n$. Form discrete distribution π_0^n from those samples.

   $$\pi_0^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_0^i}$$

   Set $k = 1$.

2. **Iteration :**

   At time k, the algorithm uses the quantized measurement at k, $\mathbf{y}_k$ to update $\{\mathbf{x}_{k-1}^i\}_{i=1}^n$ into $\{\mathbf{x}_k^i\}_{i=1}^n$.

   (a) Prediction :

   Obtain n independent samples $\{\mathbf{w}_{k-1}^i\}_{i=1}^n$ from the distribution of w_{k-1}. For $1 \le i \le n$, get the i-th prior sample $\bar{\mathbf{x}}_k^i$ by

   $$\bar{\mathbf{x}}_k^i = f(\mathbf{x}_{k-1}^i, u_{k-1}, \mathbf{w}_{k-1}^i)$$

   Form discrete distribution ρ_k^n from prior samples $\{\bar{\mathbf{x}}_k^i\}_{i=1}^n$.

   $$\rho_k^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\bar{\mathbf{x}}_k^i}$$

   (b) Evaluation of likelihoods :

   Evaluate the likelihoods of prior samples $\{\bar{\mathbf{x}}_k^i\}_{i=1}^n$ based on measurement $\mathbf{y}_k$ in

the following way. In the formulas below, $l_k^{\mathbf{y}_k}(\bar{\mathbf{x}}_k^i)$ denotes the likelihood of the prior sample $\bar{\mathbf{x}}_k^i$ based on $\{y_k = \mathbf{y}_k\}$.

Define $G(\mathbf{y}_k)$ to be $Q_{\mathbf{y}_k}$. Define $G(\mathbf{y}_k) - C_k\bar{\mathbf{x}}_k^i$ to be

$$G(\mathbf{y}_k) - C_k\bar{\mathbf{x}}_k^i \triangleq \{x | x \in \mathcal{R}^{d_o}, x + C_k\bar{\mathbf{x}}_k^i \in G(\mathbf{y}_k)\}$$

Then, $l_k^{\mathbf{y}_k}(\bar{\mathbf{x}}_k^i)$ is evaluated by integrating $p(v_k)$ on the region $G(\mathbf{y}_k) - C_k\bar{\mathbf{x}}_k^i$

$$l_k^{\mathbf{y}_k}(\bar{\mathbf{x}}_k^i) = \int_{G(\mathbf{y}_k) - C_k\bar{\mathbf{x}}_k^{(i)}} p(v_k)dv_k \tag{3.8}$$

After evaluating likelihoods of all prior samples $\{\bar{\mathbf{x}}_k^i\}_{i=1}^n$ based on measurement $\{y_k = \mathbf{y}_k\}$, normalize the likelihoods as

$$u_k^i = \frac{l_k^{\mathbf{y}_k}(\bar{\mathbf{x}}_k^i)}{\sum_{i=1}^n l_k^{\mathbf{y}_k}(\bar{\mathbf{x}}_k^i)}$$

(c) Estimation :

The estimated $\hat{h}(x_k)$ is obtained by

$$\hat{h}(x_k) = \sum_{i=1}^n u_k^i h(\bar{\mathbf{x}}_k^i)$$

(d) Resampling :

Form discrete distribution $\bar{\pi}_k^n$ from prior samples $\{\bar{\mathbf{x}}_k^i\}_{i=1}^n$ weighted by their respective normalized likelihoods.

$$\bar{\pi}_k^n = \sum_{i=1}^n u_k^i \delta_{\bar{\mathbf{x}}_k^i}$$

Draw n independent samples $\{\mathbf{x}_k^i\}_{i=1}^n$ from the distribution $\bar{\pi}_k^n$ as follows.

i. Form accumulative normalized likelihoods vector $U = \{U_k^j\}_{j=0}^n$. $U_k^0 = 0$ and for $1 \leq j \leq n$, $U_k^j = \sum_{i=1}^j u_k^i$.

ii. Draw n independent samples from random variable a uniformly distributed over $[0, 1)$. Denote them by $\{\mathbf{a}_i\}_{i=1}^n$. For each i, find the index b_i such that $U_k^{b_i - 1} \leq \mathbf{a}_i < U_k^{b_i}$. Set $\mathbf{x}_k^i = \bar{\mathbf{x}}_k^{b_i}$

(e) Form discrete distribution π_k^n from $\mathbf{x}_k^i$.

$$\pi_k^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_k^i}$$

(f) Set $k = k + 1$. If $k \leq K$, go to (a).

3. **Estimation :**

$$E[h(x_K)|\{y_i\}_{i=1}^K = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)] = \hat{h}(x_K)$$

3.4.3 Random Discrete Probability Measure

We introduce two definitions in this subsection. The meaning of asymptotical optimality relies on Definition 3.

Definition 3.4.2 (Random discrete probability measure). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the probability space. $\{u_i\}_{i=1}^n$ are random variables satisfying the following constraints: $0 \leq u_i \leq 1$ for $i = 1, \cdots, n$ and $\sum_{i=1}^n u_i = 1$. $\{z_i\}_{i=1}^n$ are $\mathcal{R}^d$-valued random variables. A random discrete probability measure on $\mathcal{R}^d$, ν, is a function from Ω to $\mathcal{P}(\mathcal{R}^d)$ such that*

$$\nu = \sum_{i=1}^{n} u_i \delta_{z_i}$$

For a random discrete probability measure $\nu = \sum_{i=1}^n u_i \delta_{z_i}$ and $\forall g \in C_b(\mathcal{R}^d)$, easy to see that $\nu g = \sum_{i=1}^n u_i g(z_i)$ is a random variable.

Definition 3.4.3 (Convergence of random discrete probability measures). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the probability space and let $\{\nu_n : \Omega \to \mathcal{P}(\mathcal{R}^d)\}_{n=1}^\infty$ be a sequence of random discrete probability measures on $\mathcal{R}^d$ and ν a probability measure on $\mathcal{R}^d$. $\{\nu_n\}_{n=1}^\infty$ converges to ν in $\mathcal{L}^p$ sense, if for $\forall g \in C_b(\mathcal{R}^d)$*

$$\lim_{n\to\infty} ||\nu_n g - \nu g||_{\mathcal{L}^p} = 0$$

We abbreviate $\lim_{n\to\infty} ||\nu_n g - \nu g||_{\mathcal{L}^p} = 0$ as $\nu_n g \xrightarrow{\mathcal{L}^p} \nu g$. $||\cdot||_{\mathcal{L}^p}$ denotes $\mathcal{L}^p$ norm of a random variable.

3.4.4 Proof of Asymptotical Optimality of QMSMC Method

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the probability space on which stochastic processes $\{x_k\}_{k=0}^{\infty}$ and $\{y_k\}_{k=1}^{\infty}$ are defined. Define $y_{1:k}$ to be $\{y_i\}_{i=1}^{k}$. Denote by $\mathbf{y}_{1:k}$ a given sample path $(\mathbf{y}_1, \ldots, \mathbf{y}_k)$ of $y_{1:k}$.

Define conditional probability measure $\pi_k^{\mathbf{y}_{1:k}}$ for $k \geq 1$ to be $\pi_k^{\mathbf{y}_{1:k}}(A) = \mathbf{P}(x_k \in A | y_{1:k} = \mathbf{y}_{1:k})$ for $\forall A \in \mathcal{B}(\mathcal{R}^{d_x})$. π_0 is defined to be the known distribution of x_0. Define the predicted conditional probability measure $p_k^{\mathbf{y}_{1:k-1}}$ for $k \geq 2$ to be $p_k^{\mathbf{y}_{1:k-1}}(A) = \mathbf{P}(x_k \in A | y_{1:k-1} = \mathbf{y}_{1:k-1})$ for $\forall A \in \mathcal{B}(\mathcal{R}^{d_x})$. Define p_1 to be $p_1(A) = \mathbf{P}(x_1 \in A)$ for $\forall A \in \mathcal{B}(\mathcal{R}^{d_x})$.

We notice that for a given control sequence $\{u_k\}_{k=0}^{\infty}$, the stochastic process $\{x_k\}_{k=0}^{\infty}$ defined in system (1) is a Markov process. Additionally, we assume that for the control sequence $\{u_k\}_{k=0}^{\infty}$, the transition kernel of process $\{x_k\}_{k=0}^{\infty}$ has Feller property.

Definition of Feller property can be recalled.

The transition kernel of a Markov process $\{x_k\}_{k=0}^{\infty}$ is the sequence of functions $\{K_k(x, A) : \mathcal{R}^{d_x} \times \mathcal{B}(\mathcal{R}^{d_x}) \mapsto \mathcal{R}^+\}_{k=0}^{\infty}$ such that for all $A \in \mathcal{B}(\mathcal{R}^{d_x})$ and $x \in \mathcal{R}^{d_x}$, $K_k(x, A) = \mathbf{P}(X_{k+1} \in A | X_k = x)$.

We begin with three lemmas which are essential for our proof of asymptotical optimality of QMSMC method.

Lemma 3.4.1. *$l_k^{\mathbf{y}_k}(x) : \mathcal{R}^{d_x} \to \mathcal{R}^+$ defined in algorithm 1 is continuous, positive and bounded for $\forall \mathbf{y}_k$ and $\forall k \in N$.*

Proof. For $\forall \mathbf{y}_k$ and $\forall k \in N$, easy to see $|l_k^{\mathbf{y}_k}(x)| \leq 1$ for all x. So, $l_k^{\mathbf{y}_k}(x)$ is bounded.

To prove it is continuous, take $x_i \to x \in \mathcal{R}^{d_x}$. Obviously, $C_k x_i \to C_k x$.

$$
\begin{aligned}
|l_k^{\mathbf{y}_k}(x_i) - l_k^{\mathbf{y}_k}(x)| &= |\int_{G(\mathbf{y}_k) - C_k x_i} p(v_k)dv_k - \int_{G(\mathbf{y}_k) - C_k x} p(v_k)dv_k| \\
&= |\int_{G(\mathbf{y}_k) - C_k x_i \setminus G(\mathbf{y}_k) - C_k x} p(v_k)dv_k \\
&\quad - \int_{G(\mathbf{y}_k) - C_k x \setminus G(\mathbf{y}_k) - C_k x_i} p(v_k)dv_k| \\
&\leq |\int_{G(\mathbf{y}_k) - C_k x_i \setminus G(\mathbf{y}_k) - C_k x} p(v_k)dv_k| + \\
&\quad |\int_{G(\mathbf{y}_k) - C_k x \setminus G(\mathbf{y}_k) - C_k x_i} p(v_k)dv_k| \\
&= \int_{\{G(\mathbf{y}_k) - C_k x_i \setminus G(\mathbf{y}_k) - C_k x\} \cup \{G(\mathbf{y}_k) - C_k x \setminus G(\mathbf{y}_k) - C_k x_i\}} p(v_k)dv_k
\end{aligned}
$$

$$= \int_{G(\mathbf{y}_k)-C_k x_i \Delta G(\mathbf{y}_k)-C_k x} p(v_k)dv_k$$
$$= \mu_k(G(\mathbf{y}_k) - C_k x_i \Delta G(\mathbf{y}_k) - C_k x)$$

Δ denotes the symmetric difference between two sets. μ_k denotes the distribution of v_k. In reasoning above, we used as a probability density function, $p(v_k) \geq 0$ for $\forall v_k$. From the ass 1, we see that the boundary of $G(\mathbf{y}_k)$, $\partial G(\mathbf{y}_k)$, consists of a finite union of subsets of $d_o - 1$ dimensional hyperplanes in $\mathcal{R}^{d_o}$. Let $H_i = \{x | x \in \mathcal{R}^{d_o}, \exists a \in [0,1], x + [aC_k x_i + (1-a)C_k x] \in \partial G(\mathbf{y}_k)\}$. We show that $G(\mathbf{y}_k) - C_k x_i \Delta G(\mathbf{y}_k) - C_k x \subset H_i$.

To see this, we notice $\partial G(\mathbf{y}_k)$ separates $\mathcal{R}^{d_o}$ into two disjoint open subsets. Denote them by S_1 and S_2. $\mathcal{R}^{d_o} \setminus \partial G(\mathbf{y}_k) = S_1 \cup S_2$. For $a_1 \in S_1$ and $a_2 \in S_2$, a path $p_c = \{x | x = ba_1 + (1-b)a_2, 0 \leq b \leq 1\}$ joins a_1 and a_2. Easy to see p_c is path connected and then connected. Suppose p_c doesn't intersect with $\partial G(\mathbf{y}_k)$. Then, $p_c = (p_c \cap S_1) \cup (p_c \cap S_2)$. Since S_1 and S_2 are disjoint open sets, $(p_c \cap S_1)$ and $(p_c \cap S_2)$ are disjoint and open in relative topology. However, p_c is connected, it can not be expressed as the union of two nonempty disjoint open sets. So, one of them must be empty set which is contradictory to the fact that $a_1 \in (p_c \cap S_1)$ and $a_2 \in (p_c \cap S_2)$. So, p_c intersects with $\partial G(\mathbf{y}_k)$, i.e. $p_c \cap \partial G(\mathbf{y}_k)$ is nonempty.

For $\forall o \in G(\mathbf{y}_k) - C_k x_i \Delta G(\mathbf{y}_k) - C_k x$, suppose $o \in G(\mathbf{y}_k) - C_k x_i$ but $o \notin G(\mathbf{y}_k) - C_k x$. Then, $o + (C_k x_i - C_k x) \in G(\mathbf{y}_k) - C_k x$. If either o or $o + (C_k x_i - C_k x) \in \partial G(\mathbf{y}_k) - C_k x$, $o + \alpha(C_k x_i - C_k x) \in \partial G(\mathbf{y}_k) - C_k x$ for $\alpha = 0$ or $\alpha = 1$. If neither of them $\in \partial G(\mathbf{y}_k) - C_k x$, from reasoning in the paragraph above, the path $p_c = \{x | x = bo + (1-b)(o + (C_k x_i - C_k x)) = o + (1-b)(C_k x_i - C_k x), 0 \leq b \leq 1\}$ will intersect with the boundary $G(\mathbf{y}_k) - C_k x$. So, $o + \alpha(C_k x_i - C_k x) \in \partial G(\mathbf{y}_k) - C_k x$ for some $\alpha \in [0,1]$. That means $o + [\alpha C_k x_i + (1-\alpha)C_k x] \in \partial G(\mathbf{y}_k)$. Following the same reasoning, for $o \notin G(\mathbf{y}_k) - C_k x_i$ but $o \in G(\mathbf{y}_k) - C_k x$, we also can get $o + [\alpha C_k x_i + (1-\alpha)C_k x] \in \partial G(\mathbf{y}_k)$ for some $\alpha \in [0,1]$. So, $G(\mathbf{y}_k) - C_k x_i \Delta G(\mathbf{y}_k) - C_k x \subset H_i$.

Let $E_n = \bigcup_{i=n}^{\infty} H_i$. Clearly, $E_i \supset E_{i+1}$ for $i \geq 1$. $\partial G(\mathbf{y}_k) - C_k x \subset H_i \subset E_i$ for $\forall i$. For any point $o \notin \partial G(\mathbf{y}_k) - C_k x$. Since $C_k x_i \to C_k x$, we always can find $M \in N$ such that for $i \geq M$, $o \notin E_i$. So, $\bigcap_{i=1}^{\infty} E_i = \partial G(\mathbf{y}_k) - C_k x$.

Since $\mu_k(E_1) \leq 1 < \infty$, from theorem 1.8 in Folland (1999), $\lim_{i\to\infty} \mu_k(E_i) = \mu_k(\partial G(\mathbf{y}_k) - C_k x)$. d_o dimensional Lebesgue measure is translation invariant, so $m^{d_o}(\partial G(\mathbf{y}_k) - C_k x) = m^{d_o}(\partial G(\mathbf{y}_k))$. Since $\partial G(\mathbf{y}_k)$ consists of a finite union of subsets of $d_o - 1$ dimensional hyperplanes, $m^{d_o}(\partial G(\mathbf{y}_k)) = 0$. Easy to see that μ_k is absolute continuous with respect to m_{d_o}, so, $\lim_{i\to\infty} \mu_k(E_i) = \mu_k(\partial G(\mathbf{y}_k) - C_k x) = 0$. $0 \leq \mu_k(G(\mathbf{y}_k) - C_k x_i \Delta G(\mathbf{y}_k) - C_k x) \leq$

$\mu_k(H_i) \leq \mu_k(E_i)$. So, $\mu_k(G(\mathbf{y}_k) - C_k x_i \Delta G(\mathbf{y}_k) - C_k x) \to 0$ as $x_i \to x$. This proves the continuity of $l_k^{\mathbf{y}_k}(x)$.

Since $m^{d_o}(G(\mathbf{y}_k) - C_k x) = m^{d_o}(G(\mathbf{y}_k)) > 0$, we can choose a compact set $A \subset G(\mathbf{y}_k) - C_k x$ with $m^{d_o}(A) > 0$. Since $p(v_k)$ is positive and continuous on $\mathcal{R}^{d_o}$, $p(v_k)$ has a positive minimum on A. Denote it by δ. Then, $\int_{G(\mathbf{y}_k)-C_k x} p(v_k)dv_k \geq \int_A p(v_k)dv_k \geq \delta m^{d_o}(A) > 0$. This holds true for $\forall k$, $\mathbf{y}_k$ and x. ∎

The second lemma is regarding the positivity of a few quantities which will appear as denominators in the proof of Theorem 1.

Lemma 3.4.2. *For any given* $\mathbf{y}_{1:\infty}$, $p_k^{\mathbf{y}_{1:k-1}} l_k^{\mathbf{y}_k} > 0$ *for* $k \geq 2$ *and* $p_1 l_1^{\mathbf{y}_1} > 0$.

Proof. Since p_1 is a probability measure, there exists a Borel set $E \subset \mathcal{R}^{d_x}$ such that $p_1(E) > 0$. Define $E_n = \{x | x \in E, l_1^{\mathbf{y}_1}(x) > \frac{1}{n}\}$. Since $l_1^{\mathbf{y}_1}(x) > 0$ for $\forall x$, $E = \bigcup E_n$. From theorem 1.8 in Folland (1999), $p_1(E) = \lim_{n\to\infty} p_1(E_n)$. So, there exist $M \in N$ such that $p_1(E_M) > 0$. $p_1 l_1^{\mathbf{y}_1} \geq \int_{E_N} l_1^{\mathbf{y}_1}(x) p_1(dx) > 0$.

Similarly we can prove that $p_k^{\mathbf{y}_{1:k-1}} l_k^{\mathbf{y}_k} > 0$ for $k \geq 2$. ∎

The third lemma presents the recurrence formula of conditional probability measures.

Lemma 3.4.3. *Denote by* $\{K_k\}_{k=0}^{\infty}$ *the transition kernel of Markov process* $\{x_k\}_{k=0}^{\infty}$ *in system* (1) *for a given control sequence* $\{u_k\}_{k=0}^{\infty}$. *The following recurrence relations hold true:*

$$\begin{cases} p_1 = \pi_0 K_0 \\ \frac{d\pi_1^{\mathbf{y}_{1:1}}}{dp_1} = \frac{l_1^{\mathbf{y}_1}}{\int l_1^{\mathbf{y}_1}(x) p_1(dx)} \end{cases}$$

For $k \geq 2$*:*

$$\begin{cases} p_k^{\mathbf{y}_{1:k-1}} = \pi_{k-1}^{\mathbf{y}_{1:k-1}} K_{k-1} \\ \frac{d\pi_k^{\mathbf{y}_{1:k}}}{dp_k^{\mathbf{y}_{1:k-1}}} = \frac{l_k^{\mathbf{y}_k}}{\int l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)} \end{cases}$$

Proof. For $k \geq 2$,

1. The first equation.

   For $\forall A \in \mathcal{B}(\mathcal{R}^{d_x})$,

$$p_k^{\mathbf{y}_{1:k-1}}(A) = \mathbf{P}(x_k \in A | y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$\pi_{k-1}^{\mathbf{y}_{1:k-1}} K_{k-1}(A) = \int_{\mathcal{R}^{d_x}} K_{k-1}(x, A) \pi_{k-1}^{\mathbf{y}_{1:k-1}}(dx)$$

$$= \int_{\mathcal{R}^{d_x}} \mathbf{P}(x_k \in A | x_{k-1} = x) \pi_{k-1}^{\mathbf{y}_{1:k-1}}(dx)$$

$$= \int_{\mathcal{R}^{d_x}} \mathbf{P}(x_k \in A | x_{k-1} = x)$$

$$\mathbf{P}(x_{k-1} \in dx | y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$= \mathbf{P}(x_k \in A | x_{k-1} \in \mathcal{R}^{d_x}, y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$= \mathbf{P}(x_k \in A | y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$= p_k^{\mathbf{y}_{1:k-1}}(A)$$

So, $p_k^{\mathbf{y}_{1:k-1}} = \pi_{k-1}^{\mathbf{y}_{1:k-1}} K_{k-1}$.

2. The second equation.

From lemma 2, $l_k^{\mathbf{y}_k} p_k^{\mathbf{y}_{1:k-1}} > 0$ for $k \geq 2$. So, the second equation is defined.

$$\frac{d\pi_k^{\mathbf{y}_{1:k}}}{dp_k^{\mathbf{y}_{1:k-1}}} = \frac{l_k^{\mathbf{y}_k}}{\int l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)} \Leftrightarrow$$

for $\forall A \in \mathcal{B}(\mathcal{R}^{d_x})$,

$$\pi_k^{\mathbf{y}_{1:k}}(A) = \frac{\int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)}{\int l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)} \tag{3.9}$$

$D \triangleq \int l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)$, then,

$$(3) \Leftrightarrow \forall \mathbf{y}_{1:k}, \ \ \forall A \in \mathcal{B}(\mathcal{R}^{d_x}),$$

$$\pi_k^{\mathbf{y}_{1:k}}(A) \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k}) = \frac{\int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)}{D} \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k}) \tag{3.10}$$

Left hand side of (4) $= \mathbf{P}(\{x_k \in A\} | y_{1:k} = \mathbf{y}_{1:k}) \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k})$

$$= \mathbf{P}(\{y_{1:k} = \mathbf{y}_{1:k}\} \cap \{x_k \in A\})$$

$$= \mathbf{P}(\{y_k = \mathbf{y}_k\} \cap \{y_{1:k-1} = \mathbf{y}_{1:k-1}\} \cap \{x_k \in A\})$$

$$= \mathbf{P}(\{y_k = \mathbf{y}_k\} \cap \{x_k \in A\} | \{y_{1:k-1} = \mathbf{y}_{1:k-1}\}) \mathbf{P}(y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$= \int_A \mathbf{P}(y_k = \mathbf{y}_k | x_k = x, y_{1:k-1} = \mathbf{y}_{1:k-1}) p_k^{\mathbf{y}_{1:k-1}}(dx) \mathbf{P}(y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$= \int_A \mathbf{P}(y_k = \mathbf{y}_k | x_k = x) p_k^{\mathbf{y}_{1:k-1}}(dx) \mathbf{P}(y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$= \int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx) \mathbf{P}(y_{1:k-1} = \mathbf{y}_{1:k-1})$$

$$= \int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx) \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k}) \frac{\mathbf{P}(y_{1:k-1} = \mathbf{y}_{1:k-1})}{\mathbf{P}(y_{1:k} = \mathbf{y}_{1:k})}$$

$$= \int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx) \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k}) \frac{1}{\mathbf{P}(y_k = \mathbf{y}_k | y_{1:k-1} = \mathbf{y}_{1:k-1})}$$

$$= \int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx) \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k})$$

$$\frac{1}{\int \mathbf{P}(y_k = \mathbf{y}_k | x_k = x, y_{1:k-1} = \mathbf{y}_{1:k-1}) p_k^{\mathbf{y}_{1:k-1}}(dx)}$$

$$= \int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx) \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k})$$

$$\frac{1}{\int l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)}$$

$$= \frac{\int_A l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)}{\int l_k^{\mathbf{y}_k}(x) p_k^{\mathbf{y}_{1:k-1}}(dx)} \mathbf{P}(y_{1:k} = \mathbf{y}_{1:k})$$

which is the right hand side of (6). This proves the second equation.

For $k = 1$, the formula can be proved in the same way as the proof for $k \geq 2$ except minor notation difference. ∎

3.4.5 General Convergence Theorem

We first prove a general convergence theorem which is a revised and generalized version of theorem 2.3.1 in Crisan (2001).

Theorem 3.4.1 (General Convergence Theorem). *Consider system (1). For a given sample path $\mathbf{y}_{1:\infty}$ of $y_{1:\infty}$, probability measures π_0, p_1, $\pi_k^{\mathbf{y}_{1:k}}(k \geq 1)$ and $p_k^{\mathbf{y}_{1:k-1}}$ $(k \geq 2)$ are defined as before. Let $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ be the probability space. Let $\{\tau_0^n : \Omega_1 \to \mathcal{P}(\mathcal{R}^{d_x})\}_{n=1}^{\infty}$ be a sequence of random discrete probability measures on $\mathcal{R}^{d_x}$. For every $k \geq 1$, let $\{\tau_k^n\}_{n=1}^{\infty}$, $\{q_k^n\}_{n=1}^{\infty}$ and $\{\bar{\tau}_k^n\}_{n=1}^{\infty}$ be sequences of random discrete probability measures mapping from Ω_1 to $\mathcal{P}(\mathcal{R}^{d_x})$. For $\forall k, n \geq 1$, $\bar{\tau}_k^n$ depends on q_k^n in the following way. If $q_k^n = \sum_{i=1}^m u_i \delta_{z_i}$,*

$$\bar{\tau}_k^n = \frac{1}{\sum_{i=1}^m u_i l_k^{\mathbf{y}_k}(z_i)} \sum_{i=1}^m u_i l_k^{\mathbf{y}_k}(z_i) \delta_{z_i}$$

where $l_k^{\mathbf{y}_k}$ is defined as in (2).

The following claims hold true:

1. **a1.** *$\{\tau_0^n\}_{n=1}^{\infty}$ converges to π_0 in $\mathcal{L}^p$ sense.*

2. **a2.** $\{q_1^n\}_{n=1}^{\infty}$ *converges to* p_1 *in* $\mathcal{L}^p$ *sense.*

3. **a3.** $\{\bar{\tau}_k^n\}_{n=1}^{\infty}$ *converges to* $\pi_k^{\mathbf{y}_{1:k}}$ *in* $\mathcal{L}^p$ *sense for* $k \geq 1$.

4. **a4.** $\{q_k^n\}_{n=1}^{\infty}$ *converges to* $p_k^{\mathbf{y}_{1:k-1}}$ *in* $\mathcal{L}^p$ *sense for* $k \geq 2$.

if the following three conditions are satisfied.

1. **b1.** $\{\tau_0^n\}_{n=1}^{\infty}$ *converges to* π_0 *in* $\mathcal{L}^p$ *sense.*

2. **b2.** *For* $\forall g \in C_b(\mathcal{R}^{d_x})$, $\lim_{n\to\infty} ||q_k^n g - \tau_{k-1}^n K_{k-1} g||_{\mathcal{L}^p} = 0$ *for* $k \geq 1$.

3. **b3.** *For* $\forall g \in C_b(\mathcal{R}^{d_x})$, $\lim_{n\to\infty} ||\tau_k^n K_k g - \bar{\tau}_k^n K_k g||_{\mathcal{L}^p} = 0$ *for* $k \geq 1$.

Proof. Define $\mathcal{L}^p(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ to be the space of all random variables defined on $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ with finite $\mathcal{L}^p$ norm. To simplify notations, we write p_k for $p_k^{\mathbf{y}_{1:k-1}}$, π_k for $\pi_k^{\mathbf{y}_{1:k}}$ and l_k for $l_k^{\mathbf{y}_k}$ in this proof.

We use induction to prove the claims from **a1** to **a4**.

b1 and **a1** are same statements.

From lemma 3, $p_1 = \pi_0 K_0$. For $\forall g \in C_b(\mathcal{R}^{d_x})$,

$$||q_1^n g - p_1 g||_{\mathcal{L}^p} \leq ||q_1^n g - \tau_0^n K_0 g||_{\mathcal{L}^p} + ||\tau_0^n K_0 g - \pi_0 K_0 g||_{\mathcal{L}^p}$$

From **b2** with $k = 1$, the first item on the right hand side of the inequality goes to zero as $n \to \infty$. From **b1** and Feller property of $\{K_k\}_{k=0}^{\infty}$, the second item on the right hand side of the inequality goes to zero as $n \to \infty$. So,

$$q_1^n g \xrightarrow{\mathcal{L}^p} p_1 g$$

which is **a2**. From lemma 2, $p_1 l_1 > 0$. Following the same argument in lemma 2, we have $q_1^n l_1 > 0$.

$$\begin{aligned}
||\bar{\tau}_1^n g - \pi_1 g||_{\mathcal{L}^p} &= ||\frac{q_1^n g l_1}{q_1^n l_1} - \frac{p_1 g l_1}{p_1 l_1}||_{\mathcal{L}^p} \\
\leq ||\frac{q_1^n g l_1}{q_1^n l_1} - \frac{q_1^n g l_1}{p_1 l_1}||_{\mathcal{L}^p} &+ ||\frac{q_1^n g l_1}{p_1 l_1} - \frac{p_1 g l_1}{p_1 l_1}||_{\mathcal{L}^p} \\
= ||\frac{q_1^n g l_1}{(q_1^n l_1)(p_1 l_1)} p_1 l_1 - q_1^n l_1||_{\mathcal{L}^p} &+ \\
||\frac{q_1^n g l_1}{p_1 l_1} - \frac{p_1 g l_1}{p_1 l_1}||_{\mathcal{L}^p} &
\end{aligned}$$

$$\leq \frac{||g||_u}{|p_1 l_1|}||p_1 l_1 - q_1^n l_1||_{\mathcal{L}^p} +$$
$$\frac{1}{|p_1 l_1|}||p_1 g l_1 - q_1^n g l_1||_{\mathcal{L}^p}$$

where $||g||_u = sup_x\{|g(x)|\} < \infty$. By lemma 1, l_1 and $l_1 g$ are both continuous and bounded. From **a2**, $\bar{\tau}_1^n g \xrightarrow{\mathcal{L}^p} \pi_1 g$. **a3** holds true for $k = 1$.

Suppose the claims **a3** and **a4** hold true for $M-1$ with $M \geq 2$. Then, for $\forall g \in C_b(\mathcal{R}^{d_x})$,

$$q_{M-1}^n g \xrightarrow{\mathcal{L}^p} p_{M-1} g$$
$$\bar{\tau}_{M-1}^n K_{M-1} g \xrightarrow{\mathcal{L}^p} \pi_{M-1} K_{M-1} g \tag{3.11}$$

In (5), we used ass 2. For $\forall g \in C_b(\mathcal{R}^{d_x})$,

$$||q_M^n g - p_M g||_{\mathcal{L}^p} \leq ||q_M^n g - \tau_{M-1}^n K_{M-1} g||_{\mathcal{L}^p} +$$
$$||\tau_{M-1}^n K_{M-1} g - \bar{\tau}_{M-1}^n K_{M-1} g||_{\mathcal{L}^p} + ||\bar{\tau}_{M-1}^n K_{M-1} g - p_M g||_{\mathcal{L}^p}$$
$$= ||q_M^n g - \tau_{M-1}^n K_{M-1} g||_{\mathcal{L}^p} + ||\tau_{M-1}^n K_{M-1} g - \bar{\tau}_{M-1}^n K_{M-1} g||_{\mathcal{L}^p}$$
$$+||\bar{\tau}_{M-1}^n K_{M-1} g - \pi_{M-1} K_{M-1} g||_{\mathcal{L}^p}$$

By **b2** with $k = M$, the first item on the right hand side of equation goes to zero as $n \to \infty$. By **b3** with $k = M-1$, the second item on the right hand side goes to zero as $n \to \infty$. By **a3** with $k = M-1$ and Feller Property of K_{M-1}, the third item on the right hand side goes to zero as $n \to \infty$. So, for $\forall g \in C_b(\mathcal{R}^{d_x})$,

$$q_M^n g \xrightarrow{\mathcal{L}^p} p_M g \tag{3.12}$$

a4 holds for $k = M$.

From lemma 2, $p_M l_M > 0$. Following the same argument in lemma 2, we have $q_M^n l_M > 0$.

$$||\bar{\tau}_M^n g - \pi_M g||_{\mathcal{L}^p} = ||\frac{q_M^n g l_M}{q_M^n l_M} - \frac{p_M g l_M}{p_M l_M}||_{\mathcal{L}^p}$$
$$\leq ||\frac{q_M^n g l_M}{q_M^n l_M} - \frac{q_M^n g l_M}{p_M l_M}||_{\mathcal{L}^p} + ||\frac{q_M^n g l_M}{p_M l_M} - \frac{p_M g l_M}{p_M l_M}||_{\mathcal{L}^p}$$
$$= ||\frac{q_M^n g l_M}{(q_M^n l_M)(p_M l_M)} p_M l_M - q_M^n l_M||_{\mathcal{L}^p} +$$
$$||\frac{q_M^n g l_M}{p_M l_M} - \frac{p_M g l_M}{p_M l_M}||_{\mathcal{L}^p}$$

$$\leq \frac{||g||_u}{|p_M l_M|}||p_M l_M - q_M^n l_M||_{\mathcal{L}^p} +$$
$$\frac{1}{|p_M l_M|}||q_M^n g l_M - p_M g l_M||_{\mathcal{L}^p}$$

where $||g||_u = sup_x\{|g(x)|\} < \infty$.

That is,

$$||\bar{\tau}_M^n g - \pi_M g||_{\mathcal{L}^p} \leq \frac{||g||_u}{|p_M l_M|}||p_M l_M - q_M^n l_M||_{\mathcal{L}^p}$$
$$+\frac{1}{|p_M l_M|}||q_M^n g l_M - p_M g l_M||_{\mathcal{L}^p} \tag{3.13}$$

By lemma 1, l_M and gl_M are both continuous. From (6), we see $||p_M l_M - q_M^n l_M||_{\mathcal{L}^p} \to 0$ and $||p_M g l_M - q_M^n g l_M||_{\mathcal{L}^p} \to 0$ as $n \to \infty$. So,

$$||\bar{\tau}_M^n g - \pi_M g||_{\mathcal{L}^p} \to 0$$

as $n \to \infty$. **a3** holds true for $k = M$.

■

3.4.6 Convergence Theorem for QMSMC Method

To apply the general convergence theorem to QMSMC method, we first need to show that ρ_k^n, $\bar{\pi}_k^n$ and π_k^n generated in QMSMC method are random discrete probability measures. This is confirmed in the following lemma.

Lemma 3.4.4. *Consider Problem 1 and QMSMC method. For a given sample path* $\mathbf{y}_{1:K}$ *of* $y_{1:K}$. ρ_k^n, $\bar{\pi}_k^n$ *and* π_k^n *generated in QMSMC method are random discrete probability measures for* $1 \leq k \leq K$.

Proof. Let $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ be the probability space on which random variables x_0, $\{w_k\}_{k=0}^{K-1}$, $\{v_k\}_{k=1}^{K}$ and uniform random variable a used in resampling steps are defined.

We use induction.

For $k = 1$, since $\bar{\mathbf{x}}_1^i = f(\mathbf{x}_0^i, u_0, \mathbf{w}_0^i)$ and f measurable, $\bar{\mathbf{x}}_1^i$ is a random variable for every $1 \leq i \leq n$. So, $\rho_1^n = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{\mathbf{x}}_1^i}$ is a random discrete probability measure.

From lemma 1, for $\forall \mathbf{y}_1$, $l_1^{\mathbf{y}_1}(x)$ is continuous and then measurable. $g : x \to \frac{1}{x}$ is measurable for $x > 0$. $u_1^i = \frac{l_1^{\mathbf{y}_1}(\bar{\mathbf{x}}_1^i)}{\sum_{i=1}^n l_1^{\mathbf{y}_1}(\bar{\mathbf{x}}_1^i)}$ is a random variable for all $1 \leq i \leq n$ because the denominator is always strictly positive by lemma 1. Easy to see $\sum_{i=1}^n u_1^i = 1$. $\bar{\pi}_1^n$ is a random discrete probability measure.

Define set A to be $\{(u_1, u_2, \cdots, u_n) \in \mathcal{R}^n | 0 \leq u_i \leq 1, \sum_{i=1}^n u_i = 1\}$. Construct $f_a : A \times \{\mathcal{R}^{d_x}\}^n \times [0,1) \to \mathcal{R}^d$ as the following. For $f_a(u_1, u_2, \cdots, u_n, x_1, x_2, \cdots, x_n, \alpha)$, we set $u_0 = 0$ and find l such that $\sum_{i=0}^{l-1} u_i \leq \alpha < \sum_{i=0}^{l} u_i$. $f_a(u_1, u_2, \cdots, u_n, x_1, x_2, \cdots, x_n, \alpha) = x_l$. Easy to see that f_a is a function. For $\forall E \in \mathcal{B}(\mathcal{R}^{d_x})$, $f_a^{-1}(E) = \bigcup_{i=1}^n \{x_i \in E \cap \sum_{j=0}^{i-1} u_j \leq \alpha < \sum_{j=0}^{i} u_j\}$. For every i, $\mathbf{x}_1^i$ can be viewed as $f_a(u_1^1, u_1^2, \cdots, u_1^n, \bar{\mathbf{x}}_1^1, \bar{\mathbf{x}}_1^2, \cdots, \bar{\mathbf{x}}_1^n, a)$. Since $u_1^1, u_1^2, \cdots, u_1^n, \bar{\mathbf{x}}_1^1, \bar{\mathbf{x}}_1^2, \cdots, \bar{\mathbf{x}}_1^n, a$ are all random variables, $f_a^{-1}(E) = \bigcup_{i=1}^n \{x_i \in E \cap \sum_{j=0}^{i-1} u_j \leq \alpha < \sum_{j=0}^{i} u_j\}$ is measurable for $\forall E \in \mathcal{B}(\mathcal{R}^{d_x})$, which means that $\mathbf{x}_1^i, 1 \leq i \leq n$ are random variables. Since $\sum_{i=1}^n u_1^i = 1$, π_1^n is a random discrete probability measure.

Suppose $\mathbf{x}_{M-1}^i, 1 \leq i \leq n$ are random variables for $M \geq 2$, we can just follow the steps for $k = 1$ to prove the ρ_M^n, $\bar{\pi}_M^n$, π_M^n are random discrete probability measures and $\mathbf{x}_M^i, 1 \leq i \leq n$ are random variables. ∎

Then, we can prove the convergence of $\bar{\pi}_k^n$ in QMSMC method to $\pi_k^{\mathbf{y}_{1:k}}$ in $\mathcal{L}^1$ sense as $n \to \infty$.

Theorem 3.4.2 (Convergence Theorem for QMSMC Method). *Consider system (1). Given a sample path $\mathbf{y}_{1:\infty}$ of $y_{1:\infty}$. Probability measures π_0, p_1 and $\pi_k^{\mathbf{y}_{1:k}} (1 \leq k \leq K)$ and $p_k^{\mathbf{y}_{1:k-1}}$ $(2 \leq k \leq K)$ are defined as before.*

Consider QMSMC method for $n = 1, 2, \cdots, \infty$. The following claims regarding random discrete probability measure sequences $\{\bar{\pi}_k^n\}_{n=1}^\infty$ and $\{\rho_k^n\}_{n=1}^\infty$ generated in QMSMC method hold true.

1. *$\{\rho_1^n\}_{n=1}^\infty$ converges to p_1 in $\mathcal{L}^1$ sense.*
2. *$\{\bar{\pi}_k^n\}_{n=1}^\infty$ converges to $\pi_k^{\mathbf{y}_{1:k}}$ in $\mathcal{L}^1$ sense for $1 \leq k \leq K$.*
3. *$\{\rho_k^n\}_{n=1}^\infty$ converges to $p_k^{\mathbf{y}_{1:k-1}}$ in $\mathcal{L}^1$ sense for $2 \leq k \leq K$.*

Proof. Let $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ be the probability space on which random variables x_0, $\{w_k\}_{k=0}^{K-1}$, $\{v_k\}_{k=1}^K$ and uniform random variable a used in resampling steps are defined. This probability space is also the space on which $\{\pi_0^n\}_{n=1}^\infty$, $\{\rho_k^n\}_{n=1}^\infty$, $\{\bar{\pi}_k^n\}_{n=1}^\infty$ and $\{\pi_k^n\}_{n=1}^\infty$ are defined for $1 \leq k \leq K$. From QMSMC method, we see that $\bar{\pi}_k^n$ depends on ρ_k^n in the same way as $\bar{\tau}_k^n$ depends on q_k^n. So, we only need to verify that:

1. **b1**. $\{\pi_0^n\}_{n=1}^\infty$ converges to π_0 in $\mathcal{L}^1$ sense.
2. **b2**. For $\forall g \in C_b(\mathcal{R}^{d_x})$, $\lim_{n\to\infty} ||\rho_k^n g - \pi_{k-1}^n K_{k-1} g||_{\mathcal{L}^1} = 0$ for $k \geq 1$.
3. **b3**. For $\forall g \in C_b(\mathcal{R}^{d_x})$, $\lim_{n\to\infty} ||\pi_k^n K_k g - \bar{\pi}_k^n K_k g||_{\mathcal{L}^1} = 0$ for $k \geq 1$.

The verification is as follows:

1. For $\forall g \in C_b(\mathcal{R}^{d_x})$, $\pi_0^n g = \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_0^i)$. Since $\{\mathbf{x}_0^i\}_{i=1}^n$ are independent samples of random variable x_0, $\{g(\mathbf{x}_0^i)\}_{i=1}^n$ are independent samples of random variable $g(x_0)$. Mean of $g(x_0)$ is $\pi_0 g \leq ||g||_u < \infty$. By Strong Law of Large Numbers, $\pi_0^n g \overset{a.s.}{\to} \pi_0 g$ as $n \to \infty$. So, $\pi_0^n g \overset{\mathcal{L}^1}{\to} \pi_0 g$. **b1** holds true.

2. For $k \geq 1$ and $\forall g \in C_b(\mathcal{R}^{d_x})$,

$$\rho_k^n g = \frac{1}{n}\sum_{i=1}^n g(\bar{\mathbf{x}}_k^i)$$

$$\pi_{k-1}^n K_{k-1} g = \frac{1}{n}\sum_{i=1}^n E[g(x_k)|x_{k-1} = \mathbf{x}_{k-1}^i]$$

$$||\rho_k^n g - \pi_{k-1}^n K_{k-1} g||_{\mathcal{L}^2} = \frac{1}{n}||\sum_{i=1}^n \{g(\bar{\mathbf{x}}_k^i) - E[g(x_k)|x_{k-1} = \mathbf{x}_{k-1}^i]\}||_{\mathcal{L}^2}$$

$$= \frac{1}{n}||\sum_{i=1}^n \{g(f(\mathbf{x}_{k-1}^i, u_{k-1}, \mathbf{w}_{k-1}^i)) - E[g(f(\mathbf{x}_{k-1}^i, u_{k-1}, w_{k-1}))]\}||_{\mathcal{L}^2}$$

From the independence of $\{\mathbf{w}_{k-1}^i\}_{i=1}^n$, we can see that

$$||\sum_{i=1}^n \{g(f(\mathbf{x}_{k-1}^i, u_{k-1}, \mathbf{w}_{k-1}^i)) - E[g(f(\mathbf{x}_{k-1}^i, u_{k-1}, w_{k-1}))]\}||_{\mathcal{L}^2}$$

$$= (\sum_{i=1}^n ||g(f(\mathbf{x}_{k-1}^i, u_{k-1}, \mathbf{w}_{k-1}^i)) - E[g(f(\mathbf{x}_{k-1}^i, u_{k-1}, w_{k-1}))]||_{\mathcal{L}^2}^2)^{\frac{1}{2}}$$

$$\leq \sqrt{n}||g^2||_u^{\frac{1}{2}}$$

So,

$$||\rho_k^n g - \pi_{k-1}^n K_{k-1} g||_{\mathcal{L}^2} \leq \frac{1}{\sqrt{n}}||g^2||_u^{\frac{1}{2}}$$

$||\rho_k^n g - \pi_{k-1}^n K_{k-1} g||_{\mathcal{L}^2} \to 0$ as $n \to \infty$. So, $\lim_{n\to\infty} ||\rho_k^n g - \pi_{k-1}^n K_{k-1} g||_{\mathcal{L}^1} = 0$. **b2** holds true.

3. For $\forall k \geq 1$ and $\forall g \in C_b(\mathcal{R}^{d_x})$,

$$\pi_k^n g = \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_k^i)$$

Since $\mathbf{x}_k^i$ are independent samples from distribution $\bar{\pi}_k^n$

$$E[g(\mathbf{x}_k^i)] = \frac{1}{n}\sum_{i=1}^{n} u_k^i g(\bar{\mathbf{x}}_k^i)$$
$$= \bar{\pi}_k^n g \leq ||g||_u < \infty$$

By Strong Law of Large Numbers and Feller property of $\{K_k\}_{k=0}^{\infty}$,

$$\pi_k^n K_k g \overset{a.s.}{\to} \bar{\pi}_k^n K_k g$$

as $n \to \infty$ in π_k^n. So $\lim_{n\to\infty} ||\pi_k^n K_k g - \bar{\pi}_k^n K_k g||_{\mathcal{L}^1} = 0$. **b3** holds true.

Theorem 3.4.3. *In QMSMC method, as the size of samples $n \to \infty$, random variable $\hat{h}(x_K)$ converges to $E[h(x_K)|y_{1:K} = \mathbf{y}_{1:K}]$ in probability.*

Proof. The corollary is interpreting that $\bar{\pi}_K^n$ in QMSMC method converges to $\pi_K^{\mathbf{y}_{1:K}}$ as $n \to \infty$ in $\mathcal{L}^1$ sense.

As a stochastic algorithm, QMSMC method generates a sequence of random variables indexed by size of samples used. The sequence of random variables converges to the value sought in the optimal filtering problem in probability. This is the asymptotical optimality of QMSMC method.

Theorem 3.4.4. *In QMSMC method, if f's range spaces are uniformly bounded for a given control sequence $\{u_k\}$, then, $\hat{h}(x_K)$ converges to $E[x_K|y_{1:K} = \mathbf{y}_{1:K}]$ in probability.*

Proof. Choose $h(x_k)$ to be a smooth function with compact support such that $h(x_k) = x_k$ if$x_k \in D$ and $h(x_k) = 0$ if $x_k \in S(U, d)$. where D is a compact set including f's range spaces with respect to control sequence $\{u_k\}$. $S(u, d) = \{x | x \in \mathcal{R}^d, dist(U, x) > d\}$. Then, the conclusion is obvious. ∎

Remark 3.4.1. *For this problem we are considering, sequential Monte Carlo is trading computation for accuracy in estimation. QMSMC's asymptotical optimality implies that the computation power required to gain significant margin over other simple but effective approximation would be extraordinary, in some cases.*

3.4.7 Variants in QMSMC

Resampling step is arguably the most important step in SMC. Different variants exist in this step.

Algorithm 3.4.1.

3.4.8 On its way to fame: Open problems and future work for SMC

Several crucial properties are still missing

1. Uniform Convergence
2. Convergence rate

Chapter 4

Simulation for Noisy Measurement

In chapter 3, we proposed QMKF, QMEKF and QMSMC for filtering from quantized measurement. Asymptotical optimality of QMSMC in a defined sense is proven under certain assumptions. However, it is very difficult, if possible, to analytical compare the performances of these methods for a given system. This chapter will demonstrate the relative effectiveness of QMKF, QMEKF and QMSMC by simulation. We apply QMKF and QMSMC for linear time invariant systems with quantized measurement and QMEKF and QMSMC for nonlinear systems with quantized measurements. The relative performances of applicable methods will certainly depend on the choice of systems to which they are applied to. We will choose a few sample linear time invariant systems with order 2, 3 and 4. We don't intend to do exhaustive simulation for linear time invariant systems of all orders since even if it is possible in a meaningful way, that huge effort is not necessarily worthwhile. Since in most practical cases, quantizer is under control of the designer of the system, we choose quantizer for those sample systems based on our experience rather than an arbitrary one.

Vastly different nonlinearities exist within the category of nonlinear system. Those nonlinear systems which are models of real world systems are of greatest interests to our study. In applying QMEKF and QMSMC to nonlinear systems, we focus the problem of filtering from quantized measurement for the navigation model of MIT instrumented X-60 SE Cell helicopter. Simulation confirmed the asymptotical optimality of QMSMC and showed that under certain conditions, with still reasonable computation costs, QMSMC will significantly outperform QMEKF.

4.1 QMKF and QMSMC for Linear Systems

Consider the following canonical form of continuous-time linear time invariant system with single input and single output.

$$\begin{aligned} \dot{x} &= Ax + B\dot{w} \\ y &= Cx + D\dot{v} \end{aligned} \tag{4.1}$$

where

$$A = \left[\begin{array}{c|c} & I_{n-1} \\ \hline & \\ -a_n & -a_{n-1} \cdots - a_1 \end{array}\right]$$

$$B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

$$C = \begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix}$$

I_{n-1} is the $n-1$ dimensional identity matrix. D is a constant. w and v are wiener processes.

Discretizing this system by time interval dt and considering quantization in measurement, we can obtain the following discrete-time system.

$$\begin{aligned} x_{k+1} &= e^{Adt}x_k + Bw_k \\ y_k &= q(Cx_k + Dv_k) \end{aligned} \tag{4.2}$$

where dt is time step. w_k, v_k are white stationary zero mean gaussian processes independent with each other. At each time index k, w_k and v_k are of variance dt.

We choose $dt = 0.01$ in our simulation. Different time steps will result in different simulation results. For simplicity, we fix dt in our simulation and will not consider the effect of time step on simulation results.

Choosing a canonical form simplifies the comparison between different systems. We don't know whether a specific choice of canonical form will affect our study. In this thesis, we only consider the canonical form as above.

4.1.1 Second Order Systems

We choose one particular second order linear time invariant system model with specified parameters and will change noise variance in measurement and sample size in QMSMC.

System Specification

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

Transfer function of this system without considering observation noise is

$$H(s) = \frac{\sqrt{2}}{s^2 + s + 1}$$

Eigenvalues of A are $s_{1,2} = -\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$. Damping ratio of this system is 0.5. We let D vary from 4 to 16 with multiplication of $\sqrt{2}$ at each step. Steady state distribution of Cx_k is the standard normal distribution. Easy to see that steady state distribution for $Cx_k + Dv_k$ will be zero mean normal distribution with variance $1 + D^2 * 0.01$.

Parameters

We fixed quantizer to be of the following boundary points.

$$B_q = \left\{ \begin{array}{ccccccc} -1.5, & -1, & -0.5, & 0, & 0.5, & 1, & 1.5 \end{array} \right\}$$

Choice of this particular quantizer is based on its simple structure and closeness to the optimal quantizer for static standard normal distribution. The sample size of QMSMC will be 100, 400, 1600. Each realization of the system is 1000 time unit long and we simulate 10 realizations for each combination of noise variance and QMSMC's sample size. The average cost function for the 10 realizations is plotted.

Simulation results

The following figure shows the performances of QMKF and QMSMC with different sample sizes for different levels of observation noise. The cost function is defined in the following way:

$$P = \sqrt{E(e_k^T e_k)}$$

where E denotes expectation, $e_k = \hat{x}_k - x_k$. To calculate $E(e_k^T e_k)$, we assume that e_k is a stationary vector stochastic process and ergodic. $E(e_k^T e_k)$ is calculated by time average.

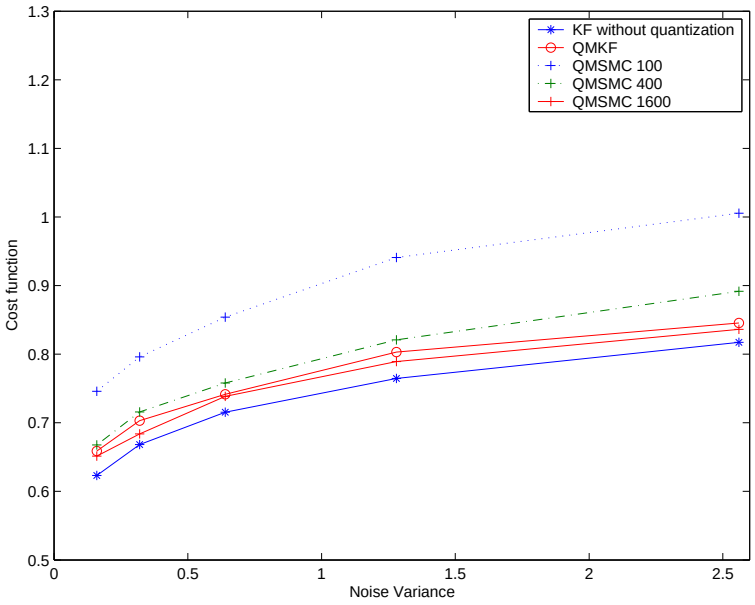


Figure 4.1: Simulation Results for System I

In this figure, we also plot the cost function by Kalman filter when it is applied to the system model which is same as in system specification except no quantization in measurement model.

The table below lists all the data used to draw the pictures above as well as the estimated standard deviation of each of the item.

Analysis

Simulation for this system demonstrates the asymptotical optimality of QMSMC and gives us an approximate idea about the size of samples needed in QMSMC algorithm for it to

Table 4.1: Table of standard deviation of simulated results

Noise Variance	KF	QMKF	σ_{QMKF}	QMSMC 100	σ_{100}
0.160000	0.6232	0.658803	0.002854	0.745832	0.001855
0.320000	0.6682	0.703094	0.004548	0.796086	0.002483
0.640000	0.7153	0.741482	0.003516	0.854065	0.004414
1.280000	0.7647	0.803052	0.003426	0.940981	0.004680
2.560000	0.8172	0.845374	0.004381	1.005394	0.007102
Noise Variance	QMSMC 400	σ_{400}	QMSMC 1600	σ_{1600}	
0.160000	0.667665	0.001907	0.651310	0.002180	
0.320000	0.715787	0.002002	0.683837	0.001692	
0.640000	0.757987	0.002005	0.738392	0.002385	
1.280000	0.820808	0.002958	0.789307	0.003159	
2.560000	0.891733	0.003355	0.836184	0.002643	

match the performance of QMKF for this specific system. From simulation, the sample size would be around 800.

Figure 4.1 also provides us information about the limit performance of QMSMC algorithm as computation goes to infinity which, by theory, is the error associated with optimal filtering in least mean-square-error sense. This limit arguably should be greater than the cost function caused by Kalman filter for the same system expect quantization in measurement model. The gap between them is small, which means that the deterioration in the quality of state estimation caused by quantization is small. It is interesting to see that allowing a relatively larger error in state estimation, we can reduce the requirement of measurement from infinite precision to 3 bits representation. For quantized measurement, with about three orders of magnitude more computation, QMSMC can outperform QMKF for this system. Whether this cost is worthy the reduction in state estimation error is to be determined by practitioners.

4.1.2 Third Order Systems

System Specification

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -2 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ \sqrt{3} \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

Transfer function of this system without considering observation noise is

$$H(s) = \frac{\sqrt{3}}{s^3 + 2s^2 + 2s + 1}$$

Eigenvalues of A are $s_{1,2} = -\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$ and $s_3 = -1$. We let D vary from 4 to 16 with multiplication of $\sqrt{2}$ at each step. Steady state distribution of Cx_k is the standard normal distribution. Easy to see that steady state distribution for $Cx_k + Dv_k$ will be zero mean normal distribution with variance $1 + D^2 * 0.01$.

Parameters

We fixed quantizer to be of the following boundary points.

$$B_q = \left\{ \begin{array}{ccccccc} -1.5, & -1, & -0.5, & 0, & 0.5, & 1, & 1.5 \end{array} \right\}$$

Choice of this particular quantizer is based on its simple structure and closeness to the optimal quantizer for static standard normal distribution. The sample size of QMSMC will be 100, 200, 400, 800, 1600.

Simulation results

4.1.3 Fourth Order Systems

System Specification

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -3 & -4 & -3 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

Transfer function of this system without considering observation noise is

$$H(s) = \frac{2}{s^4 + 3s^3 + 4s^2 + 3s + 1}$$

. Eigenvalues of A are $s_{1,2} = -\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$ and $s_{3,4} = -1$. We let D vary from 4 to 16 with multiplication of $\sqrt{2}$ at each step. Steady state distribution of Cx_k is the standard normal distribution. Easy to see that steady state distribution for $Cx_k + Dv_k$ will be zero mean normal distribution with variance $1 + D^2 * 0.01$.

Parameters

We fixed quantizer to be of the following boundary points.

$$B_q = \begin{Bmatrix} -1.5, & -1, & -0.5, & 0, & 0.5, & 1, & 1.5 \end{Bmatrix}$$

Choice of this particular quantizer is based on its simple structure and closeness to the optimal quantizer for static standard normal distribution. The sample size of QMSMC will be 100, 200, 400, 800, 1600.

Simulation results

4.2 QMEKF and QMSMC for MIT instrumented X-60 SE Cell helicopter

We consider the problem of filtering with quantized measurement for the navigation model of MIT instrumented X-60 SE Cell helicopterGavrilets (2003).

4.2.1 System Specification

For given discretization interval δt, the discrete-time state-space navigation model is as follows:

$$x_{k+1} = f_h(x_k, u_k, w_k) \tag{4.3}$$

$$y_k = q_h(Dx_k + v_k) \tag{4.4}$$

where the state vector $x_k = [p_k; v_k; q_k; a_k^b; w_k^b]$. p_k, v_k, a_k^b, w_k^b are 3×1 vectors. q_k is 4×1 vector. Control vector $u_k = [a_k; \omega_k]$ where a_k and ω_k are 3×1 vectors. $w_k = [w_{1,k}; w_{2,k}; w_{3,k}]$

is system noise vector. $w_{1,k}$, $w_{2,k}$ and $w_{3,k}$ are 3×1 vectors. $D = \begin{bmatrix} I_6 & O \end{bmatrix}$ where I_6 is 6-dimension identity matrix. O is 6×9 zero matrix.

The system function f_h is as the following:

$$p_{k+1} = p_k + v_k \delta t$$

$$v_{k+1} = v_k + (C_k(a_k - a_k^b) + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} g)\delta t + w_{1,k}$$

$$q_{k+1} = \Phi(\omega_k - \omega_k^b, \delta t) q_k + k\delta t(1 - \|q_k\|^2)$$

$$a_{k+1}^b = a_k^b + w_{2,k}$$

$$\omega_{k+1}^b = \omega_k^b + w_{3,k}$$

In these equations, $g = 9.81$. $q_k = \begin{bmatrix} q_{k,1} & q_{k,2} & q_{k,3} & q_{k,4} \end{bmatrix}^T$. T denotes transpose. C_k is obtained from q_k as follows.

$$C_k = \begin{bmatrix} 1 - 2(q_{k,3}^2 + q_{k,4}^2) & 2(q_{k,2}q_{k,3} - q_{k,1}q_{k,4}) & 2(q_{k,2}q_{k,4} + q_{k,1}q_{k,3}) \\ 2(q_{k,2}q_{k,3} + q_{k,1}q_{k,4}) & 1 - 2(q_{k,2}^2 + q_{k,4}^2) & 2(q_{k,3}q_{k,4} - q_{k,1}q_{k,2}) \\ 2(q_{k,2}q_{k,4} - q_{k,1}q_{k,3}) & 2(q_{k,3}q_{k,4} + q_{k,1}q_{k,2}) & 1 - 2(q_{k,3}^2 + q_{k,3}^2) \end{bmatrix}$$

$\Phi(\omega_k - \omega_k^b, \delta t)$ is obtained as follows:

$$\phi = (\omega_k - \omega_k^b)\delta t = \begin{bmatrix} \phi_x & \phi_y & \phi_z \end{bmatrix}^T$$

$$\Phi_4 = \begin{bmatrix} 0 & \phi_x & \phi_y & \phi_z \\ -\phi_x & 0 & -\phi_z & \phi_y \\ -\phi_y & \phi_z & 0 & -\phi_x \\ -\phi_z & -\phi_y & \phi_x & 0 \end{bmatrix}$$

$$s = \frac{1}{2}\|\phi\|_2$$

$$\Phi(\omega_k - \omega_k^b, \delta t) = I_4 \cos s + \frac{1}{2}\Phi_4 \frac{\sin s}{s}$$

where I_4 is the 4-dimension identity matrix and $||\cdot||_2$ is the L^2 norm. Quantizer q_h quantizes each component of $Dx_k + v_k$ as a scalar. The scalar quantizer for each component is uniform with rate b. Their granular regions are chosen to fit the specific trajectory we simulate.

Table 4.2: Standard deviation(Std) of every component of noises

Noise	$w_{k,1}$	$w_{k,2}$	$w_{k,3}$	$w_{k,4}$	$w_{k,5}$	$w_{k,6}$
Std	7×10^{-3}	7×10^{-3}	7×10^{-3}	$2 \times d \times 10^{-5}$	$2 \times d \times 10^{-5}$	$5 \times d \times 10^{-6}$
Noise	$w_{k,7}$	$w_{k,8}$	$w_{k,9}$	$v_{k,1}$	$v_{k,2}$	$v_{k,3}$
Std	5×10^{-5}	5×10^{-5}	5×10^{-5}	2	2	0.304
Noise	$v_{k,4}$	$v_{k,5}$	$v_{k,6}$			
Std	0.45	0.45	1			

4.2.2 Parameters

We simulate the system for 200 seconds with discretization interval $\delta t = 0.01$. In the simulation, we apply an artificially generated control sequences $\{u_k\}_{k=0}^{20,000}$. The initial distribution for x_0 is Gaussian. Each component of vectors w_k and v_k is chosen to be zero-mean stationary white Gaussian process which is independent of each other. The standard deviations of every components at any time index k is listed in Table 1 in which $d = \pi/180$. The components of w_k and v_k are numbered as follows.

$$w_k = [\ w_{k,1} \quad w_{k,2} \quad w_{k,3} \quad w_{k,4} \quad w_{k,5} \quad w_{k,6} \quad w_{k,7} \quad w_{k,8} \quad w_{k,9}\ ]^T$$

$$v_k = [\ v_{k,1} \quad v_{k,2} \quad v_{k,3} \quad v_{k,4} \quad v_{k,5} \quad v_{k,6}\ ]^T$$

We treat quantization as additive noise with zero mean and variance $\frac{1}{12}L_c^2$ where L_c is the length of a cell in corresponding scalar quantizer.

Since q_k, a_k^b, ω_k^b have very different units with p_k and v_k, we only compute the square root of mean square error of $[p_k; v_k]$. Because QMSMC method's estimation result at any time index k is a random variable for a given sample size n, we run QMSMC method 10 times for a given n and compute the mean of the 10 resulting values as the cost function. This process is repeated for different quantizer rate b to show the effect of b. The result is shown in Figure 1.

The cost function of error in state estimation is the square root of sum of estimation errors variances of the first 6 components of the state vector, which are positions and velocities. The error of attitude is excluded because of no proper way to combine the error in position and velocity and error in attitudes.

The rate of scalar quantizer of each component are chosen to be the same for all 6 observation components and range from 5 to 10. The corresponding communication requirement between the helicopter and the controller is from $3k$ bits per second(bps) to $6k$ bps.

The choice of the number of samples used in SMC algorithm depends on the computational capacity of the controller and the required accuracy. As a random algorithm, SMC's estimation result is random variable. We simulate the SMC for each given quantizer and number of samples for 10 times and compute the mean and standard deviation.

4.2.3 Simulation results

Figure 4.2: Performance of QMSMC method with $n = 1000, 2000$ and EKF

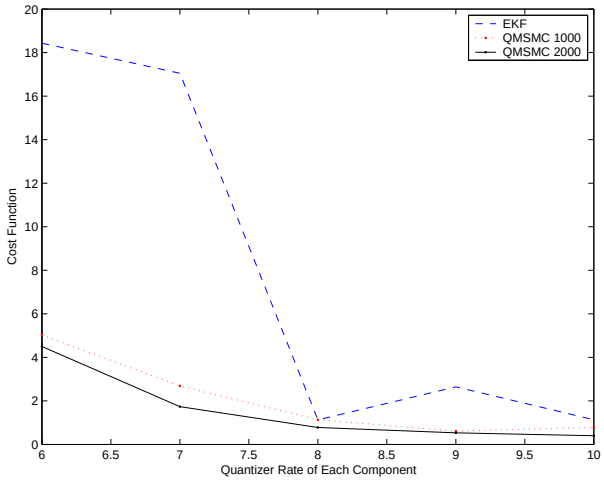


Figure 1 shows as n increases from 1000 to 2000, the cost function decreases for each quantizer rate. This is expected from Corollary 1. This figure also shows that for quantizer rate $b = 6, 7$, QMSMC method outperforms EKF by large margin.

The trajectory in north and east direction is shown in figure 1 and the trajectory in the down direction is shown in figure 2:

Extended Kalman Filter and the Sequential Monte Carlo method are both applied to estimate the state from the quantized noisy observation. In sequential monte carlo method, since the system and observation noises are gaussian distributed, gaussian random variable generator in matlab is used to generate independent system and observation noises samples.

4.2.4 Analysis

Easy to see that (8) and (9) satisfy Assumption 1. Furthermore, we can show that for the given control sequence, transition kernel of process $\{x_k\}_{k=0}^{20,000}$ has Feller property. We

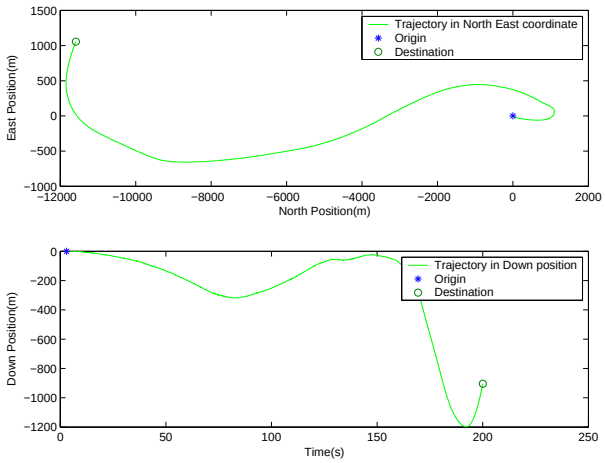


Figure 4.3: Simulated Trajectory

apply QMSMC to the problem of filtering in this example. In applying QMSMC method, we choose $h(x)$ in Problem 1 to be continuous with compact support and is identity function within a large enough region which makes the difference between $E[h(x_K)|y_{1:K} = \mathbf{y}_{1:K}]$ and $E[x_K|y_{1:K} = \mathbf{y}_{1:K}]$ negligible for $\forall K = 1, 2, \cdots, 20000$. Then, from corollary 1, as the size of samples $n \to \infty$, random variable $\hat{h}(x_K)$ converges to $E[x_K|y_{1:K} = \mathbf{y}_{1:K}]$ in probability. As a comparison, we also apply Extended Kalman Filter(EKF). Details of EKF can be found in Gavrilets (2003).

4.2.5 Under the first assumption

Figure 2 presents the simulation results of EKF and SMC with 1000 and 2000 particles for different quantization rates under the first assumption.

Notice that at high quantization rate, the noise introduced by quantization is relatively small, the EKF and SMC yields similar results for state estimation. However, as the rate decreases, which means quantization becomes coarser, SMC can increasingly outperform EKF. This also means to achieve the same estimation error, SMC requires less communication than EKF.

4.2.6 Under the second assumption

Figure 3 presents the simulation results of EKF for different quantization rates under the second assumption and those of SMC with 1000 and 2000 particles under the first assump-

tion.

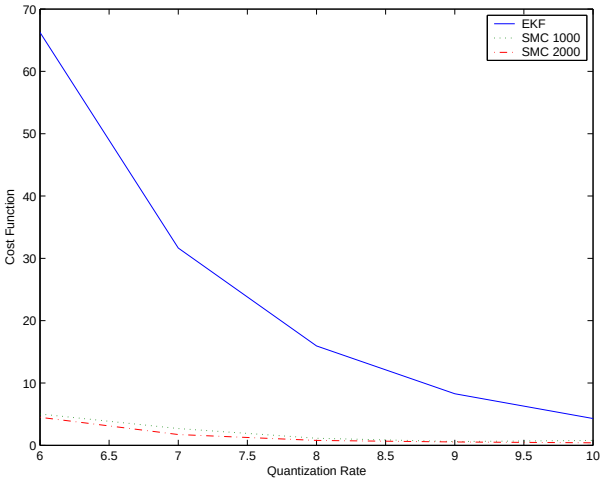


Figure 4.4: Comparison of EKF under second assumption and SMC with 1000 and 2000 particles under first assumption

This plot clearly shows that to apply SMC under first assumption greatly outperform EKF under second assumption which suppose the helicopter estimate its state on board and transmits the estimation results.

4.2.7 Conclusion

We proposed a new sequential monte carlo method to estimate state in navigation model of MIT instrumented X-Cell 60 SE helicopter under quantized observation. The performances of the proposed sequential monte carlo method and extended kalman filter in terms of defined cost function are considered and compared for two different assumptions. Simulation showed that the proposed sequential monte carlo method gives outperform the extended kalman filter under both assumptions. This provides a way to obtain more accurate state estimation under severe communication constraints at the cost of high computational load.

Chapter 5

State Estimation from Quantized Measurement II : Noiseless Measurement

5.1 Motivation and Formulation

In Chapter 2, under a set of assumptions, we formulated system with limited communication into a specific class of models as below.

$$\begin{aligned} x_k &= f(x_{k-1}, u_{k-1}, w_{k-1}) \\ y_k &= q(C_k x_k + v_k) \end{aligned} \tag{5.1}$$

In this class of models there exists nonzero additive noise v_k in measurement before quantization.

In this chapter, we will study the following class of models which are different from models (5.1) in the sense that there is no additive noise in measurement before quantization.

$$\begin{aligned} x_k &= f(x_{k-1}, u_{k-1}, w_{k-1}) \\ y_k &= q(C_k x_k) \end{aligned} \tag{5.2}$$

Absence of additive noise in measurement is an ideal model which, in a strict sense, does not apply to any practical systems with analog measurement. Theoretically, it is still meaningful to study this case since it presents the opportunity to study the limit

version of filtering from quantized noisy measurements when the additive noise goes to zero. As the additive noise in measurement goes to zero, QMSMC becomes increasingly difficult to implement. The difficulty comes from singularity in the likelihood evaluation step. QMSMC uses the distribution of measurement noise to evaluate the likelihood of each prior sample based on received quantized measurement of current step k, $\mathbf{y}_k$. When the variance of the additive noise in measurement becomes small compared with that of $C_k x_k$, for a given sample size, it becomes more frequent that all the prior samples generated at step k will have almost zero likelihood. Theoretically, the algorithm can still continue even the likelihoods are almost zero. However, in implementation, once a likelihood is sufficiently small, it has to be treated as zero because of limited codeword length in digital computation. Singularity refers to the case that in implementation of QMSMC, all prior samples have zero likelihoods based on currently received quantized measurement.

Singularity will cause QMSMC algorithm to stop and certain procedures have to be adopted to keep QMSMC continuing. Besides halting QMSMC, singularity itself indicates significant error in state estimation. We are unable to study singularity and its effect on performance of QMSMC theoretically. It is a consensus within SMC literature that systems with weak measurement noise poses challenges to SMC. In our case, it poses challenges to QMSMC.

The limit of decreasing measurement noise is its absence. Conceptually, absence of measurement noise makes the filtering problem easier than the one with noisy measurement since the Bayesian update part in the Fokker-Planck and Bayesian framework only needs to set the propagated probability density function to be zero outside the corresponding quantization region((or the measure of the complement set of corresponding quantization region to be zero) and then re-normalizing the function. In theory, QMSMC can still be applied to those models. The likelihood evaluation step would produce either 1 or 0. However, frequent singularity makes the algorithm much less practical than when being applied to noisy measurement.

Consider the following discrete-time state space model:

$$\begin{aligned} x_k &= A_{k-1}x_{k-1} + B_{k-1}w_{k-1}) \\ y_k &= q(C_k x_k) \end{aligned} \tag{5.3}$$

We begin our study of filtering for this model with linear time invariant scalar system. Though lack of realistic complexity, linear time invariant scalar systems provides a successful application of Fokker-Plank and Bayesian update scheme which conceptually solves this

problem.

Then, we focus on two ad hoc methods. The first one is modified Quantized Measurement Kalman Filter which is essentially the QMKF used for noisy measurement except in treatment of measurement noise.

The second one is Quantized Measurement Conditional Sampling(QMCS) algorithm which is a sample based method similar with QMSMC. We add projection step and replace the likelihood evaluation and resampling steps of QMSMC by conditional sampling based on quantized measurement. The conditional sampling step maintains a certain sample group size and those samples are used to do state estimation. It is easy to understand that the new sample would be close to the real a posterior distribution in some sense. However, we are unable to prove it theoretically and exact number of trials required in conditional sampling step is unknown.

5.2 Scalar System

We focus on the following Itô equation.

$$\dot{x} = ax + b\dot{w} \tag{5.4}$$

The evolution of probability density function of x at time t, $\rho(t,x)$, is governed by Fokker-Planck equation as the following:

$$\rho_t = \frac{b^2}{2}\rho_{xx} - ax\rho_x - a\rho$$

where

$$\rho_t \triangleq \frac{\partial \rho}{\partial t}, \rho_{xx} \triangleq \frac{\partial^2 \rho}{\partial x^2}, \rho_x \triangleq \frac{\partial \rho}{\partial x}$$

with boundary condition

$$\rho(x,0) = \rho_0$$

Adding quantized measurement to (5.5), we have the following system:

$$\begin{aligned} \dot{x} &= ax + b\dot{w} \\ y &= q(cx) \end{aligned} \tag{5.5}$$

where $q : \mathcal{R} \to \mathbf{Z}_M$ is a scalar quantizer.

Discretizing the system with sufficiently small time interval dt, we can see that Fokker-Planck partial differential equation can be numerically approximated by

5.2.1 Propagation and Update for discrete time

5.2.2 Simulation Results

5.3 QMKF(QMEKF) for Quantized Noiseless Measurement

Consider the following state space model:

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k w_k + G_k u_k \\ y_k &= q(C_k x_k) \end{aligned} \tag{5.6}$$

where A_k, B_k and C_k are matrices with proper dimensions. System noise w_k is white gaussian process with known covariance and mean for each time index k.

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k w_k + G_k u_k \\ y_k &= C_k x_k \\ z_k &= C_k x_k + n(C_k x_k) \end{aligned} \tag{5.7}$$

From Assumption 3.3.1, to apply Kalman filter, we need to know the variance of $n(C_k x_k)$ at each time step. Variance of $n(C_k x_k)$ depends on quantizer q, inverse mapping i and the distribution of $C_k x_k$ at time index k. Suppose the distribution of $o_k = C_k x_k$ is continuous and has probability density function f.

Suppose quantizer q is given, then, to minimize variance of $n(C_k x_k)$

Assume the quantizer to be time variant. At each time index, the conditional variance and mean can be represented as.

QMKF for the model is stated as below. It includes two steps in each iteration Prediction, Update.

Algorithm 5.3.1. *1.* **Prediction :**

$$\hat{x}_{k/k-1} = A_{k-1}\hat{x}_{k-1/k-1} + G_{k-1}u_{k-1}$$

$$\Sigma_{k/k-1} = A_{k-1}\Sigma_{k-1/k-1}A_{k-1}^T + B_{k-1}Q_{k-1}B_{k-1}^T$$

2. **Update :**

$$K_k = \Sigma_{k/k-1}C_k^T(C_k\Sigma_{k/k-1}C_k^T + R_q(y_k))^{-1}$$

$$\Sigma_{k/k} = \Sigma_{k/k-1} - K_k C_k \Sigma_{k/k-1}$$

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k(z_k - C_k\hat{x}_{k/k-1})$$

where $R_q(y_k)$ *is the conditional variance of* $n(C_kx_k + v_k)$ *conditioned on* $q(C_kx_k + v_k) = y_k$.

This algorithm is different from QMKF for quantized noisy measurement only in the step involving measurement noise. The variance of quantization noise is absent in the update part.

5.4 Quantized Measurement Conditional Sampling for State Estimation

Consider the following linear state space model:

$$\begin{aligned} x_{k+1} &= A_kx_k + B_kw_k + G_ku_k \\ y_k &= q(C_kx_k) \end{aligned} \tag{5.8}$$

The filtering problem for this system is as follows:

5.4.1 Algorithm

The Quantized Measurement Conditional Sampling algorithm provides a state filter for system. In order to state the algorithm, we first introduce definition of boundary between two quantization regions.

1. **Initiation :**
   Draw a set of n independent samples from the distribution of x_0, denoted by $\{\mathbf{x}_0^i\}_{i=1}^n$.

2. **Iteration :**
   For step k,

   (a) Projection:
   Obtain $\mathbf{y}_k$. If $\mathbf{y}_k \neq \mathbf{y}_{k-1}$, project $\mathbf{x}_{k-1}^i$ onto the boundary of Cell $Q_{\mathbf{y}_k}$. Denote it by $\partial(Q_{\mathbf{y}_k})$. By definition of boundary, $\partial(Q_{\mathbf{y}_k})$ is closed. Projection means to replace $\mathbf{x}_{k-1}^i$ by the closest element in $\partial(Q_{\mathbf{y}_k})$.

   $$\mathbf{x}_{k-1}^i = arg \min_{p \in \partial(Q_{\mathbf{y}_k})} ||\mathbf{x}_{k-1}^i - y||$$

   $||\cdot||$ denotes Euclidean distance.

(b) Conditional Sampling:

i. Set number counter $s = 1$.

ii. Uniformly choose $j \in \{1, 2, \cdots, n\}$. Then, choose $\mathbf{x}_{k-1}^{j}$ from set $\{\mathbf{x}_{k-1}^{i}\}_{i=1}^{n}$.

iii. Sample $\mathbf{w}_{k-1}$ from distribution of w_k. Obtain trial prior sample $\mathbf{x}_k^s$ by

$$\mathbf{x}_k^s = A_{k-1}\mathbf{x}_{k-1} + B_{k-1}\mathbf{w}_{k-1}$$

iv. If $q(C_k\mathbf{x}_k^s) = \mathbf{y}_k$, accept $\mathbf{x}_k^s$. Set $s = s + 1$. Otherwise reject $\mathbf{x}_k^s$, go to (iii).

v. If $s < n$, go to (ii).

(c) Estimation:

The state estimation at step k is obtained as the following:

$$\hat{x}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_k^s$$

(d) $k = k + 1$, go to (a).

5.4.2 Simulation

5.4.3 Discussion

Chapter 6

Optimization in Quantization

Among previous discussions of filtering problem from quantized measurement, we implicitly assumed the quantizer under consideration is a given one within a specific class of quantizers. Various methods are proposed to provide asymptotically optimal or suboptimal solutions to the problem.

Assuming each entry in the codebook of a quantizer is uniformly distributed, the size of codebook in a quantizer will determines the communication required to transmit a codeword. Effort in investigating the impact of limited communication on state estimation will inevitably bring up the question that what is the lower bound of communication required to achieve certain task in the system?

Since different communication protocols and tradeoffs between precisions and time delay will be involved in this system, we just focus on one building block in this chapter. Our central theme is how to quantitatively measurement the efficiency of a quantizer with a given codebook size for our purpose. What is the optimal quantizer and

6.1 Definition of Optimality

Essentially, this is an optimization problem with respect to a special class of functions, i.e. quantization functions.

6.2 Optimal Quantizer for Standard Normal Distribution

6.3 Optimal Quantizer for Gauss-Markov Systems

Previous section studied the optimal quantizer for standard normal distribution. A natural and more interesting question next is what is the optimal quantizer for a Gauss Markov process?

In order to proceed, we first need to define the optimality of the quantizer.

We only study second order Gauss-Markov process which can be modelled as a system without observation noises. The reason for this is...

Consider second order LTI system with quantized measurement as below,

$$dx = Axdt + Bdw$$

$$y = q(Cx)$$

where $x = \begin{bmatrix} x1 \\ x2 \end{bmatrix}$. A, B, C have proper dimensions. w denotes Wiener process. q denotes the quantizer in measurement model. $q : \mathcal{R} \to \mathbf{Z}_M$, in which M denotes the codebook size of q. We only consider noiseless observation before quantization and scalar quantization. Thus, C is 1 2 matrix. We only consider systems where A = 0 1 1 b B = 0 1 C = 1 0 As usual, dene damping ratio of the system to be = b 2 . Within the class of systems we are considering, damping ratio characterizes to which extent the system is oscillatory and characterizes the autocorrelation function of the observation process y(t). Discretize system (1) with time step dt, we obtain discrete LTI system as below 1 xk = Adxk1 + Bwk1 yk = q(Cxk) (2) where Ad = exp(Adt) This system can be rewritten as xk = Adxk1 + Bwk1 yk = Cxk + Q(CXk) (3) where Q(CXk) = q(Cxk)Cxk is the quantization noise sequence. Recall that Quantized Measurement Kalman Filter(QMKF) assumes Q(CXk) to be a white process. Problem 1 Find the optimal quantizer q in Least Mean Squared Error(LMSE) sense for system (3) using QMKF.

Choosing QMKF as the suboptimal filter.

8level quantizer.

The strategy for numerical examination.

As studied in Report 42, Q(CX_k) is close to white process in the sense that its autocorrelation function's second largest value is smaller than 10% of the largest value for the systems we studied. It is also close to be uncorrelated with Cx_k since correlation function

between $\mathrm{Q}(CX_k)$ and CX_k is weak in a defined sense. If $\mathrm{Q}(CX_k)$ is white process uncorrelated with CX_k, then, the quantizer which minimizes the variance of $\mathrm{Q}(CX_k)$ will result in LMES for system (3) using QMKF. That quantizer is the optimal quantizer for the static equilibrium distribution of y_k. However, the assumption that $\mathrm{Q}(CX_k)$ is a white process uncorrelated with CX_k is not perfect. So, the optimal quantizer we are seeking should be different from the optimal quantizer for static equilibrium distribution of y_k. Since $\mathrm{Q}(CX_k)$ is close to white process in the sense defined before, we expect that the optimal quantizer in LMES sense should also be close to the optimal quantizer for static equilibrium distribution of y_k in some distance dense. So, to numerically search for the optimal quantizer, we start with the associated optimal quantizer for static equilibrium distribution of y_k. Here, we only consider $\mathrm{M} = 8$. From numerical computational by Lloyd algorithm, we know that the optimal quantizer for static equilibrium distribution of y_k is symmetric(I haven't found an argument for this theoretically). Denote the set of boundary points by $[c, b, a, 0, a, b, c]$

Assumption 6.3.1. *Assume the optimal quantizer sought in problem 1 is symmetric with respect to origin, i.e. its boundary points have the structure* $[c, ba, 0, a, b, c]$ *and Mean Squared Estimation Error function,* $E(a, b, c)$*, is a smooth function of* a, b, c*.*

By the continuity assumption,

$$E(a,b,c) = \frac{\partial E}{\partial a}|_{a_0}(a-a_0) + \frac{\partial E}{\partial b}|_{b_0}(b-b_0) + \frac{\partial E}{\partial c}|_{c_0}(c-c_0) + h.o.t$$

Then, to obtain the optimal quantizer numerically, we numerically determine $E(a, b, c)$ for given a, b, c.

We simulate Quantized Measurement Kalman Filter for the system and estimate the mean-squared-error in state estimation for one simulation. Since the mean-squared-error for one simulation is itself a random variable, we repeat the simulation N times, obtaining an estimate of its standard deviation, σ. Assuming each simulation is independent with each other, the mean of these N mean-squared-errors has standard deviation $\frac{\sigma}{\sqrt{N}}$. When N is large enough, the standard deviation can be reduced significantly so that we can get an more accurate estimate of mean-squared-error for this particular system and quantizer.

Modifying the parameter a in the quantizer with b, c held constant, we can obtain noisy observation of mean-squared-error function for different as. When the change in a is small(by experience), we can assume the mean-squared-error function is linear in a neighborhood of some constant a_0. Using linear regression, we can obtain an estimate of the partial derivative of the mean-squared-error function with respect to a at a_0.

To verify the linear regression is appropriate, we can calculate the standard deviation of error by using linear regression. If the error is close to the standard deviation of those observations, then, we have confidence in its validity. However, how to quantitatively measure the confidence is not clear to me right now.

We can apply the same method to b around b_0 and c around c_0.

6.4 Results

Bibliography

Crisan, D. (2001). *Sequential Monte Carlo Methods in Practice*, chapter 2, pages 17–41. Springer-Verlag.

Crisan, D., and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746.

A. Doucet, N. de Freitas, and N. Gordon, editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

Folland, G. B. (1999). *Real Analysis, Modern Techniques and Their Applications*. John Wiley Sons, Inc, second edition.

Gavrilets, V. (2003). *Autonomous Aerobatics Maneuvering of Miniature Helicopters*. PhD thesis, MIT.

Gersho, A., and Gray, R. M. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers.

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings-F.*, 140(2):107–113.

Haykin, S. (1988). *Digital Communications*. John Wiley and Sons, Inc.

Liu, J., and Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of American Statistical Association*, 93:1032–1044.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8).

Wong, W. S., and Brockett, R. W. (1997). Systems with finite communication bandwidth constraints - part i: State estimation problems. *IEEE Transactions on Automatic Control*, 42(9):1294 – 1299.