IMPORTANCE SAMPLING METHODS WITH MULTIPLE SAMPLING DISTRIBUTIONS

BY WENTAO LI

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Dr. Rong Chen and Dr. Zhiqiang Tan

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

Importance Sampling Methods with Multiple Sampling Distributions

by Wentao Li

Dissertation Director: Dr. Rong Chen and Dr. Zhiqiang Tan

The complexity of integrands in modern scientific, industrial and financial problems increases rapidly with the development of data collection technologies. Monte Carlo method is widely used for complicated integration. In Monte Carlo integration, it is a natural and flexible method to consider multiple simulation mechanisms instead of one to address different aspects of the integrand. New methods are needed to combine the multiple mechanisms efficiently.

Monte Carlo integration methods are reviewed, with focus on importance sampling methods (IS) and sequential Monte Carlo methods (SMC). The former is commonly used for low-dimension problems. The latter is a variation of IS, which has been developed to be a new branch itself in the recent two decades, and promising for high-dimension problems with sequential nature.

For IS, techniques for combining multiple proposal distributions have been well developed, including Owen and Zhou (2000) and Tan (2004). Important implementation issues are needed to be resolved, including the allocation of sample budgets and the selection of proposals. A two-stage procedure is proposed to optimize the sample allocation, and although little theoretical investigation has been done for such a two-stage

procedure in literatures, its optimality among current approaches is theoretically justified. The choice of the first stage sample size is also discussed through investigating the high order performance of estimators. About the construction of proposals, suggestions are given to approximate the perfect case.

For SMC, only the plain vanilla combination of multiple proposals has been used in literatures. A novel SMC filtering scheme is proposed to combine the multiple proposals through the control variates approach in Tan (2004). Control variates are used in both resampling and estimation. The new algorithm is shown to be asymptotically more efficient than the direct use of multiple proposals and control variates. The guidance for selecting multiple proposals and control variates is also given. Numerical studies of the AR(1) model observed with noise and the stochastic volatility model with AR(1) dynamics show that the new algorithm can significantly improve over the bootstrap filter and auxiliary particle filter.

Acknowledgements

I would like to express my gratitude to my advisors, Dr. Rong Chen, for his encouragement, ideas and guidance, and to Dr. Zhiqiang Tan for his continuous help and consideration in the past three years. Without their supervision and support, the completion of this dissertation would not have been possible.

I am greatly indebted to the department of Statistics at Rutgers University for the continuous financial support throughout my PhD study. Without the support, I would not have had the opportunities that I have had to date.

I would like to thank Dr. Minge Xie and Dr. Xiaodong Lin for their time and efforts to be my committee members. I would also like to thank Dr. John Kolassa for providing many valuable advices for the graduate study and life.

Finally, I would like to thank my wife, Yejing Liu, for her continuous help and support over my whole graduate study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistical Integration

Since the introduction of calculus by Newton and Leibnitz, the evaluation of integrals stays in the central role of many science and engineering problems. Without exception, the proper interpretation and evaluation of integrals is the key to many fundamental problems of statistics. For examples, the cumulative distribution function and moments, which characterize and sometimes identify the probability distribution, are in the form of integrals; the posterior density, which is the central topic in Bayesian statistics, requires to integrate the joint density. Integrals in statistical problems are usually in the following form:

$$\mu = \int_{\Omega} h(x)\pi^*(x)dx$$

where $\pi^*(x)$ is a probability density with domain Ω and $h(x)$ is a real function. μ can be treated as the expectation of $h(x)$ with respect to density $\pi^*(x)$. Even some statistical integrals that seldom interpreted as expectations can be written in this form. For examples, the standard normal CDF at x_0 can be treated as the normalizing constant of standard normal density truncated in $(-\infty, x_0)$; the posterior density can be treated as the expectation of likelihood function with prior density.

In most practical problems, integrals are evaluated by numerical approximation, since only a few simple functions can be integrated analytically with techniques taught in the college calculus course. The approximation is usually in the form of weighted average of integrands evaluated at multiple points $\{x_1, \cdots, x_n\}$, such as the midpoint rule for univariate integration. In other words, the underlying Lebesgue measure is approximated by discrete counting measure with $\{x_1, \cdots, x_n\}$ as support and weights

as measures. Based on the choice of $\{x_1, \cdots, x_n\}$, approximation methods have two categories: deterministic method and Monte Carlo method. The former makes the selection in deterministic way, including Newton-Cotes rules, Gaussian quadrature, quasi Monte Carlo method and etc. See Press et al. (2007) for an overview. The latter generates $\{x_1, \cdots, x_n\}$ from some probability distributions, including MCMC, importance sampling, acceptance-rejection methods and etc. See Robert and Casella (2004) for an overview. None of them has overwhelming advantages over the other. The deterministic method usually takes into account the analytical characteristic of the integrand, e.g. gradient or Lipschitz continuity, and can give more accurate results for regular and small dimension problems (Geweke, 1996). But as a result, strong assumptions are needed for the integrand, and if multiple integrals are of interest, separate implementations are needed. In contrast, the Monte Carlo method usually involves less or no analytical characteristics of the integrand and may be inferior in analytically tractable problems. But due to its milder assumptions, it is more robust for non-regular or high-dimension problems (Geweke, 1996). For multiple integrals, Monte Carlo method can also be designed in flexible way so that one batch of random sample can be applied to different integrals. For more comparison and examples, see Robert and Casella (2004) and Caflisch (1998).

Due to its probabilistic nature, Monte Carlo method can be designed specifically to consider the statistical aspect of the integral. For the target integral μ, the basic idea of Monte Carlo integration is to generate random sample $\{x_1, \cdots, x_n\}$ from $\pi^*(x)$ and approximate μ by

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \tag{1.1}$$

where the average converges to μ by the law of large number. From the expression of μ, it is easy to see that its value is mostly contributed by integrating the area where the product $h(x)\pi^*(x)$ is large. Since the majority of $\{x_1, \cdots, x_n\}$ falls in the high density area of $\pi^*(x)$, the average avoids to evaluate $h(x)$ in the less important area for the integral. In the case that the support of $h(x)$ overlaps with the low density area of $\pi^*(x)$, such as evaluating tail probability, importance sampling method can

be employed to generate $\{x_1, \cdots, x_n\}$ from density emphasizing the support of $h(x)$ instead of $\pi^*(x)$. Compared to the grid-based deterministic methods, Monte Carlo method can be adapted better to the practical background of problems and therefore is well accepted by practitioners like physicists, systems engineers and statisticians. In addition, the Monte Carlo method requires much less sophisticated mathematics compared to the deterministic methods and is straightforward to be implemented given developed random number generators (Caflisch, 1998).

1.2 State Space Form

Originated in engineering, the state space form is used to model dynamic systems with time-varying inputs and outputs. Examples of input-output pair in a dynamic system include the market volatility and stock price, original and received signal in wireless communication, running speed and position in a real-time tracking, and many others. Assume x_t is the underlying signal input at time t, y_t is the measurable output, and the system runs from time 1 to n. Let $x_{1:t} = \{x_1, \cdots, x_t\}$ and $y_{1:t} = \{y_1, \cdots, y_t\}$. The state space form contains two equations, with one modeling the signal process $x_{1:t}$ and the other modeling the measurements $y_{1:t}$, as following:

$$x_t = F(x_{1:t-1}, w_t) \text{ or } x_t \sim f(\cdot|x_{1:t-1}),$$
$$y_t = G(x_{1:t}, v_t) \text{ or } y_t \sim g(\cdot|x_{1:t}),$$

where w_t is the innovation, v_t is the measurement noise and all are independent (Chen, 2005). At time t, given the previous signals, x_t is evolved through the equation F or the conditional density f, and the output of system is measured with y_t through G or the conditional density g.

In practice, the Markovian state space form is usually employed:

$$x_t = F(x_{t-1}, w_t) \text{ or } x_t \sim f(\cdot|x_{t-1}), \tag{1.2}$$
$$y_t = G(x_t, v_t) \text{ or } y_t \sim g(\cdot|x_t), \tag{1.3}$$

where the underlying state x_t is a Markov process and the distribution of the observation y_t can be determined solely by the current signal. Not only because of the simplicity, the widespread application of the Markovian form is also due to the Markov chain observed with noise structure well described in many real problems. For example, in a target-tracking problem, the actual position and speed are underlying states and the noisy position in the device is the observation. The current position only depends on the position and speed of last time point, and the device only measures the current position (Haug, 2012). In the capture-recapture problem in population study, suppose the moving pattern of some species is of interest. Their resident location is the underlying state and the capture record is the observation. Then their current resident only depends on the previous resident and the capture record only depends on their current location (Dupuis, 1995). Another well-known terminology, "Hidden Markov Model", also describes the same Markovian structure, and is often used when the state takes discrete values (Cappé et al., 2005). In some cases, a non-Markovian signal process can be reparameterized to be have the Markovian structure. In Fearnhead (1998), the AR(2) model is reformulated to have AR(1) structure, therefore the noisy AR(2) model has the following equivalent state space forms:

$$\begin{cases} x_t + a_1 x_{t-1} + a_2 x_{t-2} = \epsilon_t \\ y_t = b x_t + \eta_t \end{cases} \Longleftrightarrow \begin{cases} \Theta_t = A\Theta_{t-1} + W_t \\ y_t = B\Theta_t + \eta_t \end{cases},$$

$$where\ \Theta_t = \begin{pmatrix} -a_2 x_{t-1} \\ x_t \end{pmatrix},\ W_t = \begin{pmatrix} 0 \\ \varepsilon_t \end{pmatrix},\ A = \begin{pmatrix} 0 & -a_2 \\ 1 & -a_1 \end{pmatrix}\ and\ B = (0,\ 1).$$

Another example which is about the blind deconvolution of wireless communication can be seen in Miguez and Djuric (2002).

1.2.1 Integration Problems in State Space Form

The main goal of state space form is to make inference for the unobservable state x_n conditional on the given observations $\{y_1, \cdots, y_t\}$. Intuitively, it is natural to select the conditional density $p(x_n|y_{1:t})$ as the inference target. From the perspective of Bayesian analysis, if the state equation (1.2) is treated as the prior information of x_t, $p(x_n|y_{1:t})$ is

the posterior density and the posterior mean is the Bayesian solution of minimizing the mean square error of estimating x_n with $y_{1:t}$ available (Fearnhead, 1998). Depending on the observations given, the inference tasks can be divided into three categories, and the target posterior density of each is closely related to the joint posterior density $p(x_{1:t}|y_{1:t})$ (Chen, 2005).

1. Filtering: The aim is to update the knowledge of state when new observations come in, i.e. $t = n$. The posterior density $p(x_t|y_{1:t})$ can be obtained by marginalizing $p(x_{1:t}|y_{1:t})$ over $x_{1:t-1}$;

2. Smoothing: The aim is to estimate the previous state with all available observations, i.e. $t > n$. The posterior density $p(x_n|y_{1:t})$ can be obtained by marginalizing $p(x_{1:t}|y_{1:t})$ over $x_{n+1:t-1}$ and $x_{1:n-1}$;

3. Predicting: The aim is to predict the future state with currently available observations, i.e. $t < n$. Let $n = t+k$. The posterior density $p(x_n|y_{1:t})$ can be obtained by marginalizing

$$\begin{aligned} p(x_{1:t+k}|y_{1:t}) &= p(x_{1:t}|y_{1:t})p(x_{t+1:t+k}|x_{1:t}, y_{1:t}) \\ &= p(x_{1:t}|y_{1:t}) \prod_{i=1}^{k} p(x_{t+i}|x_{1:t+i-1}, y_{1:t}) \\ &= p(x_{1:t}|y_{1:t}) \prod_{i=1}^{k} p(x_{t+i}|x_{t+i-1}), \end{aligned}$$

   where the last equation holds by the Markovian property in (1.2).

Therefore, the treatment of $p(x_{1:t}|y_{1:t})$ stays the central role of these three tasks. The related work in this thesis only considers estimation with $p(x_{1:t}|y_{1:t})$ and the filtering task. For treatments on the other two tasks, see Durbin and Koopman (2012) and Cappé et al. (2005). Kitagawa (1996), Briers et al. (2010) and Doucet and Johansen (2009) also give some reviews of the smoothing task. In this thesis it is assumed that the state and observation densities f and g are fully known, i.e. inference is only needed to be made on x_t. The inference problem for unknown parameters of f and g is totally different from filtering. Since unknown parameters can be treated separately, one can

first estimate parameters then apply any filtering method in this thesis with estimated parameters. For treatment of parameter estimation, see Liu and West (2001), Cappé et al. (2007), Pitt and Shephard (1999) and Gilks and Berzuini (2001).

The target integral of filtering is the posterior expectation of some real function $h(x_t)$, i.e. $E[h(x_t)|y_{1:t}]$. While in most cases the target function h has only x_t as the argument, the algorithms in later chapters can be applied to more general cases. Therefore we assume h takes $x_{1:t}$ as arguments. Also for the ease of representation, assume there is an initial state x_0 following $p(x_0)$ and the integration is over $x_{0:t}$, i.e. the target integral is

$$E[h(x_{1:t})|y_{1:t}] = \int h(x_{1:t})p(x_{0:t}|y_{1:t})dx_{0:t}. \tag{1.4}$$

The integration requires evaluation of the joint posterior density, and by standard Bayesian theorem,

$$p(x_{0:t}|y_{1:t}) \propto p(x_{0:t})p(y_{1:t}|x_{0:t}), \tag{1.5}$$

where "$\propto$" means "proportional to" and the scale is a constant with respect to $x_{0:t}$. Since t is usually very large, it would be expensive to evaluate the density jointly hence alternative evaluation is needed.

1.2.2 Recursive Solution

With the standard conditional theorem, the RHS of (1.5) can be expanded as following:

$$\begin{aligned} p(x_{0:t})p(y_{1:t}|x_{0:t}) &= p(y_t|y_{1:t-1}, x_{0:t})p(x_t|y_{1:t-1}, x_{0:t-1})p(y_{1:t-1}, x_{0:t-1}) \\ &= p(x_0)\prod_{k=1}^{t} p(y_k|y_{1:k-1}, x_{0:k})p(x_k|y_{1:k-1}, x_{0:k-1}) \end{aligned}$$

By the Markovian properties in (1.2) and (1.3), the last expression above can be simplified and (1.5) is equivalent to

$$p(x_{0:t}|y_{1:t}) \propto p(x_0)\prod_{k=1}^{t} p(y_k|x_k)p(x_k|x_{k-1}), \tag{1.6}$$

where both $p(y_k|x_k)$ and $p(x_k|x_{k-1})$ are known in the models. The sequential expression has two benefits. Firstly, it is computationally easier to be programmed and evaluated. Secondly, since at time t the value is equal to the product of densities of x_t and y_t and the value at time $t-1$, it can be calculated recursively as new observations come in, which is an ideal expression for the filtering task.

1.3 Outline of Thesis

For integration problems mentioned in previous sections, Monte Carlo method is heavily employed in the practical computation due to the benefits indicated in Section 1.1. Among the many developed Monte Carlo algorithms, importance sampling (IS) is a classical scheme, dated back to the era of the first modern computer (Kahn, 1949; Kahn and Harris, 1949), and has been employed by many practitioners since then. For a low-dimension problem, with a properly designed sampling distribution, which is called the proposal distribution, the standard importance sampling can have promising performance. The particle filter method is a high-dimension variation of importance sampling, with initial idea in Rosenbluth and Rosenbluth (1955) and formally introduced in Gordon et al. (1993). Initially it was designed for the state space form and now can accommodate wider range of high-dimension problems (Del Moral et al., 2006). It has been developed to be a whole new branch of computational methods, generally called sequential Monte Carlo (SMC).

Section 2 reviews these two main Monte Carlo integration methods and focuses on the design of proposal distribution which is a central implementation issue. The design of a satisfactory sampling mechanism has several requirements which pose difficulties in practice. The practice of considering multiple proposal distributions is reviewed, including many literatures of IS and SMC. Then for combining multiple proposals, the implementation issues in IS and limitations in SMC of current approaches are summarized, according to which new methods are proposed in later chapters to make significant improvement. Section 3 reviews the historical approaches to combine multiple proposals in IS and SMC and their limitations in details.

For the state-of-art approaches in IS including, including Owen and Zhou (2000)'s regression estimator and Tan (2004)'s MLE estimator, the main implementation issues are the sample allocations for multiple proposals and the selection of proposals. Section 4 proposes a two-stage procedure to optimize the sample allocation and investigates its theoretical properties. At the first stage, the optimal mixture proportions of the sampling proposals are estimated using pilot sample; at the second stage, formal sample is generated according to the optimal mixture proportions and the estimator is constructed using samples from both stages. It is shown that the two-stage estimator achieves the best performance among the up-to-date approaches. The suggestions on constructing proposals to approximate the perfect case are also given.

When designing the proposal distributions in SMC, how to handle the multimodality of target distribution or how to control the tails of proposals often remains unsolved. Section 5 proposes a novel algorithm to combine multiple proposals into SMC methods through the control variate approach in Tan (2004), which is called the likelihood approach, so that the previous issues can be easily handled by including multiple proposals addressing different aspects. It is shown that the likelihood approach has the exclusive benefit that control variates can be included in the resampling step, which makes the algorithm makes significant improvement over the standard algorithms. We also give the suggestions on constructing proposals and control variates to approximate the perfect case, making the new algorithm practical to use.

Chapter 2

Monte Carlo Integration

Consider the target integral μ and the Monte Carlo estimator (1.1), the key problem is how to simulate from the target density $\pi^*(x)$. If the inverse CDF of $\pi^*(x)$ is available, the simulation can be done by standard inverse transformation method, or accept-reject method if appropriate instrumental density $q(x)$ can be found so that the ratio $\pi^*(x)/q(x)$ is bounded by a not too large constant (Robert and Casella, 2004). For many non-regular or high-dimension target densities, it is difficult to implement these two methods. The importance sampling (IS) and the sequential Monte Carlo (SMC) are two classes of methods widely used in such more complicated situations. They generate sample from alternative densities instead of $\pi^*(x)$ and adjust the difference by assigning weights to observations. SMC is a method originating from high-dimension IS and has become a new branch itself in the recent two decades. The review below focuses in the selection of sampling mechanism which is critical to the performance of both methods.

2.1 Importance Sampling

The idea of IS for approximating μ is based on the identity

$$\int_\Omega h(x)\pi^*(x)dx = \int_\Omega \frac{h(x)\pi^*(x)}{q(x)} q(x)dx,$$

where $q(x)$ is a probability density, called the proposal density. Through the addition of $q(x)$, μ can be treated as the expectation with respect to density $q(x)$ instead of $\pi^*(x)$. With a sample $(x_1, \cdots, x_n)$ from $q(x)$, μ can be approximated by the sample average

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)w_i, \tag{2.1}$$

where $w_i = \pi^*(x_i)/q(x_i)$ is called the importance weight. IS is usually applied to the following two variations of μ:

(I) $Z = \int_\Omega \pi(x)dx$, where $\pi(x) \propto \pi^*(x)$ is an unnormalized density;

(II) μ, where only $\pi(x) \propto \pi^*(x)$ can be evaluated.

Z can be treated as the normalizing constant of $\pi(x)$. Although Z is a special case of μ, the estimator (1.1) does not work since $h(x)$ is an unknown constant, therefore it can not be estimated by methods generating sample from $\pi^*(x)$, such as MCMC, accept-reject method and etc. The need of calculating Z arises in many areas, including missing data analysis, marginal likelihood calculation, estimation of free energies in physics(Gelman and Meng, 1998), and communication system (Smith et al., 1997). The second form is frequently of interest in Bayesian analysis in many fields, where $\pi^*(x)$ is the posterior density, such as rare event simulation (Denny, 2001), reliability (Hesterberg, 1995), computational finance (Owen and Zhou, 1999) and computer graphics (Veach and Guibas, 1995). Since $\pi^*(x)$ can only be evaluated up to a constant, (2.1) does not apply. Instead, by expressing μ as

$$\mu = \frac{\int_\Omega h(x)\pi(x)dx}{\int_\Omega \pi(x)dx},$$

the ratio of two IS estimators, estimating the numerator and denominator respectively, can be used instead. It also converges to μ in proper conditions. This is called the ratio estimator.

Compared to the basic Monte Carlo, IS has great potential to achieve much better accuracy since it can take the target function $h(x)$ into consideration. One important application of IS is to estimate the rare event probabilities, such as the bit error rate of communication system Smith et al. (1997) and important market risk measures Hoogerheide and Van Dijk (2010). IS also benefits from the flexible design of $q(x)$ and mild conditions of convergence which makes it possible to be well adapted and robust in complex problems such as multimodality. Examples include Hesterberg (1995) which evaluates industrial strategies for public utility system and Binder and Heermann (2010) which calculates spectral densities of a physics system. Compared to MCMC,

IS is straightforward to be implemented and easy to be interpreted, since it generates i.i.d sample and the standard error can be easily estimated. See Fearnhead (2008) for more comparisons. A further advantage of IS is that through resampling the i.i.d sample according to w_i, IS can be used to generate sample from $\pi^*(x)$, which is called sampling importance resampling (SIR) (Rubin, 1988). In Bayesian problems, SIR is practical to simulate from the posterior distribution. For examples, see McAllister and Ianelli (1997) and Raftery et al. (1995). SIR is also a central component in SMC. See the introduction in the next section.

One criticism of IS is that the selection of the proposal density is too arbitrary and lack of considering important information of the integrand $h(x)\pi^*(x)$, such as derivatives (O'Hagan, 1987). Techniques are developed to combine IS with other Monte Carlo methods, including the control variates and antithetic variates, which can utilize the analytical form of the integrand. In Hesterberg (1996), the linear approximation of the integrand was selected as control variate. in Evans and Swartz (1995), Laplace approximation of the integral was used to construct the control variates. In Evans and Swartz (1995) and Evans and Swartz (1996), the antithetic variates method was generalized to involve multiple parametrization of the integrand and combined with IS.

The selection of the proposal density $q(x)$ is critical to the performance of IS. Different choices can result in the estimation variance ranging from 0 to infinity (Robert and Casella, 2004). Roughly speaking, a good $q(x)$ should satisfy several criteria. Firstly, the support of $q(x)$ should cover the support of the integrand. Secondly, in the high value support of the integrand, $q(x)$ should have high density so that the simulated sample can focus on the "important" areas for the integration. Thirdly, when Ω is not compact, $q(x)$ should decrease slower than the integrand in the tail areas, i.e. have "heavier" tail than the integrand. However, in practice it is usually challenging to design a satisfactory $q(x)$. One reason is that the complexity of the integrand makes it difficult to find a $q(x)$ that covers all its important parts. For example, when the integrand is multimodal, a unimodal $q(x)$ will not be efficient. Examples of multimodal integrand can be found in Owen and Zhou (1999). Another well known reason is that when $q(x)$ has a "lighter" tail than the integrand, i.e. $\pi(x)/q(x)$ or $h(x)\pi^*(x)/q(x)$ are

unbounded, IS estimator can have infinite variance. When there is a lack of knowledge of the integrand in some regions, unexpected large values of integrand may result in inaccurate results. See Ford and Gregory (2007) for an example.

For both problems, a general remedy is to consider multiple proposal distributions to address different aspects of the integrand. For multimodal integrands, Oh and Berger (1993) used a family of student's t distributions and Owen and Zhou (1999) used a family of beta distributions to model each mode of integrand individually. West (1993) and Givens and Raftery (1996) used a kernel estimate of the integrand as the proposal which is a mixture of normal or t distributions. Even for a unimodal target distribution, one can construct a mixture of two proposals where one mimics the center of target and the other dominates the tail. Such a construction was used in Giordani and Kohn (2010), although in a different scenario. The requirement that the tail of integrand needs to be dominated by the proposal distribution can be met by including some heavy-tailed distributions in the mixture as "protection". For example, Hesterberg (1995) included the target distribution itself as one of the components to provide an upper bound for the estimation variance, and Owen and Zhou (2000) used uniform distribution to bound the sample weights in a bounded domain case. Liang et al. (2007) divided the state domain into subregions and used the mixture of truncated target distributions in all subregions as the proposal, which leads to bounded importance weights. In Bayesian analysis, the prior density of the parameters can serve as the heavy-tailed component, as utilized in Ford and Gregory (2007). Other choices of heavy tail distributions can be found in Geweke (1989).

Given multiple potentially useful proposals, a straightforward combination method is to use their mixture as the new proposal. This method has two issues. One is that the mixture proposal may contaminate the good components in the mixture. Owen and Zhou (2000) shows that a mixture can lose efficiency by several orders of magnitude if the original proposal is nearly perfect. Another problem is that the mixture proportions need to be determined. Proper mixture proportions can increase the efficiency by an order of magnitude, as shown in Emond et al. (2001).

For the contamination problem, Owen and Zhou (2000) suggested a regression method to combine the mixture importance sampling proposal approach with some control variates. Control variate is a useful technique for variance reduction. For a review, see Rubinstein and Kroese (2008). Additional to variance reduction, Owen and Zhou's method has the property that it will not perform worse than using the best of component proposals individually, if the sample size assigned to it is the same as in the mixture case. Therefore, if half of the sample is assigned to the best proposal, the regression estimator's efficiency is at least half of the efficiency of using the best proposal's efficiency when it is used alone. Such a lower bound lessens the contamination problem. Tan (2004) proposed to use nonparametric maximum likelihood estimation in placed of regression and showed that the MLE method is the most efficient among several classes of estimators including those in Owen and Zhou (2000), Hesterberg (1995) and Veach and Guibas (1995). Some important implementation issues of Owen's regression method and Tan's MLE method are left to be discussed, including the determination of mixture proportions and the selection of proposals.

To determine appropriate proportions, Fan et al. (2006) and Hesterberg (1995) followed some heuristic rules derived from experience or interpretation of proposals. A more sophisticated approach is to use a pilot study to determine the optimal proportions via minimizing some criterion. The estimated proportions are then used to generate the sample and construct the estimators. The criterion was selected to be the asymptotic variance of IS estimator with mixture proposal in Raghavan and Cox (1998), and the variation coefficient of pilot sample in Oh and Berger (1993). However, few theoretical properties have been investigated.

2.2 Sequential Monte Carlo

In the dynamic system with state space form (1.2)-(1.3), the integration problem in filtering (1.4) is often of high dimension. Due to the curse of dimensionality, the high density area of the target distribution is like "a needle in a haystack"(Liu, 2008). Importance sampling in high-dimension problem can hardly focus the observations in the important area and suffers from heavy skewness of sample weights, i.e. the weights of

a few observations are much larger than all others, and therefore large estimation variance. The sequential importance sampling (SIS) (Liu et al., 2001) allows to design the proposal of IS sequentially according to the recursive expression (1.6) of target density $p(x_{1:t}|y_{1:t})$, but still can not avoid the weights degeneracy problem as t increases.

The particle filtering method, initially introduced in Gordon et al. (1993) as "bootstrap filter", implements the resampling regularly in the course of SIS, which mitigates the degeneracy of sample weights. The basic resampling method applies multinomial sampling on the samples with probabilities proportional to the sample weights, then drops out observations with small weights and duplicates observations with large weights. Therefore regularly implementing resampling can avoid too many samples have too small weights. Since then, the simulation-based methods for on-line filtering of dynamic systems are widely used in various fields, including target tracking (Chen and Liu, 2000), signal processing (Wang et al., 2002), estimation of economical model (Shephard, 2005) and counting contingency tables (Chen et al., 2005). Various efforts have been devoted to improve the basic particle filtering method, including improving the sampling mechanism (Doucet et al., 2006), increasing the diversity of samples (Gilks and Berzuini, 2001) and adaptively choosing the resampling schedule (Liu and Chen, 1995). Most of these techniques are unified in the sequential Monte Carlo framework (Doucet et al., 2000; Liu and Chen, 1998), with SIS with resampling in the central role, where SIS generates weighted particles from proposal distributions and resampling mitigates the degeneracy of sample weights. Reviews of the related techniques can be seen in Doucet and Johansen (2009), Cappé et al. (2007) and Chen (2005).

Design of proposal distributions is essential to the SMC methods. In Gordon et al. (1993), $p(x_t|x_{t-1})$ is used to generate particles at time t based on particles from time $t-1$, which is called bootstrap filter. The efficiency of bootstrap filter is usually limited since the sampling mechanism does not consider the information of observations y_t. When the system meets an observation outlier, since all information of observation is included in the sample weights, the particles may degenerate, i.e. the sample weights of a few particle dominate the others. But this method is popular due to its simplicity and computational efficiency. Another simple choice is the independent particle filter in Lin

et al. (2005) which uses $p(y_t|x_t)$ to generate particles based on y_t. It can be more efficient than bootstrap filter when the observational noise is small. The probability density $p(x_t|x_{t-1}, y_t)$ is known as the optimal proposal density, in the sense that the variance of its importance weights conditional on x_{t-1} is equal to 0. Intuitively, it includes the full information of both state and observation equations. But except in a few scenarios, the optimal proposal density and the corresponding sample weights are usually analytically unavailable due to the nonlinear form of the observation equation. Suboptimal choices include the probability densities constructed by approximating $p(x_t|x_{t-1}, y_t)$ through local linearization or moment approximation. See Doucet et al. (2000), Guo et al. (2005), Saha et al. (2009) and Pitt and Shephard (1999) for examples.

There are several limitations for the proposal distributions mentioned above. First, the approximation to $p(x_t|x_{t-1}, y_t)$ may not have bounded variance since the local approximation does not provide control on the tails of the proposal density. Second, the above approaches usually construct unimodal densities which are inefficient for a multimodal target density. Finally, although $p(x_t|x_{t-1}, y_t)$ includes both the information of state and observation equations, it does not consider the target function, which can make the filtering computationally expensive in some cases such as estimating the probability of rare event.

For the third limitation, specific cases such as estimating the tail probability have been discussed (Chan and Lai, 2011; Cérou et al., 2012). However, there is no guideline to deal with general target functions. For the first two limitations, a general remedy is to consider multiple proposal distributions which can include proposals for controlling tails or concentrating on multiple modes. While the usage of multiple proposals for importance sampling, which can be treated as a special case of SMC containing only one step filtering, has been well discussed in the literature as mentioned in the previous section, it is natural to consider this strategy for SMC. Meanwhile, since it has been shown in Tan (2004) and Owen and Zhou (2000) that combining multiple proposals with appropriate control variates can significantly increase the efficiency and decrease the contamination brought by mixing poor proposal distributions with good proposal distributions, control variates can also be considered in SMC. However, in

SMC framework, the usage of multiple proposals and control variates only receives limited discussions and is case-dependent. Elinas et al. (2006) and Fox et al. (2001) applied mixture proposal distribution in Monte Carlo localization of robot to combine information from camera or laser observations and the motion model. Singh et al. (2004) and Singh et al. (2007) applied the control variates to particle filter for target tracking sensor management, by replacing the target function $h(x_t)$ by $h(x_t) + \boldsymbol{\beta}^T \boldsymbol{g}(x_{1:t})$ where $\boldsymbol{g}(x_{1:t})$ is the vector of control variates and $\boldsymbol{\beta}$ is the coefficients.

Chapter 3

Review of Advanced Importance Sampling Methodologies

3.1 Importance Sampling with Multiple Proposals

Assume the target integral is in the form of (I) or (II) and $h(x)$ and $\pi(x)$ can be evaluated exactly. In this section we only consider estimating Z. The extension of estimating μ is straightforward for the first two methods, and will be discussed in the next chapter for the last two methods.

3.1.1 Mixture Importance Sampling

Assume observations $\{x_1, \cdots, x_n\}$ are taken i.i.d from a proposal distribution $q(x)$. The integral $Z = \int \pi(x)dx$ can be estimated by

$$\widehat{Z}_{IS} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(x_i)}{q(x_i)}. \tag{3.1}$$

Under mild conditions, the asymptotic variance is $Var_q[\pi(x)/q(x)]$ where Var_q is the variance under distribution $q(x)$ (Robert and Casella, 2004). The optimal proposal is $\pi(x)/Z$, suggesting that the proposal $q(x)$ should be chosen to mimic the shape of $\pi(x)$ so that the high and low density regions of $q(x)$ coincide with those of $\pi(x)$. With such a proposal, the majority of Monte Carlo sample from $q(x)$ fall in the high density region of $\pi(x)$, the importance region. In some scenarios, more than one $q(x)$ may be needed. For example, for a multimodal $\pi(x)$, it is helpful to use several proposal distributions, each targeted at one importance region. Suppose $q_1(x), \cdots, q_p(x)$ are p probability densities serving as proposals. Given a mixture proportion vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$ satisfying

$\sum_{k=1}^{p} \alpha_k = 1$, we can use the mixture distribution as the proposal and estimate Z by

$$\widehat{Z}_{MIS} = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(x_i)}{q_{\boldsymbol{\alpha}}(x_i)}, \tag{3.2}$$

where $q_{\boldsymbol{\alpha}} = \sum_{k=1}^{p} \alpha_k q_k(x)$ and $\{x_1, \cdots, x_n\}$ are generated from q_α. In addition, the variance of $\widehat{Z}_{IS}$ also demands that the ratio $\pi(X)/q(X)$ have a finite variance. A mixture distribution certainly makes it easier to satisfy such a condition as one can simply include a proposal distribution $q_1(X)$ having $\text{Var}[\pi(X)/q_1(X)] < \infty$, such as a uniform distribution if the domain is bounded. Such a proposal distribution sets an upper bound to the estimating variance and therefore plays the role of "safeguard" in importance sampling, which is the key idea of defensive importance sampling (Hesterberg, 1988).

3.1.2 Stratified Sampling

Instead of generating samples directly from the mixture distribution as that in (3.2), stratified samples $\{x_{k1}, \cdots, x_{kn_k}\}$ can be taken with deterministic size $n_k = \alpha_k n$ from the k-th proposal q_k, which leads to the estimator in Hesterberg (1988)

$$\widehat{Z}_{SIS}(\boldsymbol{\alpha}) = \frac{1}{n} \sum_{k=1}^{p} \sum_{i=1}^{n_k} \frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}. \tag{3.3}$$

Veach and Guibas (1995) consider the following estimator

$$\sum_{k=1}^{p} \frac{1}{n_k} \sum_{i=1}^{n_k} \omega_k(x_{ki}) \frac{\pi(x_{ki})}{q_k(x_{ki})}, \tag{3.4}$$

where $\{\omega_k(x)\}_{k=1}^{p}$ is a group of coefficient functions for the sample weights and satisfies $\sum_{k=1}^{p} \omega_k(x) = 1$. They showed that $\widehat{Z}_{SIS}$ is a suboptimal choice in this large class. Raghavan and Cox (1998) proposed a two-stage algorithm to construct $\widehat{Z}_{SIS}$ with estimated optimal mixture proportions in the sense of minimizing the asymptotic variance.

3.1.3 Importance Sampling with Control Variates

One problem of using a mixture proposal distribution is the possible loss of efficiency due to mixing of good proposal distributions with poor ones (Owen and Zhou, 2000). It

is a premium to pay for the insurance of valid importance sampling, but can be reduced by combining importance sampling and control variates. Given an unbiased estimator X_n of Z, improvement can be gained by constructing a proper control variate vector Y and using $X_n - \boldsymbol{\beta}^T(Y - E[Y])$ to estimate Z. The optimal $\boldsymbol{\beta}$ can be estimated using a regression approach to minimize asymptotic variance (Cochran, 1977). In Owen and Zhou (2000), combining $\widehat{Z}_{SIS}$ and control variates $\boldsymbol{g}(x) = (q_2(x) - q_1(x), \cdots, q_p(x) - q_1(x))^T$ results in the estimator

$$\widehat{Z}_{Reg}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}^T\boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}, \tag{3.5}$$

where

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \widetilde{Var}\left[\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right]^{-1}\widetilde{Cov}^T\left[\frac{\pi(X)}{q_{\boldsymbol{\alpha}}(X)}, \frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right],$$

and $\widetilde{Var}$ and $\widetilde{Cov}$ denote the pooled-sample variance and covariance. There are two appealing properties of $\widehat{Z}_{Reg}$. First, its asymptotic variance is zero when $\pi(x)$ is a linear combination of the proposals. Second, $\widehat{Z}_{Reg}$ has smaller asymptotic variance than every importance sampling estimator constructed solely with q_k with n_k samples, $k = 1, \ldots, p$. That is, $\widehat{Z}_{Reg}$ is always at least as good as the best one among the individual proposals.

3.1.4 Likelihood Approach

All previous integration methods directly approximate the target integrals. On the other hand, in Kong et al. (2003), Monte Carlo integration is treated as a statistical inference problem where the Monte Carlo sample serves as observations, the underlying measure in target integral, usually Lebesgue measure or counting measure, is treated as an unknown nonnegative measure, and the Monte Carlo sample is modeled using a semiparametric model. Then by nonparametric maximum likelihood, the unknown measure is estimated by a discrete measure with the Monte Carlo sample as support, and the target integral is estimated by the integration over the discrete measure. As an example, with $\{x_1, \cdots, x_n\}$ generated identically and independently from q_1 under Lebesgue measure, the model assumes that x_i is distributed as $q_1(x)d\nu / \int q_1(x)d\nu$ where ν is an unknown nonnegative measure. The nonparametric maximum likelihood

estimator of ν is

$$\widehat{\nu} \propto \frac{\widehat{P}(\{x\})}{q_1(x)},$$

where $\widehat{P}$ has the support on $\{x_1, \cdots, x_n\}$ with mass n^{-1} at each point. Then $Z = \int q(x)d\nu$ can be estimated by

$$\frac{\int q_1(x)d\widehat{\nu}}{\int q(x)d\widehat{\nu}} = \frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i)}{q_1(x_i)}$$

This is the same as the importance sampling estimator with proposal distribution $q_1(x)$.

Given multiple proposals $q_1, \cdots, q_p$ and control variates $\boldsymbol{g}(x)$, Tan (2004) proposed to restrict the measure ν in the set $\left\{\nu : \int q_k(x)d\nu = \int q_1(x)d\nu,\ k = 1, \cdots, p\right\}$. The nonparametric MLE of ν under such a restriction is

$$\widehat{\nu} \propto \frac{\widehat{P}\left(\{x\}\right)}{q_{\boldsymbol{\alpha}}(x) + \widehat{\boldsymbol{\zeta}}^T\boldsymbol{g}(x)}, \text{ where } \widehat{\boldsymbol{\zeta}} = \underset{\boldsymbol{\zeta}}{\text{argmax}}\sum_{k=1}^{p}\sum_{i=1}^{n_k} log\left[q_{\boldsymbol{\alpha}}(x_{ki}) + \boldsymbol{\zeta}^T\boldsymbol{g}(x_{ki})\right],$$

and the integral estimator is given by:

$$\widehat{Z}_{MLE}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widehat{\boldsymbol{\zeta}}^T\boldsymbol{g}(x_{ki})}. \tag{3.6}$$

It is shown that $\widehat{Z}_{Reg}$ is a first order approximation of $\widehat{Z}_{MLE}$ and hence has the same asymptotic efficiency (Tan, 2004). The estimator $\widehat{Z}_{MLE}$ also achieves the highest asymptotic efficiency among the class of estimators in the form of

$$\sum_{k=1}^{p}\frac{1}{n_k}\sum_{i=1}^{n_k}\omega_k(x_{ki})\frac{\pi(x_{ki}) - \boldsymbol{\beta}_k^T(x_{ki})\boldsymbol{g}(x_{ki})}{q_k(x_{ki})}, \tag{3.7}$$

where $\omega_1(x), \cdots, \omega_p(x)$ and $\boldsymbol{\beta}_1(x), \cdots, \boldsymbol{\beta}_p(x)$ satisfy that $\omega_k(x) = 0$ when $q_k(x) = 0$, $\sum_{i=1}^{p}\omega_k(x) = 1$ and $\sum_{i=1}^{p}\omega_k(x)\boldsymbol{\beta}_k(x) = \boldsymbol{b}$ for some constant vector $\boldsymbol{b}$, and therefore dominates the class of estimators in (3.4).

Because $\widehat{Z}_{Reg}$ and $\widehat{Z}_{MLE}$ asymptotically dominate the other estimators, we will only discuss the two stage procedure for these two estimators. Furthermore, there is another important benefit of using $\widehat{Z}_{Reg}$ and $\widehat{Z}_{MLE}$ in that their asymptotic variance is a convex function of $\boldsymbol{\alpha}$ and hence can be easily minimized. See remark 4.1.

3.2 SMC for Filtering

3.2.1 Basic SMC Method

For the state space model (1.2)-(1.3), at time n, denote the random vector $(x_1, \cdots, x_n)$ by $x_{1:n}$ and observations $(y_1, \cdots, y_n)$ by $y_{1:n}$. Assume the target integral is (1.4). In (1.6), let $\pi_n^*(x_{0:n}) = p(x_{0:n}|y_{1:n})$ and

$$\pi_n(x_{0:n}) \equiv p(x_0) \prod_{k=1}^{n} p(y_k|x_k)p(x_k|x_{k-1}).$$

Then $\pi_n(x_{0:n}) \propto \pi_n^*(x_{0:n})$. Due to the high dimension of π_n^*, direct application of ratio estimator of IS may result in few random draws lying in the high likelihood area of π_n^* and low estimating efficiency in most cases. By combining the sequential importance sampling (SIS) and resampling algorithms as follows (Chopin, 2004), the high dimensional problem can be mitigated:

Basic SMC method:

At time n, assume weighted particles $\{(\widetilde{x}_{0:n-1}^{(j)}; \widetilde{w}_{n-1}^{(j)})\}_{j=1}^N$ are available. For $j = 1, \cdots, N$,

1. Mutation: Generate $x_n^{(j)}$ from the proposal distribution $q(x_n|\widetilde{x}_{0:n-1}^{(j)})$ and let $x_{0:n}^{(j)} = (\widetilde{x}_{0:n-1}^{(j)}, x_n^{(j)})$.

2. Correction: Assign $x_{0:n}^{(j)}$ with weight

$$w_n^{(j)} = \widetilde{w}_{n-1}^{(j)} \frac{p(x_n^{(j)}|\widetilde{x}_{n-1}^{(j)})p(y_n|x_n^{(j)})}{q(x_n^{(j)}|\widetilde{x}_{0:n-1}^{(j)})}. \tag{3.8}$$

3. Selection: If the condition for resampling is satisfied, resample $\{x_{0:n}^{(j)}\}_{j=1}^N$ according to $\{w_n^{(j)}\}_{j=1}^N$ to obtain new weighted particles $\{(\widetilde{x}_{0:n}^{(j)}; \widetilde{\omega}_n^{(j)})\}_{j=1}^N$ where $\widetilde{\omega}_n^{(j)} = 1/N$; If the condition for resampling is not satisfied, let $(\widetilde{x}_{0:n}^{(j)}; \widetilde{\omega}_n^{(j)}) = (x_{0:n}^{(j)}; \omega_n^{(j)})$.

After the correction step, μ_n can be estimated by

$$\widehat{\mu}_{n,basic} = \frac{\sum_{j=1}^N h(x_{1:n}^{(j)})\omega_n^{(j)}}{\sum_{j=1}^N \omega_n^{(j)}}.$$

The SIS algorithm, containing the mutation and correction steps, divides the sampling into sequential steps, which can be seen from the following equation

$$\frac{\pi_n(x_{0:n})}{q(x_{0:n})} = \frac{\pi_{n-1}(x_{0:n-1})}{q(x_{0:n-1})} \frac{\pi_n(x_{0:n})}{\pi_{n-1}(x_{0:n-1})q(x_n|x_{0:n-1})},$$

where $q(x_{0:n}) = \pi_0(x_0)\prod_{t=1}^{n} q(x_t|x_{0:t-1})$ and $\pi_n(x_{0:n})/\pi_{n-1}(x_{0:n-1}) = p(x_n|x_{n-1})p(y_n|x_n)$. This sequential implementation is appropriate for on-line analysis. The difficulty of SIS is that as n goes large, a few particles will have dominating weights and the effective sample size will be small. The selection step performs resampling to drop the samples with low weights and duplicate the samples with large weights, to avoid the sample degeneracy. The schedule of performing resampling can be either fix or adaptive according to some quality indictor of particles (Liu, 2008). After resampling, the distribution of $\left\{\widetilde{x}_{0:n}^{(j)}\right\}_{j=1}^{N}$ converges to $\pi_n^*(x_{0:n})$ as $N \to \infty$ (Crisan and Doucet, 2002) and therefore equal sample weights are assigned. The algorithm may be initialized by generating i.i.d samples $\left\{x_0^{(j)}\right\}_{j=1}^{n}$ from $\pi_0(x_0)$ and setting $w_0^{(j)} = 1/N$ for all j.

In the algorithm, one needs to choose the proposal density $q(x_n|x_{0:n-1})$. A simple choice is the prior density $p(x_n|x_{n-1})$. but since the sampling does not depend on $y_{1:n}$, the algorithm may loss efficiency when the observations has outliers. The density $p(x_n|y_n, x_{n-1})$, which is the normalized $p(x_n|x_{n-1})p(y_n|x_n)$ taking x_{n-1} as fixed, is considered as the optimal proposal (e.g. Doucet et al., 2000 and Pitt and Shephard, 1999), in the sense that the variance of $w_n^{(j)}$ conditional on $\{\widetilde{x}_{0:n-1}^{(j)}\}_{j=1}^{N}$ is 0. Since the analytical form of $p(x_n|y_n, x_{n-1})$ is usually unavailable, approximation to $p(x_n|y_n, x_{n-1})$ by linearizing $log(p(x_n|x_{n-1}))$ or $log(p(y_n|x_n))$ may be used instead.

The central limit theorem of $\widehat{\mu}_{n,basic}$ is given in Chopin (2004) and stated here as the preliminary for the theoretical result of new algorithm. Assume multinomial resampling is performed at every step. Denote the domain of $x_{0:n}$ by Θ_n. Consider following conditions:

(C1) $\int |h(x_{1:n})|\pi_n^*(x_{0:n})dx_{0:n}$ and $\int |h(x_{1:n})|\pi_{n-1}^*(x_{0:n-1})q(x_n|x_{0:n-1})dx_{0:n} < \infty$;

(C2) $E_{\pi_n^*}[h^2(x_{1:n})] < \infty$;

(C3) Let Φ_0 to be the set of square integrable functions with respect to $\pi_0(x_0)$ and

$\Phi_n = \{h : \Theta_n \to \mathbb{R} \Big| E_{\pi^*_{n-1}q}\left[\frac{\pi^*_n}{\pi^*_{n-1}q}h\right]^{2+\delta} < \infty,\ E_q\left[\frac{\pi^*_n}{\pi^*_{n-1}q}h|x_{0:n-1}\right] \in \Phi_{n-1}\}$ for some positive δ. Then $h(x_{1:n}) \in \Phi_n$;

(C4) The unit function $I_n : \Theta_n \to 1$ belongs to Φ_n.

Theorem 3.1. *(Chopin, 2004) Let $V_{3,0}(h) = Var_{\pi_0}(h)$ and by induction, define*

$$
\begin{aligned}
V_{1,n}(h) &= V_{3,n-1}\left(E_q\left[h(x_{1:n})|x_{0:n-1}\right]\right) + E_{\pi^*_{n-1}}\left(Var_q\left[h(x_{1:n})|x_{0:n-1}\right]\right), && n > 0,\\
V_{2,n}(h) &= V_{1,n}\left(\frac{\pi^*_n(x_{0:n})(h(x_{1:n}) - \mu_n)}{\pi^*_{n-1}(x_{0:n-1})q(x_n|x_{0:n-1})}\right), && n \geq 0,\\
V_{3,n}(h) &= V_{2,n}\left(h\right) + Var_{\pi^*_n}\left(h\right), && n \geq 0.
\end{aligned}
$$

Suppose conditions (C1)-(C4) are satisfied. Then for any n, μ_n, $V_{2,n}(h)$ and $V_{3,n}(h)$ are finite and the following convergence hold:

$$\sqrt{N}\left[\frac{\sum_{j=1}^N h(x_n^{(j)})w_n^{(j)}}{\sum_{j=1}^N w_n^{(j)}} - \mu_n\right] \xrightarrow{\mathfrak{L}} N(0, V_{2,n}(h)), \tag{3.9}$$

$$\sqrt{N}\left[\frac{1}{N}\sum_{j=1}^N h(\widetilde{x}_{1:n}^{(j)}) - \mu_n\right] \xrightarrow{\mathfrak{L}} N(0, V_{3,n}(h)) \tag{3.10}$$

Specifically,

$$V_{2,n}(h) = \int \frac{\pi^*_n(x_{0:1})^2(\mu_1(x_{0:1}) - \mu_n)^2}{\pi_0(x_0)q(x_1|x_0)}dx_{0:1} + \sum_{t=2}^n \frac{\pi^*_n(x_{0:t})^2(\mu_t(x_{0:t}) - \mu_n)^2}{\pi^*_{t-1}(x_{0:t-1})q(x_t|x_{0:t-1})}dx_{0:t}, \tag{3.11}$$

where $\mu_t(x_{0:t}) = \int h(x_{1:n})\pi^*_n(x_{t+1:n}|x_{0:t})dx_{t+1:n}$ and $\pi^*_n(x_{0:t})$ is the marginal density of $\pi^*_n(x_{0:n})$.

Condition (C1) is required for the law of large number of triangular array $h(x_{1:n})$ conditional on $x_{0:n-1}$. Conditions (C2) to (C4) ensure the asymptotic variances are finite. By Johansen and Doucet (2008), each term in (3.11) can be interpreted as an Importance Sampling variance where the target integral is $\int \mu_t(x_{0:t})\pi^*_n(x_{0:t})dx_{0:t}$ and the importance distribution is $\pi^*_{t-1}(x_{0:t-1})q(x_t|x_{0:t-1})$.

3.2.2 Generalized SMC

In order to include the information from both state and observation equations in generating new particles, a mixture of two proposal distributions, one derived from the state equation and the other depending on the most recent observation, can be used as an alternative to $p(x_n|y_n, x_{n-1})$. Elinas et al. (2006) and Thrun et al. (2001) applied this strategy for the Monte Carlo localization problem of robot. Control variate is a general method to reduce the variance of Monte Carlo estimation. Given an unbiased estimator X of some target value Z, improvement can be gained by constructing a proper control variate vector $\mathbf{Y}$ and using $X - \boldsymbol{\beta}^T (\boldsymbol{Y} - E[\boldsymbol{Y}])$ to estimate Z. The optimal $\boldsymbol{\beta}$ can be estimated by least squares to minimize the asymptotic variance (Cochran, 1977). Although the control variate approach has been studied extensively for importance sampling, few discussions are devoted to the SMC framework where importance sampling is a special case, possible due to the lack of selection guideline for appropriate control variates and theoretical understand of control variates under SMC framework. In the basic SMC method, one implementation of control variates is to introduce control variates $\boldsymbol{S}(x_{1:n})$ satisfying $\int \boldsymbol{S}(x_{1:n})\pi_n^*(x_{0:n})dx_{0:n} = 0$ and replace the target function $h(x_{1:n})$ in $\widehat{\mu}_{n,basic}$ with $h(x_{1:n}) + \widehat{\boldsymbol{\gamma}}_n^T \boldsymbol{S}(x_{1:n})$ where $\widehat{\boldsymbol{\gamma}}_n$ is the estimated optimal coefficients (Singh et al., 2004). In this setting the estimator is still asymptotically unbiased.

One well-known limitation of using $p(x_n|y_n, x_{n-1})$ as a proposal density is that it may lose efficiency if the discrepancy between the successive densities $\pi_{n-1}^*(x_{0:n-1})$ and $\pi_n^*(x_{0:n-1})$ is large. This is because after resampling, the particles which reside in the high likelihood area of $\pi_{n-1}^*(x_{0:n-1})$ may have low values of $\pi_n^*(x_{0:n-1})$ due to the large discrepancy. Its impact can be seen from (3.8),

$$w_n^{(j)} = \widetilde{w}_{n-1}^{(j)} \frac{\pi_n(x_{0:n}^{(j)})}{\pi_{n-1}(\widetilde{x}_{0:n-1}^{(j)}) q(x_n^{(j)}|\widetilde{x}_{0:n-1}^{(j)})} = \widetilde{w}_{n-1}^{(j)} \frac{\pi_n(\widetilde{x}_{0:n-1}^{(j)})}{\pi_{n-1}(\widetilde{x}_{0:n-1}^{(j)})} \cdot \frac{\pi_n(x_n^{(j)}|\widetilde{x}_{0:n-1}^{(j)})}{q(x_n^{(j)}|\widetilde{x}_{0:n-1}^{(j)})},$$

where $\pi_n(x_n|x_{0:n-1}) = p(x_n|y_n, x_{n-1})$. Larger discrepancy results in small sample weights after mutating and increases the variance. Pitt and Shephard (1999) and Carpenter et al. (1999) proposed a look ahead method, named auxiliary particle filter,

by adjusting the distribution of $\widetilde{x}_{0:n-1}^{(j)}$ according to the new observation y_n. Specifically, since $\pi_n(x_{0:n-1})/\pi_{n-1}(x_{0:n-1})$ does not depend on x_n, by resampling according to $\{w_{n-1}^{(j)}\pi_n(x_{0:n-1}^{(j)})/\pi_{n-1}(x_{0:n-1}^{(j)})\}$ instead of $\{w_{n-1}^{(j)}\}$ in the selection step, the selected $\{\widetilde{x}_{0:n-1}^{(j)}\}$ can reside in the high likelihood area of $\pi_n^*(x_{0:n-1})$ and therefore will not increase the skewness of $w_n^{(j)}$. The analytical form of $\pi_n(x_{0:n-1})/\pi_{n-1}(x_{0:n-1})$ is usually unavailable in practice and some approximation $\eta(x_{0:n-1})$, named auxiliary variable, is used instead.

Then with $\eta(x_{0:n})$ being the auxiliary variable, $\boldsymbol{S}(x_{0:n})$ being the control variates and $q_{\boldsymbol{\alpha}}(x_n|x_{1:n-1})$ being the mixture proposal density where $q_{\boldsymbol{\alpha}}(x_n|x_{1:n-1}) = \sum_{k=1}^{p}\alpha_k q_k(x_n|x_{1:n-1})$ given p proposal densities $q_1(x_n|x_{1:n-1}), \cdots, q_p(x_n|x_{1:n-1})$ and mixture proportion vector $\boldsymbol{\alpha}$ at every time n, we have a more general SMC algorithm as follows:

Generalized SMC method:

At time n, assume weighted samples $\{(\widetilde{x}_{0:n-1}^{(j)}; \widetilde{w}_{n-1}^{(j)})\}_{j=1}^{N}$ are available. For $j = 1, \cdots, N$,

1. Mutation: Generate $x_n^{(j)}$ from the proposal distribution $q_{\boldsymbol{\alpha}}(x_n|\widetilde{x}_{0:n-1}^{(j)})$ and let $x_{0:n}^{(j)} = (\widetilde{x}_{0:n-1}^{(j)}, x_n^{(j)})$.

2. Correction: Assign $x_{0:n}^{(j)}$ with weight

$$w_n^{(j)} = \widetilde{w}_{n-1}^{(j)} \frac{p(x_n^{(j)}|\widetilde{x}_{n-1}^{(j)})p(y_n|x_n^{(j)})}{q_{\boldsymbol{\alpha}}(x_n^{(j)}|\widetilde{x}_{0:n-1}^{(j)})}. \tag{3.12}$$

3. Selection: If the condition for resampling is satisfied, resample $\{x_{0:n}^{(j)}\}_{j=1}^{N}$ according to $\{\eta(x_{0:n}^{(j)})w_n^{(j)}\}_{j=1}^{N}$ to obtain new weighted particles $\{(\widetilde{x}_{0:n}^{(j)}; \widetilde{\omega}_n^{(j)})\}_{j=1}^{N}$ where $\widetilde{\omega}_n^{(j)} = 1/\eta(\widetilde{x}_{0:n}^{(j)})$; If the condition for resampling is not satisfied, let $(\widetilde{x}_{0:n}^{(j)}; \widetilde{\omega}_n^{(j)}) = (x_{0:n}^{(j)}; \omega_n^{(j)})$.

After the correction step, μ_n can be estimated by

$$\widehat{\mu}_{n,generalized} = \frac{\sum_{j=1}^{N}\left[h(x_n^{(j)}) + \widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{S}(x_{0:n})\right]\omega_n^{(j)}}{\sum_{j=1}^{N}\omega_n^{(j)}}.$$

3.2.3 Limitations of the Generalized SMC Method

Although $p(x_n|y_n, x_{n-1})$ is well accepted as the optimal proposal density and in principle the proposal should be close to $p(x_n|y_n, x_{n-1})$, several concerns may limit the usage of a proposal that is more sophisticated than the prior density $p(x_n \mid x_{0:n-1})$.

First, it is necessary for $q(x_n|x_{0:n-1})$ to have heavier tails than $p(x_n|x_{n-1})p(y_n|x_n)$ so that sample weight $w_n^{(j)}$ has a bounded variance. But it is difficult to obtain an accurate approximation of $p(x_n|y_n, x_{n-1})$ with heavier tails, except in some special cases such as a log-concave $p(x_n|x_{n-1})p(y_n|x_n)$ approximated by a first order Taylor expansion.

Second, $p(x_n|y_n, x_{n-1})$ is not the real optimal proposal distribution. In (3.11), it is not possible to design a sequence of proposal distributions to minimize $V_{2,n}(h)$ for every n, since the optimal proposal for each term of $V_{2,n}(h)$ changes when n increases. A reasonable strategy is to construct $q(x_n|x_{0:n-1})$ by minimizing the last term in (3.11) which is

$$\int \frac{\pi_n(x_{0:n})^2(h(x_{1:n}) - \mu_n)^2}{\pi_{n-1}(x_{0:n-1})q(x_n|x_{0:n-1})} dx_{0:n},$$

since all the early terms contain "future" information. The early terms are expected to decay over time in some ergodic system, which makes this strategy valid (Johansen and Doucet, 2008). Then the minimizer is the density proportional to $p(x_n|x_{n-1})p(y_n|x_n)|h(x_{1:n}) - \mu_n|$. In this sense, $p(x_n|x_{n-1}, y_n)$ is only suboptimal, even when used with the auxiliary variable $p(y_n|x_{n-1})$, since it does not consider the target function $h(x_{1:n})$. Sometimes, using $p(x_n|x_{n-1}, y_n)$ as a proposal can be outperformed by the bootstrap filter (Johansen and Doucet, 2008).

Although the above strategy suggests the use of a density proportional to

$$p(x_n|x_{n-1})p(y_n|x_n)|h(x_{1:n}) - \mu_n|,$$

it requires the knowledge of μ_n which is the purpose of filtering in the first place and therefore cannot be used. On the other hand, the proposal choice tells that it might be beneficial to consider the target function $h(x_{1:n})$ when selecting the proposal distribution.

Chapter 4

Two-stage Importance Sampling with Mixture Proposals

This chapter proposes a two-stage procedure to optimize the sample allocation among multiple proposals and investigate its theoretical properties. In the first stage, pilot sample is drawn from a mixture proposal with predetermined proportions. The optimal mixture proportions are then estimated by minimizing the estimated asymptotic variance of Owen and Zhou (2000)'s regression estimator or Tan (2004)'s MLE estimator. In the second stage, the sample is drawn from the mixture proposal with the estimated proportions. Integral estimators are constructed using all observations, including those from the pilot stage. Then we establish a theoretical framework of such a two-stage procedure. It is shown that under very weak conditions, the integral estimators constructed by the two-stage procedure are consistent and asymptotic normal with minimum asymptotic variance over all mixture proportions. Therefore, the two-stage procedure is adaptive towards using the optimal mixture proportions. The optimal sample size used for the pilot stage is also discussed in the sense of minimizing an approximated mean square error in higher order. Furthermore, we extend Owen's regression estimator and Tan's MLE to the ratio estimators of IS. When estimating μ, if one can evaluate $\pi^*(x)$ only up to a normalizing constant, a ratio estimator is used, with the numerator being the estimated unnormalized integral and the denominator being the estimated normalizing constant. We show that the two-stage procedure for this extension also has the desirable asymptotic properties.

4.1 Two Stage Procedure

Suppose p proposal distributions $q_1, \cdots, q_p$ are given and the sample size is budgeted at n. Let $\Theta = [\delta, 1-\delta]^p$ where δ is some constant close to 0. The following algorithm

is proposed to select mixture proportions $\boldsymbol{\alpha}$ and construct estimators:

1. First stage: Given a p dimensional vector $\boldsymbol{\gamma}$ satisfying $\sum_{k=1}^{p}\gamma_k = 1$, generate n_0 independent stratified observations $\{x_i\}_{i=1}^{n_0}$ from $q_{\boldsymbol{\gamma}}(x) = \sum_{k=1}^{p}\gamma_k q_k(x)$, i.e. $n_0\gamma_k$ observations from $q_k(x)$, $k = 1, \cdots, p$. Obtain $\widehat{\boldsymbol{\alpha}}$ by minimizing

$$\widehat{\sigma}_Z^2(\boldsymbol{\alpha}) = \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\left[\pi(x_i) - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}^T\boldsymbol{g}(x_i)\right]^2}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)} \tag{4.1}$$

   where

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \left(\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\boldsymbol{g}(x_i)\boldsymbol{g}(x_i)^T}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}\right)^{-1}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\pi(x_i)\boldsymbol{g}(x_i)}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}\right)$$

   and $\boldsymbol{g}(x) = (q_2(x) - q_1(x), \cdots, q_p(x) - q_1(x))^T$, with respect to $\boldsymbol{\alpha}$ over Θ.

2. Second stage: Generate $n - n_0$ independent stratified observations $\{x_i\}_{i=n_0+1}^{n}$ from $q_{\widehat{\boldsymbol{\alpha}}}(x) = \sum_{k=1}^{p}\widehat{\alpha}_k q_k(x)$. Estimate integral Z by $\widehat{Z}(\widetilde{\boldsymbol{\alpha}})$ with all n observations, where

$$\widetilde{\boldsymbol{\alpha}} = \frac{n_0}{n}\boldsymbol{\gamma} + \frac{n - n_0}{n}\widehat{\boldsymbol{\alpha}} \tag{4.2}$$

   and $\widehat{Z}(\widetilde{\boldsymbol{\alpha}})$ can be either $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ or $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$.

Some rationale and implementation remarks are as follows:

(i) Criterion of selecting $\boldsymbol{\alpha}$: In the first stage, the optimal $\boldsymbol{\alpha}$ is estimated using the n_0 samples and it is desirable to select $\boldsymbol{\alpha}$ that gives the smallest asymptotic variance of the final estimator. Let $Var_{\boldsymbol{\alpha}}$ denotes the variance taken with respect to $q_{\boldsymbol{\alpha}}(x)$. We set the following conditions:

(C1) The union of supports of $q_k(x)$ contains the support of $\pi(x)$;

(C2) $\alpha_i > 0$ for $i = 1, \cdots, p$;

(C3) $Var_{\boldsymbol{\alpha}}\left[\pi(X)/q_{\boldsymbol{\alpha}}(X)\right] < \infty$ for some $\boldsymbol{\alpha} \in \Theta$.

Owen and Zhou (2000) and Tan (2004) showed that, under the above conditions, $\widehat{Z}_{Reg}(\boldsymbol{\alpha})$ and $\widehat{Z}_{MLE}(\boldsymbol{\alpha})$ are asymptotic normal and have the same asymptotic

variance

$$\sigma_Z^2(\boldsymbol{\alpha}) = Var_{\boldsymbol{\alpha}}\left[\frac{\pi(X) - \boldsymbol{\beta}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right] = \int \frac{\left(\pi(x) - \boldsymbol{\beta}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(x)\right)^2}{q_{\boldsymbol{\alpha}}(x)} dx - Z^2 \tag{4.3}$$

where

$$\begin{aligned}\boldsymbol{\beta}_{\boldsymbol{\alpha}} &= Var_{\boldsymbol{\alpha}}\left[\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right]^{-1} Cov_{\boldsymbol{\alpha}}^T\left[\frac{\pi(X)}{q_{\boldsymbol{\alpha}}(X)}, \frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right] \\ &= \left(\int \frac{\boldsymbol{g}(x)\boldsymbol{g}(x)^T}{q_{\boldsymbol{\alpha}}(x)} dx\right)^{-1} \left(\int \frac{\pi(x)\boldsymbol{g}(x)}{q_{\boldsymbol{\alpha}}(x)} dx\right).\end{aligned}$$

Conditions $(C1)$ to $(C3)$ are satisfied when we have at least one proposal component dominating the tail of $\pi(x)$. With the sample $\{x_i\}_{i=1}^{n_0}$ from the pilot stage, $\sigma_Z^2(\boldsymbol{\alpha}) + Z^2$ is estimated by the importance sampling estimator $\widehat{\sigma}_Z^2(\boldsymbol{\alpha})$ in (4.1) and the optimal $\boldsymbol{\alpha}$ is obtained by minimizing $\widehat{\sigma}_Z^2(\boldsymbol{\alpha})$.

(ii) Optimization range for $\boldsymbol{\alpha}$: The purpose of restricting $\boldsymbol{\alpha}$ in $[\delta, 1-\delta]^p$ for some small δ is to avoid unreliable estimators of $\sigma_Z^2(\boldsymbol{\alpha})$ or $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$. When $\alpha_i = 0$ for some i, $\int \pi(x)^2/q_{\boldsymbol{\alpha}}(x)dx$ can be infinite if q_i is the only proposal that dominates certain part of $\pi(x)$'s tail, or $\int \boldsymbol{g}(x)\boldsymbol{g}(x)^T/q_{\boldsymbol{\alpha}}(x)dx$ and $\int \pi(x)\boldsymbol{g}(x)/q_{\boldsymbol{\alpha}}(x)dx$ can be infinite if q_i is the only proposal that dominates some other proposals. In this case, if α_i is too close to 0, the estimator $\widehat{\sigma}_Z^2(\boldsymbol{\alpha})$ or $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$ is unreliable. Experience shows that $\delta = .001$ is a reasonable choice.

(iii) Choice of the initial proportions $\boldsymbol{\gamma}$: $\boldsymbol{\gamma}$ is preferred to be close to the optimal proportion vector $\boldsymbol{\alpha}^*$. If there is no any prior knowledge about $\boldsymbol{\alpha}^*$, it is recommended to use $\boldsymbol{\gamma}$ with equal components in the first stage so that pilot sample is generated from each proposal equally.

(iv) $\widehat{Z}$ in second stage: Instead of using $n - n_0$ observations to construct the estimator $\widehat{Z}(\widehat{\boldsymbol{\alpha}})$, we utilize all n observations to construct the estimator $\widehat{Z}(\widetilde{\boldsymbol{\alpha}})$ where the mixture proportions $\widetilde{\boldsymbol{\alpha}}$ account for the proportions of the combined sample.

4.2 Theoretical Properties

Let $\boldsymbol{\alpha}^*$ be the minimizer of $\sigma_Z^2(\boldsymbol{\alpha})$ under restriction $\boldsymbol{\alpha} \in \Theta$. We assume the following additional conditions:

(C4) $n_0 = o(n)$ and $n_0 \to \infty$ as $n \to \infty$;

(C5) $\pi(x)$ is not a linear combination of $q_1(x), \cdots, q_p(x)$;

(C6) $\boldsymbol{\alpha}^*$ is in the interior of Θ, that is, $\boldsymbol{\alpha}^* \in (\delta, 1-\delta)^p$.

Condition (C4) ensures $\widetilde{\boldsymbol{\alpha}}$ converges to $\boldsymbol{\alpha}^*$. Condition (C5) is necessary since if $\pi(x)$ is a linear combination of $q_1(x), \cdots, q_p(x)$, $\sigma^2(\boldsymbol{\alpha})$ will be 0 for all α_1. Some discussions of condition (C6) are given in Remark 7.

4.2.1 First Order Properties

Theorem 4.1. *Under conditions* $(C1)$ *to* $(C5)$, $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ *and* $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ *are consistent and*

$$\sqrt{n}\left(\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}}) - Z\right) \xrightarrow{\mathfrak{L}} N(0, \sigma_Z^2(\boldsymbol{\alpha}^*))$$
$$\text{and } \sqrt{n}\left(\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}}) - Z\right) \xrightarrow{\mathfrak{L}} N(0, \sigma_Z^2(\boldsymbol{\alpha}^*)).$$

Therefore, the two-stage procedure achieves the minimum asymptotic variance that Owen and Zhou's and Tan's estimators can achieve among all possible mixture proportions. Furthermore, since $\widehat{Z}_{Reg}(\boldsymbol{\alpha})$ and $\widehat{Z}_{MLE}(\boldsymbol{\alpha})$ are better than the stratified sampling estimator $\widehat{Z}_{SIS}(\boldsymbol{\alpha})$, the two-stage procedure outperforms all estimators introduced in Section 3.1 in asymptotic variance. The proof is given in the last section of this chapter.

Remark 4.1. It is important to point out that $\sigma_Z^2(\boldsymbol{\alpha})$ and its estimator $\widehat{\sigma}^2(\boldsymbol{\alpha})$ are strictly convex by Lemma 1 in the last section of this chapter. This guarantees a unique solution and applicability of convex optimization algorithms in the pilot stage. This property, or equivalently the strict convexity of the function $\sigma^2(\boldsymbol{\alpha}, \beta) = Var_{\boldsymbol{\alpha}}\left[\left(\pi(X) - \boldsymbol{\beta}^T \boldsymbol{g}(X)\right)/q_{\boldsymbol{\alpha}}(X)\right]$, also ensures the consistency and asymptotic normality with convergence rate $\sqrt{n_0}$ of random proportion vector $\widehat{\boldsymbol{\alpha}}$ under mild conditions,

by asymptotic theory for M-estimation with a convex criterion function (Haberman, 1989). Therefore larger n_0 gives more reliable $\widehat{\boldsymbol{\alpha}}$.

Remark 4.2. For $\widehat{Z}_{SIS}(\boldsymbol{\alpha})$, the optimal mixture proportions are the ones that make the mixture proposal $q_{\boldsymbol{\alpha}}$ the closest to the target distribution π. Therefore knowledge about the target density surface can help to find an approximate choice of $\boldsymbol{\alpha}$. However, the optimal mixture proportions $\boldsymbol{\alpha}^*$ for $\sigma_Z^2(\boldsymbol{\alpha})$ sometimes can be counterintuitive. For instance, in Example 1(B2) of Section 4.5, the target distribution is a mixture of a normal distribution and a t distribution, with mixing probability 0.8 and 0.2 respectively. When the same normal distribution is used as one of the proposal distributions, its optimal mixture proportion is only .1%. This is due to the fact that, for $\widehat{Z}_{Reg}(\boldsymbol{\alpha})$ and $\widehat{Z}_{MLE}(\boldsymbol{\alpha})$, the numerator of $\sigma_Z^2(\boldsymbol{\alpha})$ involves $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$, a function of $\boldsymbol{\alpha}$, which complicates the determination of the optimal proportions. Hence, an automatic selection for mixture proportions becomes necessary for $\widehat{Z}_{Reg}(\boldsymbol{\alpha})$ and $\widehat{Z}_{MLE}(\boldsymbol{\alpha})$.

Remark 4.3. If $\widetilde{\boldsymbol{\alpha}}$ in (4.2) can be replaced by some other random proportion vector, as long as it is consistent to $\boldsymbol{\alpha}^*$ as $n \to \infty$, the same asymptotic results hold. For example, one can choose the mixture proportions of the second stage so that the combined sample (of both the pilot stage and second stage) is as close to the estimated optimal proportion vector $\widehat{\boldsymbol{\alpha}}$ as possible. For example, if $n_0\gamma_k < n\widehat{\alpha}_k$ for all $k = 1, \ldots, p$, one can use $(n\widehat{\boldsymbol{\alpha}} - n_0\boldsymbol{\gamma})/(n - n_0)$ in the second stage which results in the combined sample having the exact estimated optimal proportion $\widehat{\boldsymbol{\alpha}}$. In this case, actually one should use n_0 as large as possible until it violates the above condition.

Remark 4.4. Similar asymptotic properties for $\widehat{Z}_{MIS}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{SIS}(\widetilde{\boldsymbol{\alpha}})$ are presented in Lemma 3 in the technical proof. They are always inferior to the control-variate based estimators and hence of less interest.

4.2.2 High Order Properties

Theorem 1 shows that the selection of the pilot sample size n_0 does not affect the first order property of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ as long as $n_0 = o(n)$ and $n_0 \to \infty$. Therefore an optimal choice of n_0 needs to be determined by higher order properties of

$\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$. Consider the convergence rate of $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$, a weighted average of $\boldsymbol{\gamma} - \boldsymbol{\alpha}^*$ and $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$ with weights n_0/n and $1 - n_0/n$. Since $\boldsymbol{\gamma} - \boldsymbol{\alpha}^*$ is biased, one would want to have a smaller n_0. However, a large n_0 makes $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$ closer to 0, at the rate $O\left(1/\sqrt{n_0}\right)$. Therefore the optimal n_0 is chosen to balance the effects of these two rates. The following proposition gives the higher order asymptotic expansions of $\widetilde{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widetilde{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$.

Proposition 4.1. *Under conditions (C1)-(C6),* $\widetilde{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ *and* $\widetilde{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ *can be expanded as* $\widehat{Z}^* + o\left(n_0/\left(n\sqrt{n}\right)\right) + o\left(1/\left(n_0\sqrt{n}\right)\right)$ *and* $\widehat{Z}^* = Z + g_1(\widetilde{\boldsymbol{\alpha}}) + g_2(\widetilde{\boldsymbol{\alpha}})$*, where*

$$g_1(\widetilde{\boldsymbol{\alpha}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(x_i) - \beta_{\boldsymbol{\alpha}^*} g(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int \frac{\pi(x) - \beta_{\boldsymbol{\alpha}^*} g(x)}{q_{\boldsymbol{\alpha}^*}(x)} q_{\widetilde{\boldsymbol{\alpha}}}(x)dx, \quad \textit{and}$$
$$g_2(\widetilde{\boldsymbol{\alpha}}) = O\left(\frac{n_0}{n\sqrt{n}}\right) + O\left(\frac{1}{n_0\sqrt{n}}\right).$$

The explicit forms of $g_2(\widetilde{\boldsymbol{\alpha}})$ are tedious and therefore presented in the technical proof. The selection of optimal n_0 is based on minimizing the mean square error of $\widehat{Z}^*$, which is an approximation of the mean square error of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$. Such an approximation of moments, as the criterion of second order optimality, has been widely used in higher-order asymptotic theory, e.g. Rothenberg (1984).

Theorem 4.2. *Under conditions (C1)-(C6) and*

(C7) $\int \pi(x)^4/q_{\boldsymbol{\alpha}}(x)^4 dx < \infty$ *for some* $\boldsymbol{\alpha} \in \Theta$*,*

it holds that

$$E\left[\widehat{Z}^* - Z\right] = O\left(\frac{1}{n}\right) \textit{ and } Var\left[\widehat{Z}^* - Z\right] = \frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + O\left(\frac{n_0}{n^2}\right) + O\left(\frac{1}{nn_0}\right).$$

Therefore $MSE\left[\widehat{Z}^*\right] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*) = O\left(\frac{n_0}{n^2}\right) + O\left(\frac{1}{nn_0}\right)$.

The above result gives the approximate mean squared error with higher order terms beyond the usual asymptotic variance $n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$. The order can be attributed to three sources of variability. See the technical proof for details. One source of variability is due to using the pilot sample with mixture proportions $\boldsymbol{\gamma} \neq \boldsymbol{\alpha}^*$, which leads to terms of order $O\left(n_0/n^2\right)$. The second source is the variability of estimator $\widehat{\boldsymbol{\alpha}}$, which is of the order $O\left(1/(nn_0)\right)$. The third source is the variability of estimating $\boldsymbol{\beta}_{\boldsymbol{\alpha}^*}$, which is the

optimal coefficient of control variates, in $\sigma_Z^2(\boldsymbol{\alpha}^*)$. In $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$, the estimator of $\boldsymbol{\beta}_{\boldsymbol{\alpha}^*}$ is $\widehat{\boldsymbol{\beta}}_{\widetilde{\boldsymbol{\alpha}}}$. In $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$, a similar estimator is used, as can be seen from the proof of Theorem 1. This variability is of the order $O(1/(n\sqrt{n}))$ which is also $O\left(n_0/n^2\right) + O\left(1/(nn_0)\right)$ because $2/\sqrt{n} \leq n_0/n + 1/n_0$ by the inequality $2ab \leq a^2 + b^2$.

Remark 4.5. By minimizing the order of difference, the optimal n_0 is $O(\sqrt{n})$ and hence $MSE\left[\widehat{Z}^*\right] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$ is of order $O\left(1/(n\sqrt{n})\right)$. The asymptotic rate shows that how n_0 should change with the total sample size n. In practice, another consideration of selecting n_0 is the coverage of the target distribution with pilot samples. A poor coverage can lead to poorly estimated asymptotic variance and result in inaccurate $\widehat{\boldsymbol{\alpha}}$. Our experience shows one should choose n_0 at least $\sqrt{n}$ and possibly larger according to the complexity of problem and the quality of proposal distributions. On the other hand, one can assess $\widehat{\boldsymbol{\alpha}}$ by estimating its standard error after the pilot stage. If the standard error is larger than some criterion, such as 10% of $\widehat{\boldsymbol{\alpha}}$, one can add additional pilot samples. The standard error formula is given in the last part of the technical proof.

Remark 4.6. One essential fact leading to Theorem 2 is $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* = O(1/\sqrt{n_0}) + O(n_0/n)$. Therefore when $\widetilde{\boldsymbol{\alpha}}$ is replaced by some other construction which is consistent but with different rate (e.g. Remark 3 above), the orders in Theorem 2 may change.

Remark 4.7. When some coordinates of $\boldsymbol{\alpha}^*$ are on the boundary of $[\delta, 1-\delta]$, the exact second order property is complicated. However, it is still reasonable to use the same n_0 as indicated in Theorem 2. For example, when α_1^* is on the boundary, $\widehat{\boldsymbol{\alpha}}$, as an M-estimator, will converge to $\boldsymbol{\alpha}^*$ with a rate faster than or equal to $O(1/\sqrt{n_0})$ (Geyer, 1994). In the proof of Theorem 2, when the convergence rate of $\widehat{\boldsymbol{\alpha}}_1$ changes from $O(1/\sqrt{n_0})$ to $O(1/n_0^{\varepsilon})$ with $\varepsilon \geq \frac{1}{2}$, the second order $O\left(n_0/n^2\right) + O\left(1/(n_0 n)\right)$ changes to $O\left(n_0/n^2\right) + O\left(1/(n_0^{2\varepsilon} n)\right)$. Then by choosing $n_0 = O(\sqrt{n})$, $MSE\left[\widehat{Z}^*\right] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$ is still $O\left(1/(n\sqrt{n})\right)$ and the accuracy of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ remains the same.

4.3 Extension to Ratio Estimators

4.3.1 Extension of IS Techniques to Ratio Estimators

As mentioned in Section 2.1, the integral (II) can be estimated by the ratio estimator

$$\widehat{\mu}_{IS} = \frac{\frac{1}{n}\sum_{i=1}^{n} h(x_i)\pi(x_i)/q(x_i)}{\frac{1}{n}\sum_{i=1}^{n} \pi(x_i)/q(x_i)}, \tag{4.4}$$

(Rubinstein and Kroese, 2008; Liu, 2008). By the delta method, it is easy to show that the asymptotic variance of $\widehat{\mu}_{IS}$ is

$$Var_q\left(\frac{h(x)\pi(x) - \mu\pi(x)}{q(x)}\right). \tag{4.5}$$

In the sense of minimizing (4.5), the optimal choice of $q(x)$ is the probability density proportional to $|h(x)\pi(x) - \mu\pi(x)|$. Therefore, it is preferred to choose $q(x)$ that mimics the shape of $|h(x)\pi(x) - \mu\pi(x)|$. Similar to estimating the normalizing constant, multiple proposals may be needed and the techniques in Section 3.1 may be beneficial.

Given p proposal distributions $q_1(x), \cdots, q_p(x)$ and mixture proportions $\{\alpha_k\}_{k=1}^{p}$ satisfying $\sum_{k=1}^{p}\alpha_k = 1$. Observations $\{x_{k1}, \cdots, x_{kn_k}\}$ are generated from proposal q_k with size $n_k = \alpha_k n$ for each k. In Hesterberg (1995), the mixture importance sampling and stratified sampling were applied to $\widehat{\mu}_{IS}$ by using the mixture proposal $q_{\boldsymbol{\alpha}}$ in numerator and denominator separately as follows:

$$\widehat{\mu}_{SIS} = \frac{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} h(x_i)\pi(x_{ki})/q_{\boldsymbol{\alpha}}(x_{ki})}{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \pi(x_{ki})/q_{\boldsymbol{\alpha}}(x_{ki})}.$$

Control variates and likelihood approach can also be applied to $\widehat{\mu}_{IS}$. With the same control variates $\boldsymbol{g}(x)$ as in (4.1), μ can be estimated by the following:

$$\widehat{\mu}_{Reg} = \frac{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{h(x_{ki})\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_1^T \boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}}{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\pi(x_{ki}) - \widehat{\boldsymbol{\beta}}_2^T \boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}}, \quad \widehat{\mu}_{MLE} = \frac{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{h(x_{ki})\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widetilde{\boldsymbol{\zeta}}^T \boldsymbol{g}(x_{ki})}}{\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki}) + \widetilde{\boldsymbol{\zeta}}^T \boldsymbol{g}(x_{ki})}},$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_1 &= \widetilde{Var}\left(\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)^{-1}\widetilde{Cov}^T\left(\frac{h(X)\pi(X)}{q_{\boldsymbol{\alpha}}(X)},\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right) \quad \text{and}\\
\widehat{\boldsymbol{\beta}}_2 &= \widetilde{Var}\left(\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)^{-1}\widetilde{Cov}^T\left(\frac{\pi(X)}{q_{\boldsymbol{\alpha}}(X)},\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)\\
\widetilde{\boldsymbol{\zeta}} &= \underset{\zeta}{\text{argmax}}\sum_{k=1}^{p}\sum_{i=1}^{n_k} log\left[q_{\boldsymbol{\alpha}}(x_{ki})+\zeta^T\boldsymbol{g}(x_{ki})\right].
\end{aligned}$$

Remark 4.8. The optimality of the above estimators can be seen by extending the optimality results of $\widehat{Z}_{Reg}$ in Owen and Zhou (2000) and $\widehat{Z}_{MLE}$ in Tan (2004) from scalar case to vector case. Specifically, under conditions (C1)-(C3) for $\pi(x)$ and $h(x)\pi(x)$, the two estimators

$$\begin{pmatrix} \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{h(x_{ki})\pi(x_{ki})-\widehat{\boldsymbol{\beta}}_1^T\boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \\ \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{\pi(x_{ki})-\widehat{\boldsymbol{\beta}}_2^T\boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \end{pmatrix} \text{ and } \begin{pmatrix} \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{h(x_{ki})\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})+\widetilde{\boldsymbol{\zeta}}^T\boldsymbol{g}(x_{ki})} \\ \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})+\widetilde{\boldsymbol{\zeta}}^T\boldsymbol{g}(x_{ki})} \end{pmatrix}$$

can be shown to be consistent and asymptotic normal with the minimum covariance matrix among all estimators in the form of

$$\begin{pmatrix} \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{h(x_{ki})\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \\ \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{\pi(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}_1^T \\ \boldsymbol{\beta}_2^T \end{pmatrix}\frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k}\frac{\boldsymbol{g}(x_{ki})}{q_{\boldsymbol{\alpha}}(x_{ki})}$$

for arbitrary real vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Here $A \geq B$ means $A-B$ is nonnegative definite for two square matrices A and B. Then by the delta method, it is straightforward to show the optimality of $\widehat{\mu}_{Reg}$ and $\widehat{\mu}_{MLE}$. Their asymptotic variances are identical and equal to

$$\sigma_\mu^2(\boldsymbol{\alpha}) = \frac{1}{Z^2}Var_{\boldsymbol{\alpha}}\left(\frac{h(X)\pi(X)-\mu\pi(X)-\boldsymbol{\beta}_{\boldsymbol{\alpha}}^T\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right), \tag{4.6}$$

where

$$\boldsymbol{\beta}_{\boldsymbol{\alpha}} = Var\left(\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right)^{-1}Cov^T\left(\frac{h(X)\pi(X)-\mu\pi(X)}{q_{\boldsymbol{\alpha}}(X)},\frac{\boldsymbol{g}(X)}{q_{\boldsymbol{\alpha}}(X)}\right).$$

4.3.2 Two Stage Procedure For Ratio Estimators

Take $\widehat{\mu}_{Reg}$ and $\widehat{\mu}_{MLE}$ as functions of $\boldsymbol{\alpha}$ and denote by $\widehat{\mu}_{Reg}(\boldsymbol{\alpha})$ and $\widehat{\mu}_{MLE}(\boldsymbol{\alpha})$. The two stage procedure in Section 4.1 can be applied here:

1. First stage: Given initial proportion $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)$ satisfying $\sum_{k=1}^{p}\gamma_k = 1$,

generate n_0 independent stratified sample $\{x_i\}_{i=1}^{n_0}$ from $q_{\boldsymbol{\gamma}}(x)$. Obtain $\widehat{\boldsymbol{\alpha}}$ by minimizing

$$\widehat{\tau}^2(\boldsymbol{\alpha}) = \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\left[h(x_i)\pi(x_i) - \widehat{\mu}\pi(x_i) - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}\mathbf{g}(x_i)\right]^2}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}, \tag{4.7}$$

$$\text{where } \widehat{\mu} = \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{h(x_i)\pi(x_i)}{q_{\boldsymbol{\gamma}}(x_i)} \Big/ \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\pi(x_i)}{q_{\boldsymbol{\gamma}}(x_i)},$$

$$\text{and } \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \left(\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\boldsymbol{g}(x_i)\boldsymbol{g}(x_i)^T}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}\right)^{-1} \left[\frac{1}{n_0}\sum_{i=1}^{n_0} \frac{\left(h(x_i)\pi(x_i) - \widehat{\mu}\pi(x_i)\right)\boldsymbol{g}(x_i)}{q_{\boldsymbol{\alpha}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}\right],$$

with respect to $\boldsymbol{\alpha}$ over Θ.

2. Second stage: Generate $n - n_0$ independent stratified observations $\{x_i\}_{i=n_0+1}^{n}$ from $q_{\widehat{\boldsymbol{\alpha}}}(x)$. Estimate integral μ by $\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ or $\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ with all n observations, where $\widetilde{\boldsymbol{\alpha}} = n_0/n \cdot \boldsymbol{\gamma} + (n - n_0)/n \cdot \widehat{\boldsymbol{\alpha}}$.

In the first stage, $\widehat{\tau}^2(\boldsymbol{\alpha})$ is the Monte Carlo estimate of $Z^2\sigma_\mu^2(\boldsymbol{\alpha})$. Similar to the results in Section 3.3.1, $\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ for μ have proper asymptotic results and the case for two proposal distributions is stated below.

Theorem 4.3. *Under conditions (C1)-(C5) with $\pi(x)$ replaced by $h(x)\pi(x) - \mu\pi(x)$, $\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ are consistent and*

$$\sqrt{n}\left(\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}}) - \mu\right) \xrightarrow{\mathfrak{L}} N(0, \sigma_\mu^2(\boldsymbol{\alpha}^*))$$

$$\textit{and } \sqrt{n}\left(\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}}) - \mu\right) \xrightarrow{\mathfrak{L}} N(0, \sigma_\mu^2(\boldsymbol{\alpha}^*)),$$

where $\boldsymbol{\alpha}^*$ is the minimizer of $\sigma_\mu^2(\boldsymbol{\alpha})$.

4.4 Selection of Component Proposal Distributions

In this paper we focus on finding the optimal mixture weights to construct a mixture proposal distribution for importance sampling, assuming that the set of component proposals to be included in the mixture has been preselected. Since the proposed mixture proportion determination automatically discriminates the high quality proposals from the poor ones, our procedure in a way alleviates the difficulty of selecting the set of proposal distributions. It also allows a larger set of proposals to be considered as

the procedure serves as a selection tool. Nevertheless, pre-selection of the proposals is extremely important as it provides the basis for efficient inference of optimal mixture weights. This is an area of active research. Here we provide some remarks and practical guidance.

Consider the asymptotic variances of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$, $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$, $\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ in (4.3) and (4.6). Owen and Zhou (2000) and Tan (2004) showed that when $\pi(x)$ is a linear combination of the component proposals, $\sigma_Z^2(\boldsymbol{\alpha}) = 0$ for any $\boldsymbol{\alpha}$. Therefore for estimating Z, it is preferred that the component proposals have a linear combination close to the shape of $\pi(x)$. This can be achieved by using proposals that separately approximate the modes and tails of $\pi(x)$. Alternatively, one can decompose $\pi(x)$ into a linear combination

$$\pi(x) = \sum_{k=1}^{r} c_k \pi_k(x). \tag{4.8}$$

Then the component proposals can be obtained by approximating each $\pi_k(x)$. Owen and Zhou (2000) give some illustrations of this strategy.

For the ratio estimator, it can be shown similarly that when $h(x)\pi(x) - \mu\pi(x)$ is a linear combination of the component proposals, $\sigma_\mu^2(\boldsymbol{\alpha}) = 0$ for any $\boldsymbol{\alpha}$. Therefore the strategy used for estimating Z can be used here as well. In particular, we can find a decomposition

$$h(x)\pi(x) - \mu\pi(x) = \sum_{k=1}^{r} c_k h(x)\pi_k(x) - \sum_{k=1}^{r} \mu c_k^* \pi_k^*(x).$$

and find component proposals to approximate the individual terms. If $h(x)$ takes negative values, additional terms corresponding to $h(x) = h^+(x) - h^-(x)$ will be needed. Example 3 in Section 4.5 provides an illustration of this approach.

Another consideration is the tail requirement. For estimating Z, $q_{\boldsymbol{\alpha}^*}(x)$ needs to have heavier tail than $\pi(x)$; and for estimating μ, $q_{\boldsymbol{\alpha}^*}(x)$ needs to have heavier tail than $h(x)\pi(x) - \mu\pi(x)$. In cases where $\pi(x)$'s tail decreases exponentially, the requirements can be satisfied by including some Student t distributions or other heavy tail distributions in the set of component proposals (Geweke, 1989).

Oh and Berger (1993) and West (1993) proposed adaptive procedures to find better proposal distributions. Liang et al. (2007) proposed a stochastic approximation procedure to partition the sample domain and used truncations of the target distribution in the subregions as component proposal distributions. The normalizing constant of each component is estimated in a pilot stage. These procedures can be used here for finding the component proposals in our setting. In fact, the pilot stage of our proposed procedure can also be used as well. The estimated optimal mixture weights from the pilot stage may provide hints on potentially useful proposals to be considered. For example, a large weight for a component proposal that mainly covers the tail in one direction may suggest to use additional proposals to cover the more extreme part of the tail in that direction. However, caution should be exercised when considering the removal of a proposal distribution because of its small weight, since it may be used to serve as a defensive proposal that guarantees finite variance of the IS estimator.

4.5 Empirical Studies

Here we present several examples to illustrate the performance of the proposed procedure. In all examples, the standard restricted optimization algorithm BFGS (Battiti and Masulli, 1990) is used in the pilot stage to find $\widehat{\boldsymbol{\alpha}}$.

Example 4.1. Let $\phi(\boldsymbol{x};\sigma)$ be the normal density with mean 0 and standard error σ, and $\psi_k(\boldsymbol{x})$ be the density of t distribution with degree of freedom k. In this example we consider two target distributions and two sets of proposal distributions. The combination is listed in Table 1. The case (A1) represents the situation that one of the proposal distribution, $q_2(\boldsymbol{x})$, is a good approximation to $\pi^*(\boldsymbol{x})$ by itself, and $q_1(\boldsymbol{x})$, being a product of Cauchy distributions, is a relatively poor proposal. We expect the two-stage procedure will be helpful to decrease the contamination of $q_2(\boldsymbol{x})$. The case (A2) represents the situation that both proposals are not good approximation to the target and an appropriate proportion is not immediately clear. Both (B1) and (B2) represent the situation that one of the proposals, $q_2(\boldsymbol{x})$, is a good approximation to the center of the target, but with a lighter tail, and the other proposal, $q_1(\boldsymbol{x})$, has a heavier

Target distribution	Proposal distributions $q_1 = \prod_{i=1}^{10} \psi_k(x_i)$ and $q_2 = \prod_{i=1}^{10} \phi(x_i;\sigma)$			
	$k=1$	$k=1$	$k=1$	$k=2$
	$\sigma = 1.1$	$\sigma = .4$	$\sigma = 1$	$\sigma = 1$
$\prod_{i=1}^{10} \phi(x_i;1)$	$(A1)$	$(A2)$		
$.2\prod_{i=1}^{10} \psi_4(x_i) + .8\prod_{i=1}^{10} \phi(x_i;1)$			$(B1)$	$(B2)$

Table 4.1: Parameter settings of four cases in Example 1

tail, for protection. The case (B1) uses a more conservative protection (Cauchy) and (B2) is more aggressive (t_2).

We compare five methods. The first three methods generate independent and stratified observations $\{\boldsymbol{x}_i\}_{i=1}^n$ from $q_{\boldsymbol{\alpha}_0}(\boldsymbol{x}) = \alpha_0 q_1(\boldsymbol{x}) + (1-\alpha_0)q_1(\boldsymbol{x})$ where $\boldsymbol{\alpha}_0 = (.5, 1-.5)$. The last two methods generate independent and stratified observations $\{\boldsymbol{x}_i\}_{i=1}^{n_0}$ from $q_{\boldsymbol{\alpha}_0}(\boldsymbol{x})$ and $\{\boldsymbol{x}_i\}_{i=n_0+1}^{n}$ from $q_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{x})$ where $\widetilde{\alpha}_1 = \alpha_0 n_0/n + \widehat{\alpha}_1(n-n_0)/n$ and $\widehat{\alpha}_1$ is obtained by the corresponding method. Since the simulation results of regression method are nearly identical to the likelihood approach, we only list MLE and 2MLE here. Specifically, the methods are as follows. For simplicity, only formulas for estimating Z are listed.

UIS (Unprotected Importance Sampling)**:** This is estimator (3.1) with $q(\boldsymbol{x}) = q_2(\boldsymbol{x})$.
SIS (Stratified Importance Sampling)**:** This is estimator (3.3) with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$.
MLE (MLE method)**:** This is estimator (3.6) with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$.
2SIS (Two-Stage Stratified Importance Sampling)**:**

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}, \text{ where } \widehat{\boldsymbol{\alpha}}_1 = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left(\alpha_1 \widetilde{Var}_1 \left[\frac{\pi(x)}{q_{\boldsymbol{\alpha}_0}(x)} \right] + (1-\alpha_1) \widetilde{Var}_2 \left[\frac{\pi(x)}{q_{\boldsymbol{\alpha}_0}(x)} \right] \right)$$

and $\widetilde{Var}_k$ denotes the sample variance with the subset of $\{x_i\}_{i=1}^{n_0}$ which comes from $q_k(x)$. This is the method used in Raghavan and Cox (1998).

2MLE (Two-Stage MLE)**:** This is our proposed method.

The results are shown in Table 2 for estimating Z and μ. Simulation is replicated for 1000 times independently with $n = 4000$ and $n_0 = 400$ in each simulation. We

		Z				μ			
	Method	(A1)	(A2)	(B1)	(B2)	(A1)	(A2)	(B1)	(B2)
$\overline{\widehat{\alpha}}_1$	UIS	0	0	0	0	0	0	0	0
	SIS	.50	.50	.50	.50	.50	.50	.50	.50
	2SIS	.001	.98	.21	.13	.001	.93	.40	.37
	MLE	.50	.50	.50	.50	.50	.50	.50	.50
	2MLE	.004	.98	.72	.999	.001	.91	.42	.30
$nMSE$	UIS	.16	9.4×10^3	3.0	.47	.19	1.2×10^3	66	68
	SIS	.45	28	.15	.16	.34	3.2	.38	.27
	2SIS	.15	16	.087	.028	.20	2.1	.37	.26
	MLE	.27	28	.041	.0094	.34	3.2	.37	.16
	2MLE	.15	16	.037	.0066	.20	2.1	.35	.15

Table 4.2: Comparison of methods for Example 4.1, with each column for one setting. $\overline{\widehat{\alpha}}_1$ is the mean of 1000 estimated mixture proportions and MSE is mean square error of integral estimators.

report the means of $\widehat{Z}$ or $\widehat{\mu}$, the means of $\widehat{\alpha}$ and the mean square error

$$n\widehat{V} = \frac{n}{1000}\sum_{i=1}^{1000}(\widehat{Z}_i - Z)^2 \text{ or } \frac{n}{1000}\sum_{i=1}^{1000}(\widehat{\mu}_i - \mu)^2,$$

where Z and μ are theoretical values.

It is seen that, in (A1) where q_2 is a good proposal by itself, 2SIS and 2MLE choose α_1 close or equal to the smallest allowed value (0.001) for q_1, which minimizes its contamination, and achieves the same efficiency as UIS (using the good proposal only). They are more efficient than SIS and MLE which use equal proportions for both proposal distribution.

In (A2), both 2SIS and 2MLE choose $\hat{\alpha}_1 = .98$, giving much higher proportion to the heavy tail t proposal. It is seen that the normal proposal has a much lighter tail ($\sigma = .4$) that the target ($\sigma = 1$). In this case, $q_1(x)$ is the better proposal. UIS, which uses $q_2(x)$ exclusively, does not have finite variance. Comparing to one stage MLE, the two stage procedure reduces MSE by about 43% and 34% for estimating Z and μ respectively.

In (B1) and (B2), UIS has the largest variance as expected. By using control variates, 2MLE and MLE perform much better than SIS and 2SIS. With the estimated mixture proportions, 2MLE reduces MSE by 10% and 30% for estimating Z in (B1)

	Z				μ			
	(A1)	(A2)	(B1)	(B2)	(A1)	(A2)	(B1)	(B2)
$\overline{\widehat{\alpha}}_1$	.004	.98	.72	.999	.001	.91	.42	.30
α_1^*	.001	.98	.77	.999	.001	.999	.42	.30
$n\widehat{V}$	.150	15.5	.037	.0066	.20	2.06	.35	.15
$\sigma^2(\boldsymbol{\alpha}^*)$	.155	15.9	.035	.0061	.19	1.97	.36	.15

Table 4.3: Comparison between finite sample and asymptotic results. α_1^* is the mixture proportion giving the minimum asymptotic variance, $\widehat{V}$ is the sample variance of integral estimators and $\sigma^2(\boldsymbol{\alpha}^*)$ is the minimum asymptotic variance.

and (B2) respectively, comparing the one-stage MLE. Note that 2MLE obtains a larger estimated optimal proportion for $q_1(x)$ in (B2) than in (B1). Intuitively this is because $q_1(x)$ in (B2) is "closer" to the target integrand. In estimating μ, 2MLE and MLE perform better than SIS and 2SIS, but the two stage 2MLE and one stage MLE are similar, since the estimated optimal proportions are close to .5.

To check the convergence properties of 2MLE, in all four cases we report, in Table 3, a comparisons between the theoretical minimum asymptotic variances and the sample variance of 2MLE, as well as a comparison between the optimal proportions and the average estimated proportions. It is seen that both of them are quite close to the optimal values.

Example 4.2. Consider a rare event problem in Hesterberg (1995). Let $\boldsymbol{X}$ be a three dimensional random variable with independent components (X_1, X_2, X_3) and

$$\boldsymbol{X} = (X_1, X_2, X_3) = \max(0, \boldsymbol{Y}_1 + 10\boldsymbol{d} - \boldsymbol{Z}_1 - \boldsymbol{Z}_2 - \max(500, 3000 - \boldsymbol{Y}_2 - 40\boldsymbol{d})),$$

where $\boldsymbol{Y}_1 \sim N((1600, 1650, 1600), 100^2 I_3)$, $\boldsymbol{Y}_2 \sim N((1600, 1700, 1600), 100^2 I_3)$, $\boldsymbol{Z}_1 \sim \Gamma(100\mathbf{1}_3, (5, 6, 7))$ with $\Gamma(scale, shape)$ denoting the gamma distribution, $\boldsymbol{Z}_2$ has density proportional to $e^{x/100} I_{x\in(0,300)}$, and $\boldsymbol{d} = \max(0,\ 60 - \boldsymbol{t})$, where $\boldsymbol{t} \sim N((54, 52, 55), 5^2 I_3)$. Denote the density of $\boldsymbol{X}$ to be $f(\boldsymbol{x}) = \prod_{j=1}^3 f_j(x_j)$. The targets of interest are

$$P = P\left[\sum_{i=1}^3 X_i > 1200\right] \text{ and } \mu = E\left[80 \cdot \max\left(\sum_{i=1}^3 X_i - 1200, 0\right)\right].$$

The true value of P is about 0.003 and therefore the probability measures the area in

Proposal	(I_1, I_2, I_3)	$E\left[\sum_{j=1}^{3} X_i'\right]$	α_i
$q_1(\boldsymbol{x}) = f(\boldsymbol{x})$			.5
$q_2(\boldsymbol{x})$	$(1,0,0)$	1416	.0035
$q_3(\boldsymbol{x})$	$(0,1,0)$	1266	.028
$q_4(\boldsymbol{x})$	$(0,0,1)$	1616	.0005
$q_5(\boldsymbol{x})$	$(1,1,0)$	1482	.236
$q_6(\boldsymbol{x})$	$(1,0,1)$	1832	.018
$q_7(\boldsymbol{x})$	$(0,1,1)$	1682	.0635
$q_8(\boldsymbol{x})$	$(1,1,1)$	1898	.151

Table 4.4: Parameters setting of the mixture proposal. Each $q_i(\boldsymbol{x})$ is proportional to $exp\left(\sum_{j=1}^{3} \beta_j x_j\right) f(\boldsymbol{x})$, where $\boldsymbol{\beta} = c \cdot (I_1, I_2, I_3)$ and c is selected such that the expectation is equal to the corresponding expectation, e.g. 1416. α_i is the mixture proportion for $q_i(\boldsymbol{x})$. Here (X_1', X_2', X_3') has density $q_i(\boldsymbol{x})$.

				Mixture Proportions							
	Method	mean	var	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\alpha}_5$	$\widehat{\alpha}_6$	$\widehat{\alpha}_7$	$\widehat{\alpha}_8$
P	SIS	3.4×10^{-3}	3.0×10^{-8}	.500	.0035	.028	.0005	.236	.018	.064	.151
	2MLE	3.4×10^{-3}	1.6×10^{-8}	.001	.0040	.038	.0009	.420	.051	.170	.310
μ	SIS	41	1.8	.500	.0035	.028	.0005	.236	.018	.064	.151
	2MLE	41	1.0	.001	.002	.021	.0003	.380	.040	.150	.410

Table 4.5: Comparison between two methods of Example 4.2. SIS is the method of Hesterberg (1995) and 2MLE is our method. $\overline{\widehat{\mu}}$ and $\overline{\widehat{P}}$ are the means of 1000 point estimators, $\overline{\widehat{\alpha}}_i$ are the average mixture proportions and $\widehat{V}$ is the sample variance of 1000 estimators.

the tail of $f(\boldsymbol{x})$. Hesterberg (1995) used $\widehat{Z}_{SIS}$ to estimate P and μ and constructed the proposal distributions by exponential tilting, using $q(\boldsymbol{x}) = c(\boldsymbol{\beta})exp\left(\sum_{j=1}^{3} \beta_j x_j\right) f(\boldsymbol{x})$ with parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. Seven proposals are constructed by setting $\boldsymbol{\beta} = c \cdot (I_1, I_2, I_3)$ where I_j is binary and c is set so that $E\left[\sum_{i=1}^{3} X_i'\right]$ is equal to some pre-determined value, where (X_1', X_2', X_3') follows $q(\boldsymbol{x})$. Including $f(\boldsymbol{x})$ as another proposal component, there are eight proposal components. Hesterberg (1995) provided preset mixture proportions for these proposals, listed in Table 4.

Here we compare the proposed two-stage procedure with the estimator used in Hesterberg (1995) for estimating both P and μ. The results are listed in Table 5. Again, simulation is replicated 1000 times independently with $n = 4000$ and $n_0 = 400$ in each simulation. We report the sample means and variances of $\widehat{P}$ and $\widehat{\mu}$, and the

means of the mixture proportion $\widehat{\alpha}_i$ for $i = 1, \cdots, 8$. Comparing to SIS, it is seen that while the means are the same, 2MLE reduces the variance by 47% for estimating P and 44% for estimating μ. When comparing the proportion set selected by 2MLE and the predetermined proportion set used by SIS, it is seen that some of the proposal considered to be important for SIS is also determined important by 2MLE, such as $q_5(\boldsymbol{x})$ and $q_8(\boldsymbol{x})$. The major difference is that SIS puts too much proportion on $q_1(\boldsymbol{x})$ while 2MLE only selects a very small proportion for it, indicating that only a small proportion is needed for $q_1(\boldsymbol{x})$ in order to guarantee the bounded estimating variance.

Example 4.3. In this example we examine the performance of 2MLE on estimating Value at Risk (VaR) using a Bayesian GARCH(1,1) model for S&P500 index series. Given a probability p and a time horizon d, VaR is the value that a portfolio would encounter a loss greater than or equal to, with probability p over the horizon.

Suppose at time T we have historical log returns $\boldsymbol{y} = \{y_1, \cdots, y_T\}$. Let $R(\boldsymbol{y}_d) = \sum_{k=1}^{d} y_{T+k}$ be the cumulative return in the next d periods, where $\boldsymbol{y}_d = (y_{T+1}, \cdots, y_{T+d})$ and denote $F_{\boldsymbol{y}_d}$ as the CDF of R. Then the d days ahead VaR is defined as

$$VaR_p = inf\left\{x \in \mathbb{R} | F_{\boldsymbol{y}_d}(x) \leq p\right\}.$$

VaR is a widely used measure of market risk (Duffie and Pan, 1997; Jorion, 1997). To obtain the CDF $F_{\boldsymbol{y}_d}$, we model the return series using GARCH model (Engle, 1982; Bollerslev, 1986), a commonly used model for return series and modeling volatility dynamics. Specifically, we use a Bayesian GARCH(1,1) model with normal innovations (Geweke, 1994; Bauwens and Lubrano, 2008),

$$y_t = \varepsilon_t h_t^{1/2}, \quad \varepsilon_t \overset{iid}{\sim} N(0,1), \quad h_t = \phi_0 + \phi_1 y_{t-1}^2 + \beta h_{t-1},$$

where $\phi_0 \geq 0$, $\phi_1 \geq 0$ and $\beta \geq 0$ and $\phi_1 + \beta < 1$ to ensure stationarity. Following Geweke (1994), the prior distributions of $log\phi_0$ and (ϕ_1, β) are selected to be $N(a_0, \sigma_a^2)$ and $U(\phi_1 \geq 0, \beta \geq 0, \phi_1 + \beta < 1)$. Here ϕ_0 is transformed to have the real line as domain, and (ϕ_1, β) follows a uniform distribution in the stationary domain. The hyperpamaters a_0 and σ_a^2 are set to be 1 and 2, respectively. We also use the sample variance for h_0^2 for simplicity.

The Bayesian approach has the advantage of taking into account of parameter estimation variability in the estimation of VaR. Due to the complexity, Monte Carlo method is used. Since VaR is largely a tail property, an appropriate implementation of importance sampling may significantly improve the efficiency. Although VaR is not in the form of integral, it can be estimated easily by empirical quantiles from the Monte Carlo samples. Note that, CDF and probability are in the form of integral. Related literatures about estimating VaR using Importance Sampling can be found in Hoogerheide and Van Dijk (2010), Glasserman et al. (2000) and Dunkel and Weber (2007).

The two-stage algorithm is tested to estimate VaR with $p = 0.05$ and 0.01 and horizons 1, 2 and 5 days, corresponding to 4, 5 and 8 dimensional problem, as there are three parameters in the GARCH(1,1) model. Denote $\boldsymbol{\theta} = (log\phi_0, \phi_1, \beta)$. For each VaR, following the strategy discussed in Section 4.3, we construct the proposal distributions based on the asymptotic variance of the empirical posterior CDF at VaR

$$\sigma_p^2(\boldsymbol{\alpha}) = \int \frac{\left[\left(1_{\{R(\boldsymbol{y}_d)\leq VaR\}}(\boldsymbol{y}_d) - p\right)\pi(\boldsymbol{y}_d,\boldsymbol{\theta}) - \boldsymbol{\beta}_{\boldsymbol{\alpha}}^T \boldsymbol{g}(\boldsymbol{y}_d,\boldsymbol{\theta})\right]^2}{q_{\boldsymbol{\alpha}}(\boldsymbol{y}_d,\boldsymbol{\theta})} d\boldsymbol{y}_d d\boldsymbol{\theta}, \tag{4.9}$$

where $\pi(\boldsymbol{y}_d,\boldsymbol{\theta}) = \prod_{k=1}^{T+d} p(y_k|y_{k-1},\boldsymbol{\theta})p(\boldsymbol{\theta})$, $p(y_k|y_{k-1},\boldsymbol{\theta})$ is the innovation density and $p(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$.

Expression (4.9) is not the variance of the VaR estimator. However, Hoogerheide and Van Dijk (2010) showed that the asymptotic variance of $\widehat{VaR}_p$ can be approximated by $\sigma_p^2(\boldsymbol{\alpha})$ times a constant which does not depend on the proposal density. Since it is difficult to sample from $\pi(\boldsymbol{y}_d,\boldsymbol{\theta})$ directly, we approximate it by the mixture of

$$\begin{aligned} q_1(\boldsymbol{y}_d,\boldsymbol{\theta}) &= \prod_{k=1}^{T+d} p(y_k|y_{1:k-1},\boldsymbol{\theta})q_N(\boldsymbol{\theta}) \text{ and} \\ q_2(\boldsymbol{y}_d,\boldsymbol{\theta}) &= q_*(y_{T+d}|y_{T+d-1},\boldsymbol{\theta}) \prod_{k=1}^{T+d-1} p(y_k|y_{1:k-1},\boldsymbol{\theta})q_N(\boldsymbol{\theta}), \end{aligned}$$

where $q_N(\boldsymbol{\theta})$ is the normal distribution with the mean vector being the MLE $\widehat{\boldsymbol{\theta}}$ and the covariance matrix Σ_N being the negative inverse Hessian matrix of $\pi(\boldsymbol{y},\boldsymbol{\theta})$ at $\widehat{\boldsymbol{\theta}}$, inflated by a constant to allow a wider coverage. We use $q_*(y_{T+d}|y_{1:T+d-1},\boldsymbol{\theta}) \sim N(-h_{T+d}^{1/2}, h_{T+d})$ for the proposal $q_2(\boldsymbol{y}_d,\boldsymbol{\theta})$. It tries to cover the tail (large loss on the last day of the horizon). Similar proposals can be constructed by considering other

potential situations of large loss, but only this one is included in the current example.

With the approximation of $\pi(\boldsymbol{y}_d, \boldsymbol{\theta})$, the heavier tail components can be constructed by modifying the tails of q_1 and q_2. Then the following two proposals are included as the heavier tail components:

$$\begin{aligned} q_3(\boldsymbol{y}_d, \boldsymbol{\theta}) &= \prod_{k=1}^{T+d} p(y_k|y_{1:k-1}, \boldsymbol{\theta}) q_t(\boldsymbol{\theta}) \quad \text{and} \\ q_4(\boldsymbol{y}_d, \boldsymbol{\theta}) &= q_*(y_{T+d}|y_{1:T+d-1}, \boldsymbol{\theta}) \prod_{k=1}^{T+d-1} p(y_k|y_{1:k-1}, \boldsymbol{\theta}) q_t(\boldsymbol{\theta}), \end{aligned}$$

where $q_t(\boldsymbol{\theta})$ is the product of three location-scale generalization of t_1 densities with the means being $\widehat{\boldsymbol{\theta}}$ and the squared scale parameters being the diagonal elements of Σ_N, and q_* is the same as in the construction of q_2. Since $\pi(\boldsymbol{y}_d, \boldsymbol{\theta})/q_3(\boldsymbol{y}_d, \boldsymbol{\theta}) = p(\boldsymbol{\theta})/q_t(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ is the prior distribution with exponentially decreasing tail, $q_3(\boldsymbol{y}_d, \boldsymbol{\theta})$ has heavier tail than $\pi(\boldsymbol{y}_d, \boldsymbol{\theta})$. Similarly, q_3 and q_4 have heavier tail than q_1, q_2 and proposals below, and therefore only mixture proportions for q_3 and q_4 need to be restricted.

To incorporate the integrand as discussed in Section 4.3, we further extend $q_1(\boldsymbol{y}_d, \boldsymbol{\theta})$ and $q_2(\boldsymbol{y}_d, \boldsymbol{\theta})$ to include

$$\begin{aligned} q_5(\boldsymbol{y}_d, \boldsymbol{\theta}) &\propto 1_{\{y_{T+d} \le VaR_{.05} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d}) q_1(\boldsymbol{y}_d, \boldsymbol{\theta}) \quad \text{and} \\ q_6(\boldsymbol{y}_d, \boldsymbol{\theta}) &\propto 1_{\{y_{T+d} \le VaR_{.05} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d}) q_2(\boldsymbol{y}_d, \boldsymbol{\theta}), \end{aligned}$$

Here the truncation is done only on y_{T+d}, instead of the more accurate but computationally expensive truncation of $\sum_{k=1}^{d} y_{T+k} \le VaR_{.05}$ under joint normal distribution.

The estimation of $VaR_{.01}$ can be done simultaneously by including the following component proposals

$$\begin{aligned} q_7(\boldsymbol{y}_d, \boldsymbol{\theta}) &\propto 1_{\{y_{T+d} \le VaR_{.01} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d}) q_1(\boldsymbol{y}_d, \boldsymbol{\theta}) \quad \text{and} \\ q_8(\boldsymbol{y}_d, \boldsymbol{\theta}) &\propto 1_{\{y_{T+d} \le VaR_{.01} - \sum_{k=1}^{d-1} y_{T+k}\}}(y_{T+d}) q_2(\boldsymbol{y}_d, \boldsymbol{\theta}). \end{aligned}$$

Overall, $q_1(\boldsymbol{y}_d, \boldsymbol{\theta})$ to $q_8(\boldsymbol{y}_d, \boldsymbol{\theta})$ are used as component proposal distributions.

Since our objective is to estimate $VaR_{.05}$ and $VaR_{.01}$ simultaneously, in the pilot stage we estimate the optimal mixture proportions by minimizing the sum of variances

		$p = 0.05$		$p = 0.01$	
horizon	Method	$\overline{\widehat{VaR}}$	$\widehat{V}$	$\overline{\widehat{VaR}}$	$\widehat{V}$
1 day	MLE	-1.332	$14e-5$	-1.894	$20e-5$
	2MLE	-1.333	$3.8e-5$	-1.895	$4.6e-5$
2 days	MLE	-1.886	$5.1e-4$	-2.773	$12e-4$
	2MLE	-1.886	$1.5e-4$	-2.771	$3.5e-4$
5 days	MLE	-2.997	$17e-4$	-4.432	$5.9e-3$
	2MLE	-2.996	$5.4e-4$	-4.424	$1.8e-3$

Table 4.6: Comparison between MLE and 2MLE in Example 4.3. $\overline{\widehat{VaR}}$ is the average of 300 point estimators and $\widehat{V}$ is the sample variance of 300 estimators.

of the two estimators. Since $q_5, \cdots, q_8$ involve the unknown $VaR_{.05}$ and $VaR_{.01}$, the first stage sampling is modified as follows.

1. Generate pilot samples from q_1 to q_4 with sample size $n_0/8$ each;

2. Estimate $VaR_{.05}$ using the pilot samples from step 1. Replace $VaR_{.05}$ in q_5 and q_6 with the estimate and generate pilot samples from them, with sample size $n_0/8$ each.

3. Estimate $VaR_{.01}$ using the pilot samples from steps 1 and 2. Replace $VaR_{.01}$ in q_7 and q_8 with the estimate and generate pilot samples from them, with sample size $n_0/8$ each.

4. Obtain $\widehat{\boldsymbol{\alpha}}$ by minimizing $\hat{\tau}^2_{.05}(\boldsymbol{\alpha}) + \hat{\tau}^2_{.01}(\boldsymbol{\alpha})$ where $\hat{\tau}^2_p(\boldsymbol{\alpha})$ is the estimator for $\sigma^2_p(\boldsymbol{\alpha})$ using all samples in the first three steps.

Here we compare the two-stage procedure 2MLE with the one stage MLE with equal mixture proportions. The log returns of S&P500 index from September 28, 2010 to July 13, 2011 are used, with total 200 observations. The simulation is replicated for 300 times independently with $n = 4 \times 10^6$ and $n_0 = 8 \times 10^4$ in each simulation. δ is selected to be .001.

The summary of estimation results and the estimated mixture proportions $\widehat{\boldsymbol{\alpha}}$ are listed in Table 6 and 7. From Table 6, it is seen that 2MLE's Monte Carlo variance is about 23% to 32% of the variance of MLE while there is almost no difference in

VaR	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\alpha}_5$	$\widehat{\alpha}_6$	$\widehat{\alpha}_7$	$\widehat{\alpha}_8$
1 day	1e-3	8e-4	3.4e-1	6.2e-1	9e-3	1.7e-2	8e-3	2e-3
2 days	1e-3	1e-3	3.5e-1	6.0e-1	2.4e-2	3e-3	9e-3	2e-3
5 days	1e-3	6e-4	4.3e-1	5.4e-1	2.0e-2	8e-4	4e-3	5e-4

Table 4.7: Summary of mixture proportions estimated from stage 1. The average over 300 simulations are reported. $\widehat{\alpha}_1$ to $\widehat{\alpha}_8$ correspond to the mixture proportions assigned to q_1 to q_8.

the mean. Table 7 shows that the two-stage algorithm assigns most of the mixture proportions to q_3 and q_4 which indicates that these two heavy tail component proposals are more important than the others. This is probably due to the fact that q_1 and q_2 do not cover the high density area of target distribution sufficiently, resulted in the preference to q_3 and q_4. Compared with MLE, the optimization in the pilot stage of 2MLE requires additional computing time which is about 20% more in practice.

Finally, we report some interesting insights on the comparison between MLE and 2MLE. By multiplying a scaling constant c^2 to the covariance matrix Σ_N used in the proposal q_1, all the related component proposals are made either more dispersed for $c > 1$ or more concentrated for $c < 1$. Since $\sigma_p^2(\boldsymbol{\alpha})$ is proportional to the estimation variance of VaR and the proportion does not depend on the proposal density, the trajectories of estimated $\sigma_p^2(\boldsymbol{\alpha}^*)$ and $\sigma_p^2(\boldsymbol{\alpha_0})$ as function of c are given in Figure 1 to illustrate how the quality of proposal distribution affect the performance of 2MLE and MLE.

It is seen that 2MLE is always better than MLE. Most interestingly, it shows that the performance of both methods depends on the quality of the proposal distributions, but 2MLE is much less sensitive to the proposal distributions and has more robust performance than MLE. This is due to 2MLE's ability to automatically adjust mixture proportion for the most efficient estimation. The simulation results (not shown here) show that, when c is small, 2MLE tends to assign most of the mixture proportions to the heavy tail q_3 and q_4. This insight re-enforces the notion that the two-stage approach not only improves upon the one stage approach, but also alleviate to some extend the difficulty of selecting proposal distributions.

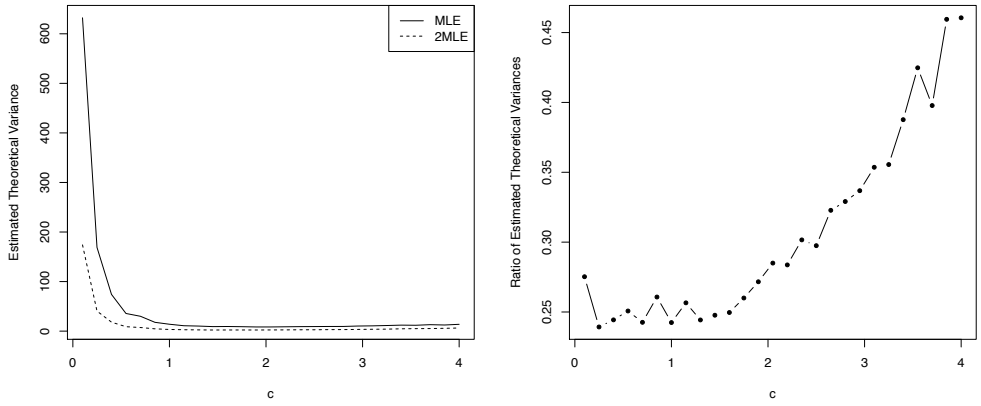

Figure 4.1: The left figure gives trajectories of estimated $\sigma_p^2(\boldsymbol{\alpha}^*)$ and $\sigma_p^2(\boldsymbol{\alpha_0})$, corresponding to MLE and 2MLE methods respectively, with respect to the scaling constant c. c ranges from .1 to 4. For each c, the theoretical variances are estimated using one Monte Carlo sample, and the average over 10 replicates is reported. The right figure gives the trajectory of ratio of estimated $\sigma_p^2(\boldsymbol{\alpha}^*)$ over estimated $\sigma_p^2(\boldsymbol{\alpha_0})$.

4.6 Summary

In this chapter, we proposed a two-stage procedure to select the optimal mixture proportions for the regression estimator in Owen and Zhou (2000) and MLE estimator in Tan (2004), and established the corresponding theoretical framework. The two-stage procedure significantly improved the existing methods in four aspects. First, the proposed estimator is asymptotically the best among all the estimators proposed in Owen and Zhou (2000), Tan (2004) and Raghavan and Cox (1998). Second, the criterion function of our pilot stage optimization is convex in its arguments, and therefore it is guaranteed that the optimization converges to the global minimum. Third, since there is no simple intuition in selecting the proportions for Owen and Zhou's (2000) regression estimator and Tan's (2004) MLE estimator, the proposed automatic procedure makes it much easier and safer to use mixture distributions for importance sampling. Finally, the automatic determination of the mixture proportion alleviates the difficulty of choosing the set of proposal distributions to be considered in the mixture, as it serves as a selection and discrimination tool and hence allows users to include more potential proposal distributions for consideration.

4.7 Technical Proof

For simplicity, we only consider two proposal distributions. Then $\boldsymbol{\alpha} = (\alpha_1, 1-\alpha_1)$ and $\boldsymbol{\gamma} = (\gamma_1, 1-\gamma_1)$. The proofs can be extended to the case of more than two proposals. To begin with, we establish the consistency of $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, 1-\widehat{\alpha}_1)$. Note that $\widehat{\alpha}_1$ is equivalently a component of the bivariate M-estimate $(\widehat{\alpha}_1, \widehat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}}\ n_0^{-1} \sum_{i=1}^{n_0} m(x_i; \alpha, \beta)$, where $m(x; \alpha_1, \beta) = \left[\pi(x) - \beta g(x)\right]^2 / \left[q_{\boldsymbol{\alpha}}(x) q_{\boldsymbol{\gamma}}(x)\right]$. Let $M(\alpha_1, \beta) = \int m(x; \alpha_1, \beta) q_{\boldsymbol{\gamma}}(x) dx$ and $(\alpha_1^*, \beta^*) = \underset{\alpha_1,\beta}{\operatorname{argmin}}\ M(\alpha_1, \beta)$. Meanwhile, $M(\alpha_1, \beta)$ and $\sigma_Z^2(\boldsymbol{\alpha})$ are strictly convex functions.

Lemma 4.1. *It holds that*

$$
\begin{aligned}
&(\widehat{\alpha}_1, \widehat{\beta}) \xrightarrow{P} (\alpha_1^*, \beta^*), \\
&(\widehat{\alpha}_1, \widehat{\beta}) = (\alpha_1^*, \beta^*) - \frac{1}{2\sqrt{n_0}} V^{-1} \widehat{U} + o_p(\frac{1}{\sqrt{n_0}}),
\end{aligned}
$$

where V and $\widehat{U}$ are given in Lemma 4.2. Then $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}_1, 1-\widetilde{\alpha}_1) \xrightarrow{P} (\alpha_1^*, 1-\alpha_1^*)$. Meanwhile, $M(\alpha_1, \beta)$, $\sigma_Z^2(\boldsymbol{\alpha})$ and $\widehat{\sigma}^2(\boldsymbol{\alpha})$ are strictly convex functions.

Proof. Note that $m(x; \alpha, \beta)$ is convex since its Hessian matrix

$$
D^2 m(x; \alpha_1, \beta) = \frac{2g(x)^2}{q_{\boldsymbol{\alpha}}(x) q_{\boldsymbol{\gamma}}(x)} \begin{pmatrix} \frac{(\pi(x)-\beta g(x))^2}{q_{\boldsymbol{\alpha}}(x)^2} & \frac{\pi(x)-\beta g(x)}{q_{\boldsymbol{\alpha}}(x)} \\ \frac{\pi(x)-\beta g(x)}{q_{\boldsymbol{\alpha}}(x)} & 1 \end{pmatrix}
$$

is a positive semidefinite matrix. Then the consistency of $\left(\widehat{\alpha}_1, \widehat{\beta}\right)$ can be proved by verifying conditions $1-3$ in Haberman (1989) for M-estimators by convex minimization. First, the parameter set $\Theta = [\delta, 1-\delta] \times \mathbb{R}$ of (α_1, β) is convex and closed. Second, (α_1^*, β^*) is unique. By Durrett (1996, Appendix 9), the differentiation and integration in $M(\alpha, \beta)$ can be exchanged so that

$$
D^2 M(\alpha_1, \beta) = \begin{pmatrix} 2\int \frac{[\pi(x)-\beta g(x)]^2 g(x)^2}{q_{\boldsymbol{\alpha}}(x)^3} dx & 2\int \frac{[\pi(x)-\beta g(x)] g(x)^2}{q_{\boldsymbol{\alpha}}(x)^2} dx \\ 2\int \frac{[\pi(x)-\beta g(x)] g(x)^2}{q_{\boldsymbol{\alpha}}(x)^2} dx & 2\int \frac{g(x)^2}{q_{\boldsymbol{\alpha}}(x)} dx \end{pmatrix}.
$$

For any bivariate vector v, $v^T \left\{D^2 M(\alpha_1, \beta)\right\} v \geq 0$ and the equality holds only when $\pi(x) \equiv c_1 q_1(x) + c_2 q_2(x)$ for some c_1 and c_2. By condition $(C5)$, $D^2 M(\alpha_1, \beta)$ is positive definite. Therefore $M(\alpha_1, \beta)$ is strictly convex and (α_1^*, β^*) is unique. Third, let $W =$

$(\delta, 1-\delta) \times \mathbb{R}$. By condition $(C3)$, $M(\alpha_1, \beta) < \infty$ for any $(\alpha_1, \beta) \in W$.

The expansion of $\left(\widehat{\alpha}_1, \widehat{\beta}\right)$ can be found in the proof of Haberman (1989, Theorem 6.1) by verifying his conditions 7 and 10. First, $D^2 M(\alpha_1^*, \beta^*)$ is positive definite as mentioned above. Second, the gradient of $m(x; \alpha_1, \beta)$ satisfies $E|Dm(x; \alpha_1, \beta)|^2 < \infty$. Therefore the convergence of $(\widetilde{\alpha}_1, 1-\widetilde{\alpha}_1)$ holds because $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\gamma} n_0/n + \widehat{\boldsymbol{\alpha}}(n-n_0)/n$ and $n_0 = o(n)$.

Finally, with the strict convexity of $M(\alpha_1, \beta)$ which is stated above, the strict convexity of $\sigma_Z^2(\boldsymbol{\alpha})$ can be seen by the facts that $\sigma_Z^2(\boldsymbol{\alpha}) = \min_{\beta} M(\alpha_1, \beta)$ and

$$\min_{\beta} M(\lambda\alpha_1 + (1-\lambda)\alpha_2, \beta) = \min_{\beta_1} \min_{\beta_2} M(\lambda\alpha_1 + (1-\lambda)\alpha_2, \lambda\beta_1 + (1-\lambda)\beta_2)$$

for any α_1, α_2 and $\lambda \in [0,1]$. The strict convexity of $\widehat{\sigma}^2(\boldsymbol{\alpha})$ can be proved similarly. □

The following expansion of $(\widehat{\alpha}_1, \widehat{\beta})$ will be used in the higher order calculation of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$.

Lemma 4.2. *It holds that*

$$(\widehat{\alpha}_1, \widehat{\beta}) = (\alpha_1^*, \beta^*) + \frac{1}{2\sqrt{n_0}} V^{-1}\widehat{U} - \frac{1}{2n_0} V^{-1}\left\{(\widehat{V} - \sqrt{n_0}V)V^{-1}\widehat{U} + \widehat{W}\right\} + o_p\left(\frac{1}{n_0}\right),$$

where $\widehat{W}$ *is a random variable of order* $O_p(1)$,

$$\widehat{U} = \begin{pmatrix} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \frac{(\pi(x_i) - \beta^* g(x_i))^2 g(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)^2 q_{\boldsymbol{\gamma}}(x_i)} \\ \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \frac{2(\pi(x_i) - \beta^* g(x_i)) g(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i) q_{\boldsymbol{\gamma}}(x_i)} \end{pmatrix},$$

$$V = \begin{pmatrix} \int \frac{[\pi(x) - \beta^* g(x)]^2 g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3} dx & \int \frac{[\pi(x) - \beta^* g(x)] g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^2} dx \\ \int \frac{[\pi(x) - \beta^* g(x)] g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^2} dx & \int \frac{g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)} dx \end{pmatrix},$$

and $\widehat{V} = \begin{pmatrix} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \frac{[\pi(x_i) - \beta^* g(x_i)]^2 g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i)^3 q_{\boldsymbol{\gamma}}(x_i)} & \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \frac{[\pi(x_i) - \beta^* g(x_i)] g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i)^2 q_{\boldsymbol{\gamma}}(x_i)} \\ \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \frac{[\pi(x_i) - \beta^* g(x_i)] g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i)^2 q_{\boldsymbol{\gamma}}(x_i)} & \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \frac{g(x_i)^2}{q_{\boldsymbol{\alpha}^*}(x_i) q_{\boldsymbol{\gamma}}(x_i)} \end{pmatrix}.$

Proof. Note that $\widehat{U} = n_0^{-1/2} \sum_{i=1}^{n_0} Dm(x_i; \alpha_1^*, \beta^*)$, $\widehat{V} = n_0^{-1/2} \sum_{i=1}^{n_0} D^2 m(x_i; \alpha_1^*, \beta^*)$ and $V = \int D^2 m(x; \alpha_1^*, \beta^*) dx$. Then by Taylor expansion around (α_1^*, β^*) on $n_0^{-1} \sum_{i=1}^{n_0} Dm(x_i; \widehat{\alpha}_1, \widehat{\beta}) = 0$ and the convergence of $(\widehat{\alpha}_1, \widehat{\beta})$, we have

$$0 = -\frac{1}{\sqrt{n_0}}\widehat{U} + \frac{2}{\sqrt{n_0}}\widehat{V}\begin{pmatrix} \widehat{\alpha}_1 - \alpha_1^* \\ \widehat{\beta} - \beta^* \end{pmatrix} + \frac{1}{n_0}\widehat{W} + o_p\left(\frac{1}{n_0}\right),$$

then

$$\begin{pmatrix} \widehat{\alpha}_1 - \alpha_1^* \\ \widehat{\beta} - \beta^* \end{pmatrix} = \frac{1}{2\sqrt{n_0}} V^{-1}\widehat{U} - \frac{1}{n_0} V^{-1} \left\{ (\widehat{V} - \sqrt{n_0}V) \cdot \sqrt{n_0} \begin{pmatrix} \widehat{\alpha}_1 - \alpha_1^* \\ \widehat{\beta} - \beta^* \end{pmatrix} + \widehat{W} \right\} + o_p\left(\frac{1}{n_0}\right).$$

The expansion of $(\widehat{\alpha}_1, \widehat{\beta})$ follows by substituting $(\widehat{\alpha}_1 - \alpha_1^*, \widehat{\beta} - \beta^*)$ to the RHS of the above equation. □

The combined sample $\{x_1, \cdots, x_n\}$ can be split into four parts by distributions q_1 or q_2 and first or second stages. Denote I_{jk} to be the index set of observations from the jth stage and q_k, i.e. $I_{11} = \{1, \cdots, n_0\gamma_1\}$, $I_{12} = \{n_0\gamma_1 + 1, \cdots, n_0\}$, $I_{21} = \{n_0 + 1, \cdots, n_0 + \lceil (n - n_0)\widehat{\alpha}_1 \rceil\}$, $I_{22} = \{n_0 + \lceil (n - n_0)\widehat{\alpha}_1 \rceil + 1, \cdots, n\}$ where $\lceil x \rceil$ means the largest integer smaller than x, and n_{jk} to be the size of I_{jk}. Here we can use for the index, where $\lfloor x \rfloor$ is the largest integer smaller than x, to define I_{jk}. But for investigating the asymptotic behavior, the difference can be ignored. We will use the decomposition

$$\begin{aligned} G_n\tau(x) &= \sqrt{\frac{n_0}{n}} \left\{ \sum_{k=1}^{2} \sqrt{\gamma_k} \cdot \sqrt{n_{1k}} \left(\frac{1}{n_{1k}} \sum_{i \in I_{1k}} \tau(x_i) - \int \tau(x) q_k(x) dx \right) \right\} \\ &\quad + \sqrt{\frac{n - n_0}{n}} \left\{ \sum_{k=1}^{2} \sqrt{\widehat{\alpha}_k} \cdot \sqrt{n_{2k}} \left(\frac{1}{n_{2k}} \sum_{i \in I_{2k}} \tau(x_i) - \int \tau(x) q_k(x) dx \right) \right\} \\ &\equiv \sqrt{\frac{n_0}{n}} \left\{ \sqrt{\gamma_1} G_{11}\tau(x) + \sqrt{\gamma_2} G_{12}\tau(x) \right\} \\ &\quad + \sqrt{\frac{n - n_0}{n}} \left\{ \sqrt{\widehat{\alpha}_1} G_{21}\tau(x) + \sqrt{\widehat{\alpha}_2} G_{22}\tau(x) \right\}. \end{aligned} \tag{4.10}$$

The following lemma shows the convergence of $\widehat{Z}_{SIS}$ with $\widetilde{\boldsymbol{\alpha}}$ as mixture proportion.

Lemma 4.3. *For any integrable function $h(x)$ satisfying $Var_{\boldsymbol{\alpha}}[h(X)/q_{\boldsymbol{\alpha}}(X)] < \infty$ for every $\alpha_1 \in [\delta, 1 - \delta]$, it holds that*

$$\frac{1}{n} \sum_{i=1}^{n} \frac{h(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)} \xrightarrow{P} \int h(x) dx,$$

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{h(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)} - \int h(x) dx \right) \xrightarrow{\mathcal{L}} N\left(0, \sum_{k=1}^{2} \alpha_k^* Var_k \left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)} \right] \right).$$

where Var_k denote the variance under distribution density $q_k(x)$.

Proof. We only need to prove asymptotic normality since it implies the consistency. Using decomposition (4.10) we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}-\int h(x)dx\right)=\sqrt{\frac{n_0}{n}}\left\{\sqrt{\gamma_1}G_{11}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)+\sqrt{\gamma_2}G_{12}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)\right\}$$
$$+\sqrt{\frac{n-n_0}{n}}\left\{\sqrt{\widehat{\alpha}_1}G_{21}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)+\sqrt{\widehat{\alpha}_2}G_{22}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)\right\} \quad (4.11)$$

The asymptotic normality will be implied by showing that the first two terms in (4.11) are of order $o_p(1)$ and the remaining is asymptotic normal.

For the first two terms, we prove that the collection of functions $\{h(x)/q_{\boldsymbol{\alpha}}(x)\}_{\alpha_1\in[\delta,1-\delta]}$ is a Donsker class under either probability measures q_1 or q_2 by verifying the three conditions in Van der Vaart (2000, Example 19.7). In fact, the parameter α is in a bounded set; $|h(x)/q_{\alpha_1}(x)-h(x)/q_{\alpha_2}(x)|\leq|m(x,\alpha_1,\alpha_2)|\cdot|\alpha_1-\alpha_2|$ for every α_1,α_2 where $m(x,\alpha_1,\alpha_2)=h(x)g(x)/(q_{\alpha_1}(x)q_{\alpha_2}(x))$, and $\int|m(x,\alpha_1,\alpha_2)|^2q_k(x)dx<\infty$. By Van der Vaart (2000, Lemma 19.24) and Lemma 4.1, we have $G_{1k}(h/q_{\widetilde{\boldsymbol{\alpha}}})=G_{1k}(h/q_{\boldsymbol{\alpha}^*})+o_p(1)$, $k=1,2$. Then by Central Limit Theorem and $n_0=o(n)$, the first two terms in (4.11) are of order $o_p(1)$.

For the last two terms, similarly, we argue that $G_{2k}(h/q_{\widetilde{\boldsymbol{\alpha}}})=G_{2k}(h/q_{\boldsymbol{\alpha}^*})+o_p(1)$ by a modification of Van der Vaart (2000, Lemma 19.24) to handle random sample size. In fact, the key condition for his results, namely weak convergence of $G_{2k}(h/q_{\boldsymbol{\alpha}^*})$, is guaranteed by Van Der Vaart and Wellner (1996, Theorem 3.5.1). Then by the independence between $\hat{\alpha}_1$ and observations in $\{x_i\}_{i=n_0+1}^{n}$ and an extension of Chow and Teicher (2003, section 9.4), we have

$$\begin{pmatrix}G_{21}(h/q_{\widetilde{\boldsymbol{\alpha}}})\\G_{22}(h/q_{\widetilde{\boldsymbol{\alpha}}})\end{pmatrix}\xrightarrow{\mathfrak{L}}N\left(0,\begin{pmatrix}Var_1\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]&0\\0&Var_2\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\end{pmatrix}\right)$$

and by Slutsky's theorem, $\sqrt{\widehat{\alpha}_1}G_{21}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)+\sqrt{\widehat{\alpha}_2}G_{22}\left(\frac{h(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right)$
$\xrightarrow{\mathfrak{L}}N\left(0,\sum_{k=1}^{2}\alpha_k^*Var_k\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\right)$. Therefore the lemma holds. □

In the above proof, only the consistency of $\widetilde{\boldsymbol{\alpha}}$ is used. If $\widetilde{\boldsymbol{\alpha}}$ is replaced by other consistent mixture proportion, the convergence properties still hold.

Corollary 4.1. *For any $\boldsymbol{\alpha}$ satisfying $\boldsymbol{\alpha} \xrightarrow{P} \boldsymbol{\alpha}^*$, it holds that*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\boldsymbol{\alpha}}(x_i)} \xrightarrow{P} \int h(x)dx,$$
$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h(x_i)}{q_{\boldsymbol{\alpha}}(x_i)} - \int h(x)dx\right) \xrightarrow{\mathfrak{L}} N\left(0, \sum_{k=1}^{2}\alpha_k^* Var_k\left[\frac{h(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]\right).$$

We also need the convergence of $\widetilde{\zeta}$ in $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ where

$$\widetilde{\zeta} = \underset{\boldsymbol{\zeta}}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left[q_{\widetilde{\boldsymbol{\alpha}}}(x_i) + \boldsymbol{\zeta}^T \boldsymbol{g}(x_i)\right].$$

Lemma 4.4. *The following convergence properties for $\widetilde{\zeta}$ hold:*

$$\widetilde{\zeta} \xrightarrow{P} 0 \text{ and } \sqrt{n}\widetilde{\zeta} \xrightarrow{\mathfrak{L}} N\left(0, \frac{\sum_{k=1}^{2}\alpha_k^* Var_k\left[g(X)/q_{\boldsymbol{\alpha}^*}(X)\right]}{\left(\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx\right)^2}\right).$$

Proof. The random variable $\sqrt{n}\widetilde{\zeta}$ is the minimizer of convex function $\psi(s) = \sum_{i=1}^{n}\log\left(q_{\widetilde{\boldsymbol{\alpha}}}(x_i) + g(x_i)s/\sqrt{n}\right)$. By verifying the condition of Hjort and Pollard (1994, basic corollary), we have the expansion

$$\sqrt{n}\widetilde{\zeta} = \frac{n^{-1/2}\sum_{i=1}^{n} g(x_i)/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx} + o_p(1)$$

and then the convergence of $\widetilde{\zeta}$ follows by Lemma 4.3. By Taylor expansion around 0, we have

$$\psi(s) = \sum_{i=1}^{n}\log\left(q_{\widetilde{\boldsymbol{\alpha}}}(x_i)\right) + \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}\right\}s - \left\{\frac{1}{2n}\sum_{i=1}^{n}\frac{g(x_i)^2}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2}\right\}s^2 + R_n(s),$$
$$\text{where } R_n(s) = \left\{\frac{1}{3n\sqrt{n}}\sum_{i=1}^{n}\frac{g(x_i)^3}{q_{\boldsymbol{\xi}}(x_i)^3}\right\}s^3 \text{ and } \xi_1 \text{ is between } \widetilde{\alpha}_1 \text{ and } \widetilde{\alpha}_1 + s/\sqrt{n}.$$

For fixed s, $\boldsymbol{\xi} \xrightarrow{P} \boldsymbol{\alpha}^*$ by $\widetilde{\boldsymbol{\alpha}} \xrightarrow{P} \boldsymbol{\alpha}^*$. Then $R_n(s) \xrightarrow{P} 0$ by Corollary 4.1, and condition of Hjort and Pollard (1994, basic corollary) holds. □

Proof of Theorem 4.1. By Taylor expansion of $n^{-1}\sum_{i=1}^{n} g(x_i)/(q_{\widetilde{\boldsymbol{\alpha}}}(x_i) + \widehat{\zeta}g(x_i)) = 0$, $\widehat{\zeta}$ can be expanded as

$$\widehat{\zeta} = \frac{\frac{1}{n}\sum_{i=1}^{n} g(x_i)/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)}{\frac{1}{n}\sum_{i=1}^{n} g(x_i)^2/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2 + \widehat{\zeta}\cdot\frac{1}{n}\sum_{i=1}^{n} g(x_i)^3/(q_{\widetilde{\boldsymbol{\alpha}}}(x_i) + \dot{\zeta}g(x_i))^3}$$
$$\equiv S_g/S_{gg}, \text{ where } \dot{\zeta} \text{ is between 0 and } \widetilde{\zeta}.$$

By Taylor expansion, we have

$$\begin{aligned}\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(x_i)}{(\widetilde{\alpha}_1+\widetilde{\zeta})q_1(x_i)+(\widetilde{\alpha}_2-\widetilde{\zeta})q_2(x_i)} \\ &= S_\pi - \left(S_{gg}^{-1}S_{\pi g}-\beta^*\right)S_g, \qquad (4.12)\\ \text{where } S_\pi &= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(x_i)-\beta^* g(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)},\ S_{\pi g}=\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(x_i)g(x_i)}{(q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+\ddot{\zeta}g(x_i))^2},\end{aligned}$$

and $\ddot{\zeta}$ is between 0 and $\widetilde{\zeta}$.

Since $\widetilde{\boldsymbol{\alpha}} \xrightarrow{P} \boldsymbol{\alpha}^*$, $\widetilde{\zeta} \xrightarrow{P} 0$ and $q_{\widetilde{\boldsymbol{\alpha}}}(x_i)+\dot{\zeta}g(x_i) = q_{\widetilde{\boldsymbol{\alpha}}+(\dot{\zeta},-\dot{\zeta})}(x_i)$, we have

$$\begin{aligned}&S_\pi \xrightarrow{P} Z,\ \sqrt{n}(S_\pi - Z) \xrightarrow{\mathfrak{L}} N(0, Var_{\boldsymbol{\alpha}^*}\left[\frac{\pi(X)-\beta^* g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]),\\ &\sqrt{n}S_g \xrightarrow{\mathfrak{L}} N(0, Var_{\boldsymbol{\alpha}^*}\left[\frac{g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]),\\ &S_{\pi g} \xrightarrow{P} Cov_{\boldsymbol{\alpha}^*}\left[\frac{\pi(X)}{q_{\boldsymbol{\alpha}^*}(X)},\frac{g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right] \text{ and } S_{gg} \xrightarrow{P} Var_{\boldsymbol{\alpha}^*}\left[\frac{g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right].\end{aligned}$$

by Lemma 4.3 and Corollary 4.1. Then plugging the above results in (4.12), Slutsky's theorem gives that

$$\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}}) \xrightarrow{P} Z \text{ and} \sqrt{n}\left(\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}}) - Z\right) \xrightarrow{\mathfrak{L}} N(0, Var_{\boldsymbol{\alpha}^*}\left[\frac{\pi(X)-\beta^* g(X)}{q_{\boldsymbol{\alpha}^*}(X)}\right]).$$

Similarly, the consistency and asymptotic normality of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ hold by the decomposition

$$\begin{aligned}\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(x_i)-\widehat{\beta}_{\widetilde{\boldsymbol{\alpha}}}g(x_i)}{q_{\boldsymbol{\alpha}}(x_i)} \\ &= S_\pi - \left[\widetilde{Var}\left(\frac{g(X)}{q_{\widetilde{\boldsymbol{\alpha}}}(X)}\right)^{-1}\widetilde{Cov}\left(\frac{\pi(X)}{q_{\widetilde{\boldsymbol{\alpha}}}(X)},\frac{g(X)}{q_{\widetilde{\boldsymbol{\alpha}}}(X)}\right)-\beta^*\right]\cdot S_g.\end{aligned}$$

□

Proof of Proposition 4.1. Denote $G_n\tau(x) = \sqrt{n}\left[n^{-1}\sum_{i=1}^{n}\tau(x_i) - \int \tau(x)q_{\widetilde{\boldsymbol{\alpha}}}(x)dx\right]$. By

(4.12) and Taylor expansion around α_1^* we have

$$\begin{aligned}
\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}}) - Z &= \frac{1}{\sqrt{n}}\left\{G_n \frac{\pi(x)-\beta^* g(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right\} + \frac{1}{\sqrt{n}}\left\{G_n \frac{g(x)}{q_{\widetilde{\boldsymbol{\alpha}}}(x)}\right\}(S_{gg}^{-1}S_{\pi g} - \beta^*) \\
&= \frac{1}{\sqrt{n}}\left\{G_n \frac{\pi(x)-\beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\} + \frac{1}{\sqrt{n}}\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}(\widetilde{\alpha}_1 - \alpha_1^*) \\
&+ \frac{1}{\sqrt{n}}\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3}\right\}(\widetilde{\alpha}_1 - \alpha_1^*)^2 + o\left(\frac{1}{\sqrt{n}}\left(\frac{n_0}{n} + \frac{1}{\sqrt{n_0}}\right)^2\right) \\
&+ \frac{1}{\sqrt{n}}\left\{G_n \frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}(\widetilde{\beta} - \beta^*) + o\left(\frac{1}{n}\right), \\
&\text{where } \widetilde{\beta} = \frac{\frac{1}{n}\sum_{i=1}^n \pi(x_i)g(x_i)/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\left\{2 - \frac{\frac{1}{n}\sum_{i=1}^n g(x_i)^2/q_{\widetilde{\boldsymbol{\alpha}}}(x_i)^2}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\right\}.
\end{aligned}$$

Note that $S_{gg}^{-1}S_{\pi g} - \beta^* = \widetilde{\beta} - \beta^* + o(1/\sqrt{n})$ by Taylor expansion. The expansion of $\widehat{\alpha}_1$ in Lemma 4.2 can be plugged into the above equation. After some algebra and note that $1/\sqrt{n} \le n_0/n + 1/n_0$ by the inequality $2ab \le a^2 + b^2$, we obtain the expansion of $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ as follows:

$$\begin{aligned}
&Z + \frac{1}{\sqrt{n}}G_n \frac{\pi(x)-\beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)} \\
&+ \frac{n_0}{n\sqrt{n}}\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}(\gamma_1 - \alpha_1^*) + \frac{1}{\sqrt{n_0}\sqrt{n}}\cdot\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}A_{n_0} \\
&+ \frac{1}{n_0\sqrt{n}}\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3}\cdot A_{n_0}^2 - G_n \frac{(\pi(x)-\beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\cdot B_{n_0}\right\} \\
&+ \frac{1}{n}\left\{G_n \frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}\left\{\sqrt{n}(\widetilde{\beta} - \beta^*)\right\} + o\left(\frac{n_0}{n\sqrt{n}}\right) + o\left(\frac{1}{n_0\sqrt{n}}\right) \\
\equiv &Z + g_1(\widetilde{\boldsymbol{\alpha}}) + g_2(\widetilde{\boldsymbol{\alpha}}) + o\left(\frac{n_0}{n\sqrt{n}}\right) + o\left(\frac{1}{n_0\sqrt{n}}\right), \qquad (4.13)
\end{aligned}$$

$$\begin{aligned}
&\text{where } A_{n_0} = (1,0)\cdot V^{-1}\widehat{U}/2,\ B_{n_0} = (1,0)\cdot V^{-1}\left((\widehat{V} - \sqrt{n_0}V)V^{-1}\widehat{U}/2 + \widehat{W}\right), \\
&g_1(\widetilde{\boldsymbol{\alpha}}) = \frac{1}{\sqrt{n}}G_n \frac{\pi(x)-\beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)} \\
&\text{and } g_2(\widetilde{\boldsymbol{\alpha}}) = \frac{n_0}{n\sqrt{n}}\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}(\gamma_1 - \alpha_1^*) \\
&\quad + \frac{1}{\sqrt{n_0}\sqrt{n}}\cdot\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}A_{n_0} \\
&\quad + \frac{1}{n_0\sqrt{n}}\left\{G_n \frac{(\pi(x)-\beta^* g(x))g(x)^2}{q_{\boldsymbol{\alpha}^*}(x)^3}\cdot A_{n_0}^2 - G_n \frac{(\pi(x)-\beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\cdot B_{n_0}\right\} \\
&\quad + \frac{1}{n}\left\{G_n \frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}\left\{\sqrt{n}(\widetilde{\beta} - \beta^*)\right\}.
\end{aligned}$$

The expansion of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ follows similarly, except the definition of $\widetilde{\beta}$ is changed to

$$\widetilde{\beta} = \frac{\widetilde{Cov}\left[\pi(X)/q_{\widetilde{\boldsymbol{\alpha}}}(X), g(X)/q_{\widetilde{\boldsymbol{\alpha}}}(X)\right]}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\left\{2 - \frac{\widetilde{Var}\left[g(X)/q_{\widetilde{\boldsymbol{\alpha}}}(X)\right]}{\int g(x)^2/q_{\boldsymbol{\alpha}^*}(x)dx}\right\}.$$

□

Proof of Theorem 4.2. The calculation of moments of $\widehat{Z}^*$ involves calculating the moments of (4.13) including

$$E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int\frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)}q_{\widetilde{\boldsymbol{\alpha}}}(x)dx\right)^{k_1}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{h_2(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int\frac{h_2(x)}{q_{\boldsymbol{\alpha}^*}(x)}q_{\boldsymbol{\gamma}}(x)dx\right)^{k_2}(\widetilde{\beta}-\beta^*)^{k_3}\right]$$

for functions $h_1(x)$ and $h_2(x)$, $k_1 = 1, 2$, $k_2 = 0, 1, 2$ and $k_3 = 0, 1, 2$,

From (4.13), note that the calculation of $E\left[\widehat{Z}^* - Z\right]$ involves calculating the cases of $k_1 = 1$. For $(k_1, k_2, k_3) = (1, 1, 0)$ and $(1, 2, 0)$, by plugging in the decomposition

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int\frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)}q_{\widetilde{\boldsymbol{\alpha}}}(x)dx &= \frac{n_0}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int\frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)}q_{\boldsymbol{\gamma}}(x)dx\right)\\ &\quad + \frac{n-n_0}{n}\left(\frac{1}{n-n_0}\sum_{i=1}^{n-n_0}\frac{h_1(x_i)}{q_{\boldsymbol{\alpha}^*}(x_i)} - \int\frac{h_1(x)}{q_{\boldsymbol{\alpha}^*}(x)}q_{\widehat{\boldsymbol{\alpha}}}(x)dx\right),\end{aligned} \tag{4.14}$$

the expectations are of order $O(1/n)$ and $O(1/(n_0 n))$, respectively, by the law of iterated expectations conditioning on $\{x_i\}_{i=1}^{n_0}$. For $(k_1, k_2, k_3) = (1, 0, 1)$, the expectation is of order $O(1/n)$ by the inequality $2ab \le a^2 + b^2$ and the fact $E\left[\sqrt{n}(\widetilde{\beta} - \beta^*)\right]^2 < \infty$. For $(k_1, k_2, k_3) = (1, 0, 0)$, the expectation is 0. Therefore $E\left[\widehat{Z}^* - Z\right] = O(1/n)$.

From (4.13), note that the calculation of $Var\left[\widehat{Z}^* - Z\right]$ involves calculating the cases of $k_1 = 2$. For $(k_1, k_2, k_3) = (2, 0, 0)$, $(2, 1, 0)$ and $(2, 2, 0)$, the expectations are of order $O(1/n)$, $O\left(n_0/n^2\right)$ and $O(1/(n_0 n))$, respectively, by (4.14) and the law of total variance conditioning on $\{x_i\}_{i=1}^{n_0}$. For$(k_1, k_2, k_3) = (2, 0, 1)$ and $(2, 0, 2)$, the expectations are of order $O(1/(n\sqrt{n}))$ and $O(1/n^2)$, respectively, by the inequality $2ab \le a^2 + b^2$ and the fact $E\left[\sqrt{n}(\widetilde{\beta} - \beta^*)\right]^4 < \infty$. The other terms are dominated by

$O\left(n_0/n^2\right) + O\left(1/(n_0 n)\right)$. Therefore by noting that $1/\sqrt{n} \leq n_0/n + 1/n_0$,

$$Var\left[\widehat{Z}^* - Z\right] = \frac{1}{n} Var\left[G_n \frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right] + O\left(\frac{1}{n n_0}\right) + O\left(\frac{n_0}{n^2}\right).$$

Again by (4.14) and the law of total variance conditioning on $\{x_i\}_{i=1}^{n_0}$, some algebra gives that

$$\begin{aligned}
&\frac{1}{n} Var\left[G_n \frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right] \\
=&\frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + \frac{n_0}{n^2}\left\{Var_{\boldsymbol{\gamma}}\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right) - \sigma_Z^2(\boldsymbol{\alpha}^*)\right\} \\
+&\frac{1}{n}(1 - \frac{n_0}{n})\left\{Var_1\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right) - Var_2\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right)\right\} E(\widehat{\alpha}_1 - \alpha_1^*) \quad (4.15) \\
\approx&\frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + O\left(\frac{n_0}{n^2}\right) + O\left(\frac{1}{n n_0}\right).
\end{aligned}$$

Therefore $Var\left[\widehat{Z}^* - Z\right] = \frac{1}{n}\sigma_Z^2(\boldsymbol{\alpha}^*) + O\left(n_0/n^2\right) + O\left(1/(n n_0)\right)$.

□

The proofs of Proposition 4.1 and Theorem 4.2 reveal the sources of the higher orders $O\left(n_0/n^2\right)$ and $O\left(1/(n n_0)\right)$ in $MSE\left[\widehat{Z}^*\right] - n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$. These two orders come from three sources which can be seen by investigating each term in (4.13) and (4.15). One source is due to using pilot samples which leads to terms

$$\begin{aligned}
&\frac{n_0}{n\sqrt{n}}\left\{G_n \frac{(\pi(x) - \beta^* g(x))g(x)}{q_{\boldsymbol{\alpha}^*}(x)^2}\right\}(\gamma_1 - \alpha_1^*) \text{ in (4.13)} \\
\text{and } &\frac{n_0}{n^2}\left\{Var_{\boldsymbol{\gamma}}\left(\frac{\pi(x) - \beta^* g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right) - \sigma_Z^2(\boldsymbol{\alpha}^*)\right\} \text{ in (4.15)},
\end{aligned}$$

and results in the order $O\left(n_0/n^2\right)$. When $\boldsymbol{\gamma} = \boldsymbol{\alpha}^*$, these two terms are equal to 0 and thus they are derived from the difference between $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}^*$. Another one is the variability of random coefficient of control variates which leads to the term

$$\frac{1}{n}\left\{G_n \frac{g(x)}{q_{\boldsymbol{\alpha}^*}(x)}\right\}\left\{\sqrt{n}(\widetilde{\beta} - \beta^*)\right\} \text{ in (4.13)}.$$

This variability results in the order $O(1/(n\sqrt{n}))$ which is also $O\left(n_0/n^2\right) + O\left(1/(n n_0)\right)$ when $n_0 = \sqrt{n}$, because $2/\sqrt{n} \leq n_0/n + 1/n_0$. The other source is the variability of estimated mixture proportion $\widetilde{\boldsymbol{\alpha}}$ which leads to all other terms in (4.13) and (4.15) except the previous 3 terms and $n^{-1}\sigma_Z^2(\boldsymbol{\alpha}^*)$. This variability results in the order $O\left(1/(n n_0)\right)$.

For the asymptotic properties of $\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}})$, the proof differs in two aspects with that of $\widehat{Z}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{Z}_{MLE}(\widetilde{\boldsymbol{\alpha}})$. One is the M-estimator $\widehat{\boldsymbol{\alpha}}$ contains an estimated parameter $\widehat{\mu}$ in the criterion function. The other one is the ratio form of $\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ and $\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}})$.

Lemma 4.5. *$\widehat{\alpha}_1 \xrightarrow{P} \alpha_1^*$ as $n \to \infty$ for $\widehat{\alpha}_1$ defined in (4.7).*

Proof. $\widehat{\alpha}_1$ can be equivalently obtained as a component of the bivariate estimator $(\widehat{\alpha}_1, \widehat{\beta}) = \underset{\alpha_1,\beta\in\Theta}{\operatorname{argmin}}\, n_0^{-1}\sum_{i=1}^{n_0}\rho(x;\alpha_1,\beta,\widehat{\mu})$, where $\rho(x;\alpha_1,\beta,\mu) = \frac{[h(x)\pi(x)-\mu\pi(x)-\beta g(x)]^2}{q_{\boldsymbol{\alpha}}(x)q_{\boldsymbol{\gamma}}(x)}$ and $\Theta = [\delta, 1-\delta]\times\mathbb{R}$. The proof of consistency of $(\widehat{\alpha}_1, \widehat{\beta})$ contains two steps.

First, although the domain of β is unbounded, $\widehat{\beta}$ stays in a compact set almost surely when $n \to \infty$, because

$$|\widehat{\beta}| = \left|\frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{(h(x_i)\pi(x_i)-\widehat{\mu}\pi(x_i))g(x_i)}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}}{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{g(x_i)^2}{q_{\widetilde{\boldsymbol{\alpha}}}(x_i)q_{\boldsymbol{\gamma}}(x_i)}}\right| \le \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(\frac{|h(x_i)\pi(x_i)|}{q_{\boldsymbol{\gamma}}(x_i)} + \frac{|\widehat{\mu}\pi(x_i)|}{q_{\boldsymbol{\gamma}}(x_i)}\right)\frac{2}{\delta}}{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{g(x_i)^2}{(q_1(x_i)+q_2(x_i))q_{\boldsymbol{\gamma}}(x_i)}}$$

and the RHS converges to a constant almost surely since $\widehat{\mu} \to \mu$ almost surely. Then the consistency of $(\widehat{\alpha}_1, \widehat{\beta})$ and the minimizer $(\widehat{\alpha}_1', \widehat{\beta}')$ restricted in some compact set $C \subset \Theta$, i.e. $\underset{\alpha_1,\beta\in C}{\operatorname{argmin}}\, n_0^{-1}\sum_{i=1}^{n_0}\rho(x;\alpha_1,\beta,\widehat{\mu})$, are equivalent since $P((\widehat{\alpha}_1, \widehat{\beta}) \in C) \to 1$.

Second, the consistency of $(\widehat{\alpha}_1', \widehat{\beta}')$ and the minimizer with $\widehat{\mu}$ replaced by μ, i.e. $\underset{\alpha_1,\beta\in C}{\operatorname{argmin}}\, n_0^{-1}\sum_{i=1}^{n_0}\rho(x;\alpha_1,\beta,\mu)$, are equivalent because

$$\begin{aligned}
&\sup_{\alpha_1,\beta\in C}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\rho(x;\alpha_1,\beta,\widehat{\mu}) - \frac{1}{n_0}\sum_{i=1}^{n_0}\rho(x;\alpha_1,\beta,\mu)\right| \\
\le&(\widehat{\mu}^2-\mu^2)\max_{\alpha_1,\beta\in C}\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\pi(x_i)^2}{q_{\alpha_1}(x_i)q_{\boldsymbol{\gamma}}(x_i)} + (\widehat{\mu}-\mu)\max_{\alpha_1,\beta\in C}\frac{2}{n_0}\sum_{i=1}^{n_0}\frac{(h(x_i)\pi(x_i)-\beta g(x_i))\pi(x_i)}{q_{\alpha_1}(x_i)q_{\boldsymbol{\gamma}}(x_i)} \\
\to&0 \text{ almost surely,}
\end{aligned}$$

and the argument similar to Van der Vaart (2000, Theorem 5.7). Then since the consistency of $\underset{\alpha_1,\beta\in C}{\operatorname{argmin}}\, n_0^{-1}\sum_{i=1}^{n_0}\rho(x;\alpha_1,\beta,\mu)$ holds by replacing $\pi(x)$ in Lemma 4.1 by $h(x)\pi(x)-\mu\pi(x)$, the consistency of $(\widehat{\alpha}_1, \widehat{\beta})$ follows. □

Proof of Theorem 4.3. The consistency and asymptotic normality of $\widehat{\mu}_{MLE}(\widetilde{\boldsymbol{\alpha}})$ follow

the extension of proof of Theorem 1 to random vector

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} h(x_i)\pi(x_i)/q_{\widetilde{\boldsymbol{\alpha}}+\widetilde{\zeta}} \\ \pi(x_i)/q_{\widetilde{\boldsymbol{\alpha}}+\widetilde{\zeta}} \end{pmatrix} - \begin{pmatrix} \int h(x)\pi(x)/q_{\boldsymbol{\alpha}^*}(x)dx \\ \int \pi(x)/q_{\boldsymbol{\alpha}^*}(x)dx \end{pmatrix}\right\}$$

and the delta method. The proof for $\widehat{\mu}_{Reg}(\widetilde{\boldsymbol{\alpha}})$ is similar. □

Variance Matrices for $\widehat{\boldsymbol{\alpha}}$

Denote

$$I_{jkl} = \int \frac{(\pi(x) - \boldsymbol{\beta}^{*T}\boldsymbol{g}(x))^j \boldsymbol{g}(x)\boldsymbol{g}(x)^T}{q_{\boldsymbol{\alpha}^*}(x)^k q_{\boldsymbol{\gamma}}(x)^l}dx,$$

$$A = I_{230} - I_{120}I_{010}^{-1}I_{120},\ B = I_{010} - I_{120}I_{230}^{-1}I_{120},$$

$$C = I_{441} - 2I_{331}I_{010}^{-1}I_{120} \text{ and } D = I_{331} - 2I_{221}I_{010}^{-1}I_{120}.$$

When estimating Z, $(\widehat{\alpha}_2, \cdots \widehat{\alpha}_p)$ has the asymptotic variance matrix

$$\frac{1}{\sqrt{n_0}}\left(A^{-1}CA^{-1} - 2I_{230}^{-1}I_{120}B^{-1}DA^{-1}\right).$$

When estimating μ, similar expression can be obtained by replacing $\pi(x)$ in I_{jkl} by $(h(x) - \mu)\pi(x)$.

Chapter 5

Efficient Sequential Monte Carlo with Multiple Proposals and Control Variates

This chapter proposes a novel algorithm to tackle the limitations of SMC proposal design mentioned in sections 2.2 and 3.2.3, using particles generated from a mixture proposal distribution and sample weights constructed by Tan (2004)'s control variate approach. For the problems of infinite variance and multimodal target density, the bounded variance can be guaranteed by including proposals with heavier tails than the target distribution, and multimodality can be handled by including proposals to address the multiple modes separately. For the problem of not considering the target function, it is dealt with by constructing control variates and proposals which incorporate the target function in the construction. Unlike the direct use of control variates, they are included in the resampling step which results in significant improvement and the better performance of mixture proposal over individual proposal. The guidelines for selecting component proposals and control variates are given. The theoretical framework of the algorithm is constructed and the asymptotic results show that the new algorithm is more efficient than the naive implementation of multiple proposals and control variates in SMC, and can be expected to increase the efficiency significantly over the standard SMC methods. Its effectiveness is illustrated through numerical studies on the AR(1) model observed with noise which is a benchmark model since all sequential distributions are analytically available, and the stochastic volatility model with AR(1) dynamics which is widely used in economics and finance.

5.1 Likelihood Based Mixture SMC

5.1.1 The Algorithm

Given p proposal densities $q_1(x_n|x_{1:n-1}), \cdots, q_p(x_n|x_{1:n-1})$ and the auxiliary variable $\eta(x_{0:n})$ for every n. For $t_0 < n$, let $q_k(x_{t_0+1:n}|x_{1:t_0}) = \prod_{t=t_0+1}^{n} q_k(x_t|x_{1:t-1})$ for $k = 1, \cdots, p$, $\boldsymbol{g}(x_{t_0+1:n}|x_{1:t_0}) = (q_1 - q_2, \cdots, q_1 - q_p)(x_{t_0+1:n}|x_{1:t_0})$ and $q_{\boldsymbol{\alpha}}(x_{t_0+1:n}|x_{1:t_0}) = \sum_{k=1}^{p} \alpha_k q_k(x_{t_0+1:n}|x_{1:t_0})$ with mixture proportion vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$. The following algorithm is proposed to utilize multiple proposals and control variates in the SMC framework:

The likelihood-based mixture SMC (LM-SMC):

At time n, assume particles $\{\widetilde{x}_{0:n-1}^{(j)}\}_{j=1}^{N}$ and indicator sets $I_1, \cdots, I_p$ are available where $\cup_{k=1}^{p} I_k = \{1, \cdots, N\}$. Let n_0 be the last time of resampling. For $j \in I_k$,

1. Mutation: Generate $x_n^{(j)}$ from the proposal $q_k(x_n|\widetilde{x}_{0:n-1}^{(j)})$ and let $x_{0:n}^{(j)} = (\widetilde{x}_{0:n-1}^{(j)}, x_n^{(j)})$.

2. Correction: Assign $x_{0:n}^{(j)}$ with weight

$$v_n^{(j)} = \frac{\pi_n(x_{0:n}^{(j)})}{\pi_{n_0}(\widetilde{x}_{0:n_0}^{(j)})\eta(\widetilde{x}_{0:n_0}^{(j)}) \left[q_{\boldsymbol{\alpha}}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)}) + \widehat{\boldsymbol{\zeta}}_n^T \boldsymbol{g}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)})\right]}. \quad (5.1)$$

$$\text{where } \widehat{\boldsymbol{\zeta}}_n = \underset{\boldsymbol{\zeta}}{\text{argmax}} \sum_{j=1}^{N} log \left[q_{\boldsymbol{\alpha}}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)}) + \boldsymbol{\zeta}^T \boldsymbol{g}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)})\right].$$

3. Selection: If the condition for resampling is satisfied, resample $\{x_{0:n}^{(j)}\}_{j=1}^{N}$ according to $\left\{\eta(x_{0:n}^{(j)})v_n^{(j)}\right\}_{j=1}^{N}$ to obtain new particles $\{\widetilde{x}_{0:n}^{(j)}\}_{j=1}^{N}$, and divide $\{1, \cdots, N\}$ with equal probabilities into new indicator sets $I_1, \cdots, I_p$ satisfying $\#I_k = \alpha_k N$; If resampling is not needed, let $\widetilde{x}_{0:n}^{(j)} = x_{0:n}^{(j)}$.

After the correction step, estimate μ_n by

$$\widehat{\mu}_{n,MLE} = \frac{\sum_{j=1}^{N} h(x_n^{(j)})v_n^{(j)}}{\sum_{j=1}^{N} v_n^{(j)}}.$$

Remark 5.1. The novelty of this algorithm is that the control variates are included in both the resampling and estimation. Another way of implementation is that in the

generalized SMC method, estimate μ_n by Owen and Zhou's estimator $\widehat{\mu}_{Reg}$ as following

$$\widehat{\mu}_{n,Reg} = \frac{\sum_{j=1}^{N}\left[h(x_n^{(j)})w_n^{(j)} - \widehat{\boldsymbol{\tau}}_{1n}^T \boldsymbol{g}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)})/q_{\boldsymbol{\alpha}}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)})\right]}{\sum_{j=1}^{N}\left[w_n^{(j)} - \widehat{\boldsymbol{\tau}}_{2n}^T \boldsymbol{g}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)})/q_{\boldsymbol{\alpha}}(x_{n_0+1:n}^{(j)}|\widetilde{x}_{0:n_0}^{(j)})\right]} \tag{5.2}$$

where $\widehat{\boldsymbol{\tau}}_{1n} = \widetilde{Var}\left[\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]^{-1}\widetilde{Cov}\left[\frac{h\pi_n}{\pi_{n_0}cq_{\boldsymbol{\alpha}}}, \frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]$, $\widehat{\boldsymbol{\tau}}_{2n} = \widetilde{Var}\left[\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]^{-1}\widetilde{Cov}\left[\frac{\pi_n}{\pi_{n_0}cq_{\boldsymbol{\alpha}}}, \frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]$.

We call this the regression-based mixture SMC (RM-SMC). It only uses the control variates in the estimation without changing the distribution of particles. Actually the regression approach can also give proper importance weights for each particle, but they are not necessarily positive and hence cannot be directly used in resampling. The likelihood approach give positive importance weights, incorporating the effect of control variates. The asymptotic results in the next section show that this makes the new algorithm outperforms both the generalized SMC method without control variates and the RM-SMC.

Remark 5.2. In the algorithm the particles are mutated within each group and only "mixed up" at the time of resampling. Hence the proposal distribution of $x_{n_0+1:n}$ is the mixture of p distributions and the number of control variates in $\boldsymbol{g}$ is $p-1$. Since the likelihood approach requires the control variates to be compatible with the component proposals, such an implementation can avoid the situation when too many control variates are needed and the optimization for $\widehat{\boldsymbol{\zeta}}_n$ is computationally expensive.

Remark 5.3. For generalized SMC method, the control variates $\boldsymbol{S}(x_{1:n})$ need to satisfy $\int \boldsymbol{S}(x_{1:n})\pi_n^*(x_{0:n})dx_{0:n} = 0$ in order to make the estimator asymptotically unbiased, which makes the construction of $\boldsymbol{S}(x_{1:n})$ not straightforward. The new algorithm does not require extra effort to construct control variates and is easy to implement.

5.1.2 Theoretical Results

Here the central limit theorems (CLT) for $\widehat{\mu}_{n,MLE}$ are presented, similar to that in Chopin (2004). For simplicity, only the scheme of multinomial resampling at every step is discussed. Let $\widetilde{v}_n^{(j)} = \eta(x_{0:n}^{(j)})v_n^{(j)}$ being the weights used in resampling, $\widetilde{\pi}_n(x_{0:n}) =$

$\eta(x_{0:n})\pi_n^*(x_{0:n})$ being the unnormalized new target density after resampling with the auxiliary variable, $\widetilde{\pi}_n^*(x_{0:n}) = e_n^{-1}\widetilde{\pi}_n(x_{0:n})$ where e_n is the normalizing constant, and $\widetilde{\mu}_n = \int h(x_{1:n})\widetilde{\pi}_n^*(x_{0:n})dx_{0:n}$. We assume the following conditions:

(C1') $\int |h(x_{1:n})|\pi_n^*(x_{0:n})dx_{0:n}$, $\int |h(x_{1:n})|\widetilde{\pi}_n^*(x_{0:n})dx_{0:n}$ and $\int |h(x_{1:n})|\widetilde{\pi}_{n-1}^*(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})dx_{0:n} < \infty$;

(C2') $E_{\pi_n^*}|h(x_{1:n})|^2$ and $E_{\widetilde{\pi}_n^*}|h(x_{1:n})|^2 < \infty$;

(C3') Let Φ_0 to be the set of square integrable functions with respect to $\pi_0(x_0)$ and $\Phi_n = \left\{ h : \Theta_n \to \mathbb{R} \,\middle|\, E_{\widetilde{\pi}_{n-1}^* q_{\boldsymbol{\alpha}}}\left[\frac{\pi_n^*}{\widetilde{\pi}_{n-1}^* q_{\boldsymbol{\alpha}}}h\right]^2 < \infty, E_{\widetilde{\pi}_{n-1}^* q_{\boldsymbol{\alpha}}}\left[\frac{\widetilde{\pi}_n^*}{\widetilde{\pi}_{n-1}^* q_{\boldsymbol{\alpha}}}h\right]^2 < \infty, \right.$ $\left. E_{q_{\boldsymbol{\alpha}}}\left[\frac{\pi_n^*}{\widetilde{\pi}_{n-1}^* q_{\boldsymbol{\alpha}}}h|x_{0:n-1}\right] \in \Phi_{n-1} \text{ and } E_{q_{\boldsymbol{\alpha}}}\left[\frac{\widetilde{\pi}_n^*}{\widetilde{\pi}_{n-1}^* q_{\boldsymbol{\alpha}}}h|x_{0:n-1}\right] \in \Phi_{n-1} \right\}$. Then $h \in \Phi_n$;

(C4') The unit function $I_n : \Theta_n \to 1$ belongs to Φ_n.

(C5) $Var_{\widetilde{\pi}_n^* q_{\boldsymbol{\alpha}}}\left(\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right) < \infty$ and is positive definite;

(C6) $\int \boldsymbol{g}(x_n|x_{0:n-1})dx_n \equiv 0$.

Let $\sigma_{3,0}^2(h) = Var_{\pi_0}[h]$ and recursively let

$$\begin{aligned}
\sigma_{1,n}^2(h) &= \sigma_{3,n-1}^2(E_{q_\alpha}[h|x_{0:n-1}]) + E_{\widetilde{\pi}_{n-1}^*}(Var_{q_{\boldsymbol{\alpha}}}[h|x_{0:n-1}]), \\
\sigma_{2,n}^2(h) &= \sigma_{1,n}^2\left(\frac{\pi_n^*(x_{0:n})(h-\mu_n)}{\widetilde{\pi}_{n-1}^*(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})} - \boldsymbol{\beta}_n^T\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right), \\
\sigma_{3,n}^2(h) &= \sigma_{1,n}^2\left(\frac{\widetilde{\pi}_n^*(x_{0:n})(h-\widetilde{\mu}_n)}{\widetilde{\pi}_{n-1}^*(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})} - \widetilde{\boldsymbol{\beta}}_n^T\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right) + Var_{\widetilde{\pi}_n^*}(h),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\beta}_n &= Var\left[\frac{\boldsymbol{g}(x_n|x_{0:n-1})}{q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})}\right]^{-1} Cov\left[\frac{\pi_n^*(x_{0:n})(h(x_{1:n})-\mu_n)}{\widetilde{\pi}_{n-1}^*(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})}, \frac{\boldsymbol{g}(x_n|x_{0:n-1})}{q_{\boldsymbol{\alpha}}(x_n|x_{10:n-1})}\right] \\
\text{and } \widetilde{\boldsymbol{\beta}}_n &= Var\left[\frac{\boldsymbol{g}(x_n|x_{0:n-1})}{q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})}\right]^{-1} Cov\left[\frac{\widetilde{\pi}_n^*(x_{0:n})(h(x_{1:n})-\widetilde{\mu}_n)}{\widetilde{\pi}_{n-1}^*(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})}, \frac{\boldsymbol{g}(x_n|x_{0:n-1})}{q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})}\right].
\end{aligned}$$

Theorem 5.1. *Suppose conditions* $(C1') - (C4')$, $(C5)$ *and* $(C6)$ *are satisfied. Then for any* n, $\sigma_{2,n}^2(h)$ *and* $\sigma_{3,n}^2(h)$ *are finite and*

$$\sqrt{N}\left(\widehat{\mu}_{n,MLE} - \mu_n\right) \xrightarrow{\mathfrak{L}} N(0, \sigma_{2,n}^2(h)), \tag{5.3}$$

$$\sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N} h(\widetilde{x}_{1:n}^{(i)}) - \widetilde{\mu}_n\right] \xrightarrow{\mathfrak{L}} N(0, \sigma_{3,n}^2(h)). \tag{5.4}$$

Specifically,

$$\sigma^2_{2,n}(h) = \sum_{t=1}^{n} \int \frac{\left[\pi^*_n(x_{0:t})(\mu_t(x_{0:t}) - \mu_n) - \boldsymbol{\beta}^T_{tn}\widetilde{\pi}^*_{t-1}(x_{0:t-1})\boldsymbol{g}(x_t|x_{0:t-1})\right]^2}{\widetilde{\pi}^*_{t-1}(x_{0:t-1})q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})} dx_{0:t}, \quad (5.5)$$

where

$$\boldsymbol{\beta}_{tn} = Var\left[\frac{\boldsymbol{g}(x_t|x_{0:t-1})}{q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})}\right]^{-1} Cov\left[\frac{\pi^*_n(x_{0:t})(\mu_t(x_{0:t}) - \mu_n)}{\widetilde{\pi}^*_{t-1}(x_{0:t-1})q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})}, \frac{\boldsymbol{g}(x_t|x_{0:t-1})}{q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})}\right],$$

and $\widetilde{\pi}^*_0(x_0) = \pi_0(x_0)$.

The first four conditions have been used for the convergence of auxiliary particle filter and analogous to the conditions for the basic SMC method. (C1') and (C2') depend on the model and target function, and in practice are usually satisfied. One sufficient condition for (C3') and (C4') to hold is

$$\frac{p(x_n|x_{n-1})p(y_n|x_n)}{\eta(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})} \text{ and } \frac{\eta(x_{0:n})p(x_n|x_{n-1})p(y_n|x_n)}{\eta(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})} \text{ are bounded from above,} \quad (5.6)$$

which can be satisfied by including heavy tail component proposals. Unlike (C3), only second moments are required in (C3'). See Cappé et al. (2005, Theorem 9.3.7) for this weaker condition. Condition (C5) ensures that the optimization for MLE weights $\widehat{\boldsymbol{\zeta}}$ gives stable results. Condition (C6) is necessary for $\sigma^2_{2,n}(h)$ to have the clean expression in Theorem 2. It is automatically satisfied in LM-SMC and is stated here to emphasize the requirement for control variates.

The analytical form of $\sigma^2_{2,n}(h)$ clearly demonstrates how the control variates take effect in two aspects. First, every term in the asymptotic variance contains the control variates, which is resulted from including control variates in the resampling step. Second, the coefficient vector for control variates is optimal for every n despite the target density for each term changes as n increases. This is due to the fact that in the likelihood approach, the estimated coefficient vector $\widehat{\boldsymbol{\zeta}}_n$ does not depend on the target density, and hence the same sample and coefficients can be adapted for different target density automatically. Therefore, LM-SMC outperforms the generalized SMC method without using control variates. Since the RM-SMC only includes the control variates in the estimator, its asymptotic variance only have them in the last term, which is shown

in the technical proof. Therefore, LM-SMC also outperforms RM-SMC.

5.1.3 The Selection of Component Proposals and Auxiliary Variable

The choices of component proposals $q_1, \cdots, q_p$ are critical to the performance of LM-SMC. As mentioned in Section 3.2.3, a reasonable strategy is to minimize the last term of the asymptotic variance which is

$$\int \frac{\left[\pi_n^*(x_{0:n})(h(x_{1:n}) - \mu_n) - \boldsymbol{\beta}_{nn}^T \boldsymbol{g}(x_n|x_{0:n-1})\widetilde{\pi}_{n-1}^*(x_{0:n-1})\right]^2}{\widetilde{\pi}_{n-1}^*(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})} dx_{0:n}. \tag{5.7}$$

By the theory of the likelihood approach (Tan, 2004), (5.7) is equal to 0 when $\pi_n^*(h - \mu_n)$ is a linear combination of $\widetilde{\pi}_{n-1}^* q_1, \cdots, \widetilde{\pi}_{n-1}^* q_p$. This property suggests to choose $q_1, \cdots, q_p$ and $\eta(x_{0:n})$ such that $\pi_n^*(h - \mu_n)$ is close to some linear combination of $\widetilde{\pi}_{n-1}^* q_1, \cdots, \widetilde{\pi}_{n-1}^* q_p$. Since $\pi_n^*(x_{0:n})(h(x_{1:n}) - \mu_n) \propto \pi_{n-1}^*(x_{0:n-1})p(x_n|x_{n-1})p(y_n|x_n)(h(x_{1:n}) - \mu_n)$, the construction of component proposals can be done by approximating and decomposing $p(x_n|x_{n-1})p(y_n|x_n)(h - \mu_n)$. Specifically, one can approximate $p(x_n|x_{n-1})p(y_n|x_n)$ by $\eta(x_{0:n-1})q(x_n|x_{0:n-1})$ where $\eta(x_{0:n-1})$ is a positive function and $q(x_n|x_{0:n-1})$ is a probability density, and find a decomposition

$$q(x_n|x_{0:n-1})(h(x_{1:n}) - \mu_n) = \sum_{k=1}^{p} c_k r_k(x_{1:n-1}) q_k(x_n|x_{0:n-1}) \tag{5.8}$$

with real functions $r_k(x_{0:n-1})$, densities $q_k(x_n|x_{0:n-1})$ and constants c_k. Then $q_1, \cdots, q_p$ and $\eta(x_{0:n-1})$ can be selected as the component proposals and auxiliary variable.

Meanwhile, to ensure the sample weights have bounded variance, one can include a heavy tail component proposal and add a positive constant into $\eta(x_{0:n-1})$ to have $\eta(x_{0:n-1})$ bounded away from 0, so that the sufficient condition (5.6) is satisfied. A simple choice of the heavy tailed proposal is the prior density $p(x_n|x_{n-1})$.

The above strategy can be illustrated using the state space model

$$\begin{aligned} x_n &= s_1(x_{n-1}) + \varepsilon_n \\ y_n &= s_2(x_n) + e_n, \end{aligned} \tag{5.9}$$

with normal ε_t and e_t and nonlinear function $s_2(x)$. Suppose the target function $h(x_{1:n})$

is x_n.

Let $log(p_2(y_n|x_n))$ be the second order Taylor expansion of $log(p(y_n|x_n))$ around the mode of $p(y_n|x_n)$, $q_2(x_n|y_n, x_{n-1})$ be the normalized $p(x_n|x_{n-1})p_2(y_n|x_n)$ and $r(x_{n-1})$ be the corresponding normalizing constant. By approximating the center of normalized $p(x_n|x_{n-1})p(y_n|x_n)$ by $q_2(x_n|y_n, x_{n-1})$, controlling its tail by $p(x_n|x_{n-1})$ and adding a positive constant c to $r(x_{n-1})$, $p(x_n|x_{n-1})p(y_n|x_n)$ can be approximated by

$$\eta(x_{0:n-1})q(x_n|x_{0:n-1}) \equiv [r(x_{n-1}) + c]\,[\gamma_1 q_2(x_n|y_n, x_{n-1}) + \gamma_2 p(x_n|x_{n-1})], \tag{5.10}$$

where the values of γ_1 and γ_2 can be arbitrary. The value c should not be too large or too small compared to $r(x_{n-1})$. The expectation $E_{\pi^*_{n-2}q_{\boldsymbol{\alpha}}}[r(x_{n-1})]$ is a reasonable choice and can be estimated by sample average.

Then the component proposal can be constructed as follows. Note that $q_2(x_n|y_n, x_{n-1})$ is a normal density with mean $\theta(x_{n-1})$. Decompose $q(x_n|x_{0:n-1})(x_n - \mu_n)$ by

$$\begin{aligned}
&[\gamma_1 q_2(x_n|y_n, x_{n-1}) + \gamma_2 p(x_n|x_{n-1})]\,(x_n - \mu_n) \\
&= \gamma_1\,[x_n - \theta(x_{n-1})]\,q_2(x_n|y_n, x_{n-1}) + \gamma_1\,[\theta(x_{n-1}) - \mu_n]\,q_2(x_n|y_n, x_{n-1}) \\
&\quad + \gamma_2\,[x_n - s_1(x_{n-1})]\,p(x_n|x_{n-1}) + \gamma_2\,[s_1(x_{n-1}) - \mu_n]\,p(x_n|x_{n-1})
\end{aligned} \tag{5.11}$$

and use the following component proposal distributions: the normalized $[x_n - \theta(x_{n-1})]^+ q_2(x_n|y_n, x_{n-1})$, normalized $[x_n - \theta(x_{n-1})]^- q_2(x_n|y_n, x_{n-1})$, $q_2(x_n|y_n, x_{n-1})$, normalized $[x_n - s_1(x_{n-1})]^+ p(x_n|x_{n-1})$, normalized $[x_n - s_1(x_{n-1})]^- p(x_n|x_{n-1})$ and $p(x_n|x_{n-1})$. All of them can be sampled directly (Weibull distribution or Normal distribution). When the mixture proportion for $p(x_n|x_{n-1})$ is non-zero ($\alpha_6 > 0$), then (5.6) is satisfied by the fact that $\eta(x_{0:n-1})$ is bounded and $\eta(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1}) > \alpha_6 c p(x_n|x_{n-1})$.

5.1.4 The Selection of Mixture Proportions

In (5.8), some proposals may be computationally expensive to sample from. The problem can be circumvented by setting the sampling proportions of these proposals to 0

and only sampling particles from a subset of $q_1, \cdots, q_p$, while keeping the control variates $\boldsymbol{g}(x_n|x_{0:n-1})$ the same. This is a reasonable strategy since, when $\pi_n^*(h-\mu_n)$ is a linear combination of $\widetilde{\pi}_{n-1}^* q_1, \cdots, \widetilde{\pi}_{n-1}^* q_p$, the variance in (5.7) equals to 0 even if some α_i are 0. Such a strategy allows more flexibility in decomposition (5.8) to include terms which are easy to be normalized but difficult to sample from. The values of nonzero proportions can follow some heuristic rules derived from experience or interpretation of proposals. Equal proportions are often a good starting point.

The strategy raised the question of whether sampling from a subset of $q_1, \cdots, q_p$ will decrease the estimation efficiency and offset the saving of computational resource. It has been noted that when the sample coverage is not of a major concern, the improvement by the likelihood approach mainly comes from the use of control variates. Therefore in practice, one can construct several easy-to-sample proposals to cover the high likelihood area of the target density with some additional more sophisticated proposal densities as covariates to achieve more accurate approximation. We illustrate this observation through the following example in importance sampling, though similar features can be seen in SMC as well.

Example. Suppose a random vector $\boldsymbol{X} = (X_1, \cdots, X_{10})$ follows

$$\pi(\boldsymbol{x}) = .8 \prod_{p=1}^{10} \phi(x_p) + .2 \prod_{p=1}^{10} \psi_4(x_p),$$

where $\phi(x)$ is the standard normal density and $\psi_t(x)$ is the student t density with degrees of freedom t. This distribution is used in Tan (2004). The target of interest is the expectation $\mu = E\left[f(\boldsymbol{X})\right]$ under $\pi(\boldsymbol{x})$, where $f(\boldsymbol{x}) = \sum_{p=1}^{10} x_p/10$. Here we compare four estimators to illustrate the effects of setting some mixture proportions to 0.

The first two estimators are based on the proposal choices in Tan (2004). Let $q_1(\boldsymbol{x}) = \prod_{p=1}^{10} \phi(x_p)$, $q_2(\boldsymbol{x}) = \prod_{p=1}^{10} \psi_2(x_p)$, $g_1(\mathbf{x}) = q_1(\boldsymbol{x}) - q_2(\boldsymbol{x})$ and $q_{.5,2}(\boldsymbol{x}) = .5q_1(\boldsymbol{x}) + .5q_2(\boldsymbol{x})$ where $q_{\boldsymbol{\alpha},k}(\boldsymbol{x})$ denotes the mixture of k proposals with proportions $\boldsymbol{\alpha}$. Among the component proposals, q_1 approximates the center of π and q_2 controls the tail of π. Suppose $\{\boldsymbol{x}_i\}_{i=1}^n$ are generated from $q_{.5,2}(\boldsymbol{x})$ in stratification. Then IS and Tan's

likelihood approach give the estimator

$$\widehat{\mu}_{P_2} = \frac{\sum_{i=1}^n f(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i)/q_{.5,2}(\boldsymbol{x}_i)}{\sum_{i=1}^n \pi(\boldsymbol{x}_i)/q_{.5,2}(\boldsymbol{x}_i)},$$
$$\widehat{\mu}_{P_2C_1} = \frac{\sum_{i=1}^n f(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i)/\left[q_{.5,2}(\boldsymbol{x}_i) + \widehat{\zeta}_{P_2C_1} g_1(\boldsymbol{x}_i)\right]}{\sum_{i=1}^n \pi(\boldsymbol{x}_i)/\left[q_{.5,2}(\boldsymbol{x}_i) + \widehat{\zeta}_{P_2C_1} g_1(\boldsymbol{x}_i)\right]},$$

where $\widehat{\zeta}_{P_2C_1} = \underset{\zeta}{\text{argmax}} \sum_{i=1}^n log\left(q_{.5,2}(\boldsymbol{x}_i) + \zeta g_1(\boldsymbol{x}_i)\right)$ and $\widehat{\mu}_{P_kC_j}$ denotes the estimator with k sampling proposals and j control variates.

The third and fourth estimators are based on more sophisticated component proposals following the discussions in Li et al. (2012) which suggested to decompose an approximation of $(f(\boldsymbol{x}) - \mu)\pi(\boldsymbol{x})$, similar to the discussions in Section 5.1.3. By approximating $\pi(\boldsymbol{x})$ with the mixture of $q_1(\boldsymbol{x})$ and $q_2(\boldsymbol{x})$ as in the first two estimators,

$$\begin{aligned}\left(\frac{1}{10}\sum_{p=1}^{10} x_p - \mu\right)\pi(\boldsymbol{x}) &\approx \left(\frac{1}{10}\sum_{p=1}^{10} x_p - \mu\right)\left[\gamma_1 q_1(\boldsymbol{x}) + \gamma_2 q_2(\boldsymbol{x})\right] \\ &= \sum_{p=1}^{10} \tau_{1p} x_p^+ \phi(x_p)\prod_{i\neq p}\phi(x_i) - \sum_{p=1}^{10}\tau_{2p}x_p^-\phi(x_p)\prod_{i\neq p}\phi(x_i) \\ &+ \sum_{p=1}^{10}\tau_{3p}x_p^+\psi_2(x_p)\prod_{i\neq p}\psi_2(x_i) - \sum_{p=1}^{10}\tau_{4p}x_p^-\psi_2(x_p)\prod_{i\neq p}\psi_2(x_i) \\ &+ \tau_5 q_1(\boldsymbol{x}) + \tau_6 q_2(\boldsymbol{x}),\end{aligned}$$

with constants $\gamma_1,\ \gamma_2,\ \tau_1,\ldots,\tau_6$. Therefore we choose the following component proposals: $q_{1j+}(\boldsymbol{x}) \propto x_j^+\phi(x_j)\prod_{i\neq j}\phi(x_i)$, $q_{1j-}(\boldsymbol{x}) \propto x_j^-\phi(x_j)\prod_{i\neq j}\phi(x_i)$, $q_{2j+}(\boldsymbol{x}) \propto x_j^+\psi_2(x_j)\prod_{i\neq j}\psi_2(x_i)$, $q_{2j-}(\boldsymbol{x}) \propto x_j^-\psi_2(x_j)\prod_{i\neq j}\psi_2(x_i)$, for $j = 1,\ldots,10$ and $q_1(\boldsymbol{x})$ and $q_2(\boldsymbol{x})$. There are total 42 component proposals and all can be sampled relatively easily. Let $q_{\boldsymbol{\alpha},42}(\boldsymbol{x})$ be the mixture of these component proposals with equal proportions and $\boldsymbol{g}(\boldsymbol{x})$ be the corresponding vector of control variates. Suppose $\{\boldsymbol{x}_i\}$ are generated from $q_{\boldsymbol{\alpha},42}(\boldsymbol{x})$ in stratification. The likelihood approach gives the third estimator

$$\widehat{\mu}_{P_{42}C_{41}} = \frac{\sum_{i=1}^n f(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i)/\left[q_{\boldsymbol{\alpha},42}(\boldsymbol{x}_i) + \widehat{\boldsymbol{\zeta}}^T_{P_{42}C_{41}}\boldsymbol{g}(\boldsymbol{x}_i)\right]}{\sum_{i=1}^n \pi(\boldsymbol{x}_i)/\left[q_{\boldsymbol{\alpha},42}(\boldsymbol{x}_i) + \widehat{\boldsymbol{\zeta}}^T_{P_{42}C_{41}}\boldsymbol{g}(\boldsymbol{x}_i)\right]},$$

	$\widehat{\mu}_{P_2}$	$\widehat{\mu}_{P_2C_1}$	$\widehat{\mu}_{P_{42}C_{41}}$	$\widehat{\mu}_{P_2C_{41}}$
MSE	$1.4E-1$	$1.4E-1$	$4.1E-3$	$5.0E-3$

Table 5.1: Comparison of the four estimators in example of Section 5.1.4. Simulation is replicated for 1000 times independently and each replicate uses 4000 draws. The mean square errors are reported.

where $\widehat{\boldsymbol{\zeta}}_{P_{42}C_{41}} = \underset{\boldsymbol{\zeta}}{\text{argmax}} \sum_{i=1}^{n} log\left(q_{\boldsymbol{\alpha},42}(\boldsymbol{x}_i) + \boldsymbol{\zeta}^T \boldsymbol{g}(\boldsymbol{x}_i)\right)$.

The fourth estimator is constructed by setting the mixture proportions of all component proposals except $q_1(\boldsymbol{x})$ and $q_2(\boldsymbol{x})$ to 0 in $\widehat{\mu}_{P_{42}C_{41}}$, which means the sample is only generated from $q_{.5,2}(\boldsymbol{x})$ but the control variates $\boldsymbol{g}(\boldsymbol{x})$ is still the same. This gives the fourth estimator

$$\widehat{\mu}_{P_2C_{41}} = \frac{\sum_{i=1}^{n} f(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i) / \left[q_{.5,2}(\boldsymbol{x}_i) + \widehat{\boldsymbol{\zeta}}^T_{P_2C_{41}} \boldsymbol{g}(\boldsymbol{x}_i)\right]}{\sum_{i=1}^{n} \pi(\boldsymbol{x}_i) / \left[q_{.5,2}(\boldsymbol{x}_i) + \widehat{\boldsymbol{\zeta}}^T_{P_2C_{41}} \boldsymbol{g}(\boldsymbol{x}_i)\right]},$$

where $\widehat{\boldsymbol{\zeta}}_{P_2C_{41}} = \underset{\boldsymbol{\zeta}}{\text{argmax}} \sum_{i=1}^{n} log\left(q_{.5,2}(\boldsymbol{x}_i) + \boldsymbol{\zeta}^T \boldsymbol{g}(\boldsymbol{x}_i)\right)$. Note that $\widehat{\mu}_{P_2C_{41}}$ has a bounded variance since the heavy tail proposal $q_2(\boldsymbol{x})$ is included.

The mean square errors (MSE) of these four estimators are reported in Table 1. it can be seen that $\widehat{\mu}_{P_2}$ and $\widehat{\mu}_{P_2C_1}$ have similar MSE, indicating no improvement by a control variate $g_1(\boldsymbol{x})$ without considering the target function $f(\boldsymbol{x})$. When control variates constructed using the information of $f(\boldsymbol{x})$ are included, the resulted estimator $\widehat{\mu}_{P_2C_{41}}$ improves the MSE of $\widehat{\mu}_{P_2}$ by more than one order of magnitude. If the sampling proposals also use the information of $f(\boldsymbol{x})$, the resulting estimator $\widehat{\mu}_{P_{42}C_{41}}$ improves the MSE of $\widehat{\mu}_{P_2C_{41}}$ by about 20%. It shows that the main contribution of improvement comes from the control variates instead of proposal distributions.

5.2 Numerical Studies

Here we present several examples to illustrate the performance of the new algorithm. In these examples, the target function is x_n, i.e. the posterior mean of state is of interest. We compare five methods: The bootstrap filter (BF), auxiliary particle filter (APF), the generalized SMC method without using control variates (GSMC), regression-based mixture SMC (RM-SMC) of (5.2), and likelihood-based mixture SMC (LM-SMC). The

generalized SMC used here does not include control variates due to the difficulties discussed in remark 5.3 of Section 5.1.1. In all examples, systematic resampling (Carpenter et al., 1999; Douc and Cappé, 2005) is used at every step and the simulation is replicated 200 times independently, and each step uses 2000 particles. The trust region optimization algorithm (Nocedal and Wright, 1999) is used for calculating MLE weights in all examples. The average of mean square error over 100 steps, i.e.

$$\overline{MSE} = \frac{1}{100}\sum_{t=1}^{100} MSE_t, \text{ where } MSE_t = \frac{1}{200}\sum_{i=1}^{200}(\widehat{\mu}_{ti} - \mu_t)^2,$$

where $\widehat{\mu}_{ti}$ is the estimator at the t_{th} step of i_{th} replication and μ_t is the theoretical posterior mean of x_t, and the comparison between LM-SMC and the i_{th} method with consideration of computing time, i.e. the ratio

$$R_i = \frac{\overline{MSE}_{LM-SMC} T_{LM-SMC}}{\overline{MSE}_i T_i},$$

where T is the computing time, for $i = 1, \cdots, 4$ are reported.

5.2.1 AR(1) Observed with Noise

Consider the following process

$$\begin{aligned} x_n &= \phi x_{n-1} + \varepsilon_n,\ \varepsilon_n \sim N(0, \sigma^2) \\ y_n &= x_n + \eta_n,\ \eta_n \sim N(0, 1). \end{aligned}$$

It is often used as a benchmark model for comparing SMC methods since all sequential distributions are analytically available through Kalman filter. The density $p(x_n|x_{n-1}, y_n)$ is normal and $\eta(x_{n-1}) = \pi_n(x_{0:n-1})/\pi_{n-1}(x_{0:n-1})$ can be evaluated. Therefore the "perfect adaption" of the auxiliary particle filter (Pitt and Shephard, 1999) can be achieved.

Construction of Component Proposals

For the mixture methods, the component proposals can be obtained by the following decomposition. Denote the mean of $p(x_n|x_{n-1}, y_n)$ by $\theta_n(x_{n-1})$. We have

$$\begin{aligned} & p(x_n|x_{n-1})p(y_n|x_n)(x_n - \mu_n) \\ &= \eta(x_{n-1}) \left[(x_n - \theta_n(x_{n-1}))^+ p(x_n|x_{n-1}, y_n) + (x_n - \theta_n(x_{n-1}))^- p(x_n|x_{n-1}, y_n)\right. \\ &\quad \left. +(\theta_n(x_{n-1}) - \mu_n)p(x_n|x_{n-1}, y_n)\right]. \end{aligned}$$

Then we can use $\eta(x_{n-1})$ as the auxiliary variable, and the following component proposals: $p(x_n|x_{n-1}, y_n)$, normalized $(x_n - \theta_n(x_{n-1}))^+ p(x_n|x_{n-1}, y_n)$ and normalized $(x_n - \theta_n(x_{n-1}))^- p(x_n|x_{n-1}, y_n)$. All proposals can be sampled directly and equal mixture proportions are assigned to all component proposals.

Results

The five methods are compared under different signal to noise ratio (SNR) settings. The values of parameter ϕ are set to be .7, .8, .9 and σ are determined so that the SNR of the process is controlled at 10, 1 and .5. For each setting, a series with length 100 is generated as observations. Results are listed in Table 2 and the trajectories of MSE for all five methods are given in Figure 1. From Table 2, several observations can be made.

In all cases, LM-SMC have the smallest $\overline{MSE}$ and the improvement over the others decreases as the value of SNR decreases and the value of ϕ increases. For the SNR=10 case, considering the computing time, LM-SMC improves BF, APF and GSMC by over 85%, except for GSMC when $\phi = .9$, and RM-SMC by over 18%. For the SNR=1 case, considering the computing time, LM-SMC also have improvement over the others for $\phi = .7$ and $\phi = .8$ but the improvement is much smaller compared to SNR=10 case. For the SNR=.5 case, the MSE decrease of LM-SMC is not significant with consideration of computing time and therefore has similar or worse performance compared with the others.

The above trends are due to an interesting phenomena. As SNR increases or ϕ

SNR=10	$\phi = .7, \sigma = 2.3$	$\phi = .8, \sigma = 1.9$	$\phi = .9, \sigma = 1.4$	Time sec
BF	$2.2E-3(.006)$	$1.0E-3(.019)$	$9.1E-4(.050)$	44
APF	$4.4E-4(.031)$	$4.1E-4(.045)$	$3.8E-4(.11)$	46
GSMC	$1.1E-4(.099)$	$1.1E-4(.14)$	$1.1E-4(.31)$	57
RM-SMC	$9.7E-6(.82)$	$14E-6(.76)$	$3.2E-5(.78)$	78
LM-SMC	$6.4E-6(1.0)$	$8.8E-6(1.0)$	$2.0E-5(1.0)$	97
SNR=1	$\phi = .7, \sigma = .7$	$\phi = .8, \sigma = .6$	$\phi = .9, \sigma = .44$	Time sec
BF	$4.7E-4(.34)$	$3.6E-4(.53)$	$3.7E-4(.78)$	44
APF	$2.9E-4(.54)$	$2.5E-4(.74)$	$2.5E-4(1.1)$	46
GSMC	$1.4E-4(.90)$	$1.4E-4(1.0)$	$1.7E-4(1.3)$	57
RM-SMC	$9.8E-5(.93)$	$1.2E-4(.95)$	$1.5E-4(1.1)$	78
LM-SMC	$7.3E-5(1.0)$	$8.8E-5(1.0)$	$1.3E-4(1.0)$	97
SNR=.5	$\phi = .7, \sigma = .5$	$\phi = .8, \sigma = .42$	$\phi = .9, \sigma = .31$	Time sec
BF	$3.5E-4(.70)$	$3.1E-4(.90)$	$2.0E-4(1.4)$	44
APF	$2.3E-4(1.0)$	$2.2E-4(1.2)$	$1.7E-4(1.6)$	46
GSMC	$1.5E-4(1.2)$	$1.6E-4(1.3)$	$1.6E-4(1.4)$	57
RM-SMC	$1.3E-4(1.0)$	$1.5E-4(1.1)$	$1.5E-4(1.1)$	78
LM-SMC	$1.1E-4(1.0)$	$1.3E-4(1.0)$	$1.3E-4(1.0)$	97

Table 5.2: Comparison of five methods in Example 5.2.1. $\overline{MSE}$ is reported and the ratio of $\overline{MSE}$ multiplied with computing time between LM-SMC and the corresponding method is reported in the parenthesis.

decreases, the average MSE of RM-SMC and LM-SMC decreases, while the average MSE of BF and APF increases. Figure 1 also indicates that such phenomena not only happens on the average values, but also on the whole trajectories. A possible reason is that since the control variates of RM-SMC and LM-SMC fully explore the information contained in x_n and SNR is a direct measure of information of x_n, under higher SNR setting, the information of x_n is more significant and therefore gives the control variates approach more advantage.

Finally, by comparing RM-SMC and LM-SMC, it can be seen that the difference of their MSE decreases as SNR increases. It means in the asymptotic variance of LM-SMC in Theorem 2, the last term contains more and more proportion, and then the variance reduction brought by the historical terms becomes less and less significant. Therefore the improvement of LM-SMC over RM-SMC decreases as SNR increases.

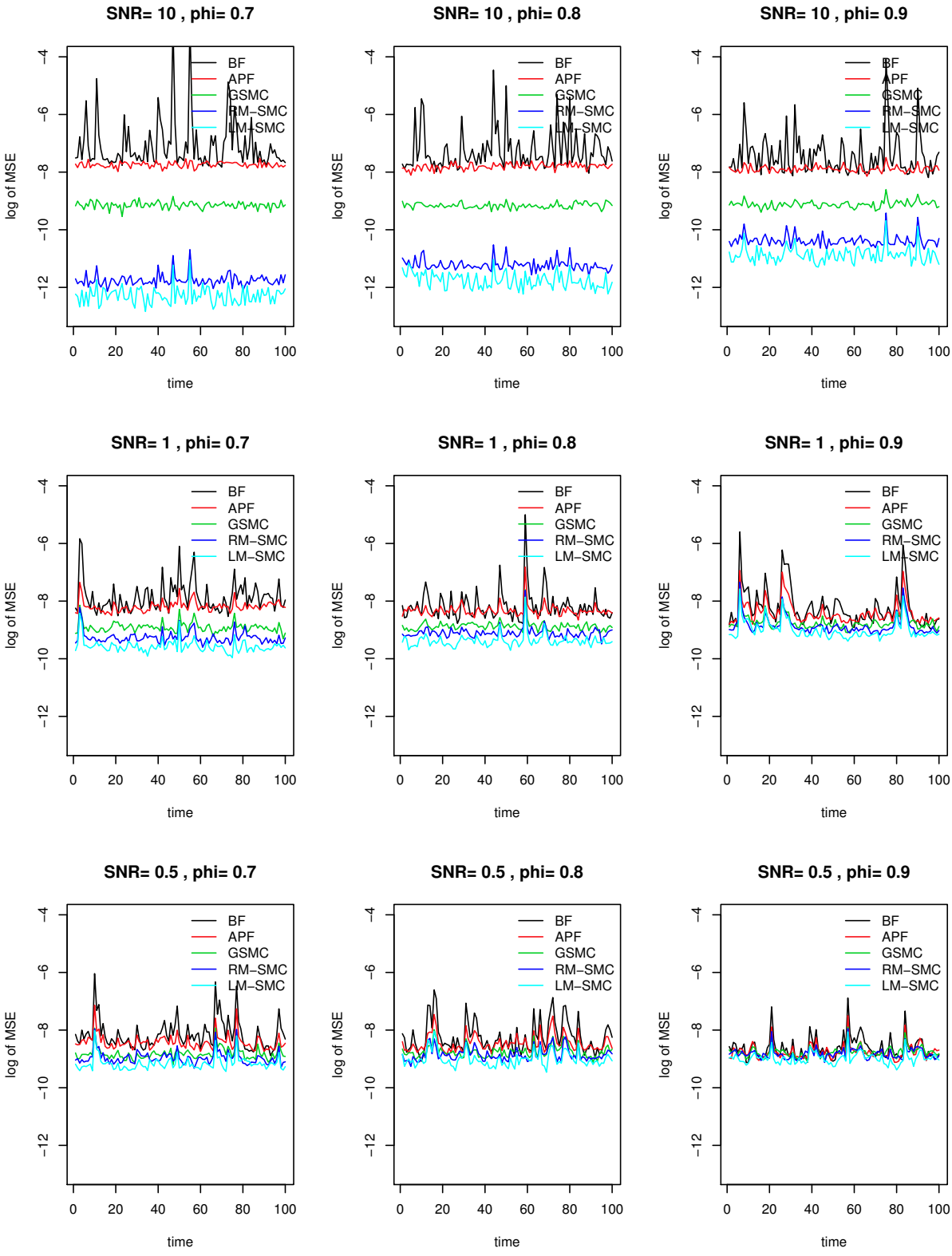

Figure 5.1: Trajectories of logarithm of MSE for the five estimators in all cases of Example 5.2.1.

5.2.2 Stochastic Volatility with AR(1) Dynamics

Consider the stochastic volatility model in Sandmann and Koopman (1998)

$$\begin{cases} x_n &= \phi x_{n-1} + \eta_n,\ \eta_n \sim N(0, \sigma_\eta^2) \\ y_n &= \overline{\sigma} e^{\frac{x_n}{2}} \xi_n,\ \xi_n \sim N(0,1), \end{cases}$$

where η_t and ξ_t are independent, y_n is the demeaned return of a portfolio, and $\overline{\sigma}$ is the average volatility level. Due to the nonlinear structure of the observation equation, the analytical form of $p(x_n|x_{n-1}, y_n)$ is unavailable.

For the APF method, Kim et al. (1998) and Pitt and Shephard (1999) suggested to use the normal density $q_1(x_n|x_{n-1}, y_n) \propto p(x_n|x_{n-1})p_1(y_n|x_n)$ as the proposal and the corresponding normalizing constant as the auxiliary variable, where $log\left(p_1(y_n|x_n)\right)$ is the first order Taylor expansion of $p(y_n|x_n)$ around $\phi \sum_{j=1}^{N} x_{n-1}^{(j)}/N$. Since $p(y_n|x_n)$ is log-concave, $p_1(y_n|x_n)$ has heavier tails than $p(y_n|x_n)$ and hence $q_1(x_n|x_{n-1}, y_n)$ and the auxiliary variable satisfy the tail requirement of the proposal distribution. In this example we select $q_1(x_n|x_{n-1}, y_n)$ and $p(x_n|x_{n-1})p_1(y_n|x_n)/q_1(x_n|x_{n-1}, y_n)$ as the proposal and auxiliary variable for the APF method.

Construction of Component Proposals

For the mixture methods, let $q_2(x_n|x_{n-1}, y_n)$ being the normalized $p(x_n|x_{n-1})p_2(y_n|x_n)$ where $log\left(p_2(y_n|x_n)\right)$ is the second order Taylor expansion of $p(y_n|x_n)$ around the maximum point of the observation likelihood, $r(x_{n-1}) \equiv p(x_n|x_{n-1})p_2(y_n|x_n)/q_2(x_n|x_{n-1}, y_n)$ and denote the mean of $q_2(x_n|x_{n-1}, y_n)$ by $\theta(x_{n-1})$. Smith and Santos (2006) discussed the benefit of using $q_2(x_n|x_{n-1}, y_n)$ as the proposal when there are extreme outliers in the observations. One problem of $q_2(x_n|x_{n-1}, y_n)$ is that it does not have heavier tail than $p(x_n|x_{n-1})p_1(y_n|x_n)$.

Since the stochastic volatility model is a special case for the example in Section 5.1.3, its construction of component proposals can be applied in here. The construction gives the auxiliary variable $r(x_{n-1})+c_{n-1}$ where $c_{n-1} = E_{\pi_{n-2}^* q_{\boldsymbol{\alpha}}}[r(x_{n-1})]$ and can be estimated by $\sum_{j=1}^{N} r(x_{n-1}^{(j)})/N$, and the component proposals $q_2(x_n|x_{n-1}, y_n)$, $p(x_n|x_{n-1})$,

CV=10	$\phi = .900$, $\sigma_\eta = .675, \overline{\sigma} = .0165$	$\phi = .950$, $\sigma_\eta = .484, \overline{\sigma} = .0164$	$\phi = .980$, $\sigma_\eta = .308, \overline{\sigma} = .0166$	Time sec
BF	$8.1E-4(.54)$	$4.3E-4(.78)$	$2.3E-4(1.4)$	53
APF	$1.0E-1(.003)$	$2.1E-2(.011)$	$2.7E-4(.80)$	78
GSMC	$5.1E-4(.56)$	$3.6E-4(.61)$	$2.6E-4(.80)$	81
RM-SMC	$3.2E-4(.58)$	$2.6E-4(.54)$	$2.0E-4(.67)$	126
LM-SMC	$1.2E-4(1.0)$	$9.3E-5(1.0)$	$8.7E-5(1.0)$	193
CV=1	$\phi = .900$, $\sigma_\eta = .363, \overline{\sigma} = .0252$	$\phi = .950$, $\sigma_\eta = .260, \overline{\sigma} = .0252$	$\phi = .980$, $\sigma_\eta = .166, \overline{\sigma} = .0253$	Time sec
BF	$2.7E-4(.61)$	$2.2E-4(.78)$	$2.4E-4(1.7)$	53
APF	$4.1E-4(.28)$	$2.5E-4(.45)$	$4.7E-4(.60)$	78
GSMC	$2.3E-4(.46)$	$2.0E-4(.57)$	$2.3E-4(1.2)$	81
RM-SMC	$1.7E-4(.41)$	$1.6E-4(.44)$	$2.1E-4(.82)$	126
LM-SMC	$4.5E-5(1.0)$	$4.6E-5(1.0)$	$1.1E-4(1.0)$	193
CV=.1	$\phi = .900$, $\sigma_\eta = .135, \overline{\sigma} = .0293$	$\phi = .950$, $\sigma_\eta = .096, \overline{\sigma} = .0293$	$\phi = .980$, $\sigma_\eta = .061, \overline{\sigma} = .0295$	Time sec
BF	$4.3E-5(.36)$	$5.2E-5(1.4)$	$3.4E-5(1.7)$	53
APF	$3.9E-5(.27)$	$4.5E-5(1.1)$	$3.2E-5(1.2)$	78
GSMC	$4.4E-5(.23)$	$5.2E-5(.93)$	$3.8E-5(.95)$	81
RM-SMC	$3.2E-5(.20)$	$5.0E-5(.63)$	$3.3E-5(.71)$	126
LM-SMC	$4.3E-6(1.0)$	$2.0E-5(1.0)$	$1.5E-5(1.0)$	193

Table 5.3: Comparison of five methods in Example 5.2.2. $\overline{MSE}$ is reported and the ratio of $\overline{MSE}$ multiplied with computing time between LM-SMC and the corresponding method is reported in the parenthesis. The theoretical posterior mean is calculated using Monte Carlo sample.

normalized $(x_n - \theta(x_{n-1}))^+ q_2(x_n|x_{n-1}, y_n)$, normalized $(x_n - \theta(x_{n-1}))^- q_2(x_n|x_{n-1}, y_n)$, normalized $(x_n - \phi x_{n-1})^+ p(x_n|x_{n-1})$ and normalized $(x_n - \phi x_{n-1})^- p(x_n|x_{n-1})$. The tail requirement can be satisfied by letting the mixture proportion of $p(x_n|x_{n-1})$ larger than 0.

Although all component proposals can be sampled directly, for the purpose of illustration, we follow the suggestion of Section 5.1.4 and set the mixture proportions of $q_2(x_n|x_{n-1}, y_n)$ and $p(x_n|x_{n-1})$ to be .5 for each, and the other proposals to be 0. With such mixture proportions, for different target functions, the sampling procedure remains the same and only the control variates need to be changed.

Results

Here we use the parameter settings in Sandmann and Koopman (1998). The values of the autoregressive parameter ϕ are set to be .90, .95 and .98, which are compatible with the range from .9 to .995 of ϕ found in empirical studies. Then for each ϕ, the values of σ_η are selected so that the coefficient of variation of the volatility $h = \overline{\sigma}^2 exp(x_n)$

$$CV = \frac{Var[h]}{E[h]^2} = exp\left(\frac{\sigma_\eta^2}{1-\phi^2}\right) - 1$$

takes the values 10, 1 and .1. High value of CV indicates the relative strength of the volatility process and low value of CV indicates the volatility is close to a constant. Finally, the average volatility level $\overline{\sigma}$ is selected such that

$$E[h] = \overline{\sigma}^2 exp\left(\frac{\sigma_\eta^2}{2(1-\phi^2)}\right)$$

is equal to .0009. This value of $E[h]$ can be interpreted as an approximately 22% annualized variance if the simulated data are taken as weekly returns. For each setting, a series with length 100 is generated as observations. Results are listed in Table 3 and the trajectories of MSE for all five methods are given in Figure 2. From Table 3, several observations can be made.

In all cases, LM-SMC has the smallest MSE. The improvement of LM-SMC over the others decreases as the value of ϕ increases, i.e. the volatility process becomes more persistent. For $\phi = .9$, considering computing time, the improvement of LM-SMC over BF, APF, GSMC and RM-SMC ranges from 39% to 80%, not including the failed APF. For $\phi = .95$, the improvement of LM-SMC is smaller than those of $\phi = .9$. For $\phi = .98$, in some cases the improvement of LM-SMC is not enough to offset the extra computing time and performs worse than BF.

In the results, the mixture methods performs more robust than APF. When CV=10 and CV=1, the APF performs worse than BF, and in two cases of CV=10, its performance is extremely bad. A possible reason is although $p(y_n|x_n)/p_1(y_n|x_n)$ is bounded, the upper bound may still be large and the sample weights are skewed. In comparison, all three mixture methods are more stable. Therefore the protection provided by the

mixture proposal outperforms the one provided by expanding the log-concave density.

Finally, it can be seen that LM-SMC reduces the MSE of RM-SMC by from 48% to 87%, without considering the computing time. It means that in (5.5), the variance reduction brought by adding the control variates to the historical terms is significant.

5.3 Summary

In this chapter, we propose a new SMC algorithm by using Tan's likelihood approach (Tan, 2004) within both resampling and estimation , and give a practical guideline of selecting control variates and proposal for sampling which are critical for efficient implementation of the algorithm. Compared to the direct use of multiple proposals and control variates, the new algorithm always has smaller asymptotic variance, which is proved in the established theoretical framework. The numerical studies show that, by including the information of target function and introducing heavy tailed proposal for protection, the new algorithm can be more efficient and stable than the bootstrap filter and auxiliary particle filter.

5.4 Technical Proof

The new algorithm differs with the basic SMC method in the use of three elements: The proposal density which is the mixture proposal $q_{\boldsymbol{\alpha}}$, the addition of auxiliary variable $\eta(x_{0:n})$ in resampling, and the new sample weights $v_n^{(j)}$. The extension of Chopin (2004)'s proof to include mixture proposal $q_{\boldsymbol{\alpha}}$ is natural, and the extension to include auxiliary variable can be referred to Johansen and Doucet (2008). Therefore the following proof is focused on including the new sample weights $v_n^{(j)}$ in the central limit theorem. The theorem is proved by inductions. At time $t-1$, assume conditions (C1')-(C4'), (C5) and (C6) are satisfied and the following consistency and the central limit

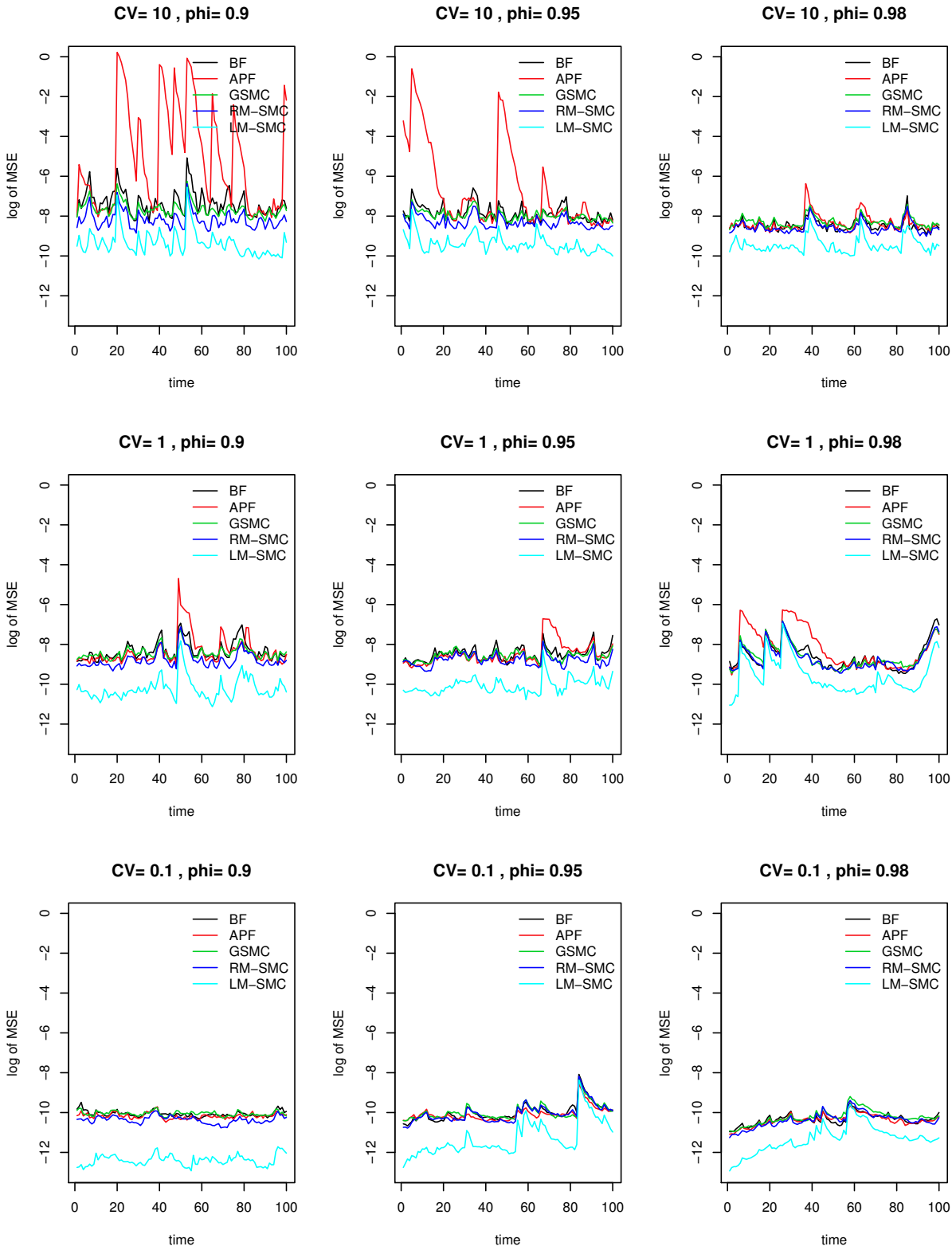

Figure 5.2: Trajectories of logarithm of MSE for the five estimators in all cases of Example 5.2.2.

theorem(CLT) hold:

$$\frac{1}{N}\sum_{j=1}^{N} h(\widetilde{x}_{1:t-1}^{(j)}) \xrightarrow{P} \int h(x_{1:t-1})\widetilde{\pi}_{t-1}^{*}(x_{0:t-1})dx_{0:t-1}, \tag{5.12}$$

$$\sqrt{N}\left(\frac{1}{N}\sum_{j=1}^{N} h(\widetilde{x}_{1:t-1}^{(j)}) - \int h(x_{1:t-1})\widetilde{\pi}_{t-1}^{*}(x_{0:t-1})dx_{0:t-1}\right) \xrightarrow{\mathfrak{L}} N(0, \sigma_{3,t-1}^{2}(h)). \tag{5.13}$$

The following lemma is an extension of the corresponding results in Chopin (2004) to include mixture proposal and auxiliary variable.

Lemma 5.1. *(Mutation) Under the inductive hypothesis, the following convergence results hold:*

$$\frac{1}{N}\sum_{j=1}^{N} h(x_{1:t}^{(j)}) \xrightarrow{P} \int h(x_{1:t})\widetilde{\pi}_{t-1}^{*}(x_{0:t-1})q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})dx_{0:t},$$

$$\sqrt{N}\left[\frac{1}{N}\sum_{j=1}^{N} h(x_{1:t}^{(j)}) - \int h(x_{1:t})\widetilde{\pi}_{t-1}^{*}(x_{0:t-1})q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})dx_{0:t}\right] \xrightarrow{\mathfrak{L}} N(0, \sigma_{1,t}^{2}(f)).$$

In the next lemma, an expansion of the MLE weights $\widehat{\boldsymbol{\zeta}}_t$ is given.

Lemma 5.2. *Under the inductive hypothesis, it holds that*

$$\widehat{\boldsymbol{\zeta}}_t = Var\left[\frac{\boldsymbol{g}}{q_\alpha}\right]^{-1} \frac{1}{N}\sum_{j=1}^{N} \frac{\boldsymbol{g}(x_t^{(j)}|\widetilde{x}_{0:t-1}^{(j)})}{q_\alpha(x_t^{(j)}|\widetilde{x}_{0:t-1}^{(j)})} + o_p\left(N^{-1/2}\right).$$

Therefore $\widehat{\boldsymbol{\zeta}}_t \xrightarrow{P} 0$ *and* $\widehat{\boldsymbol{\zeta}}_t = O_p\left(N^{-1/2}\right)$.

Proof. By definition, $\widehat{\boldsymbol{\zeta}}_t$ is the maximizer of the concave function $\psi(\boldsymbol{s})$
$= \sum_{j=1}^{N} log\left[q_{\boldsymbol{\alpha}}(x_t^{(j)}|\widetilde{x}_{0:t-1}^{(j)}) + \boldsymbol{s}^T\boldsymbol{g}(x_t^{(j)}|\widetilde{x}_{0:t-1}^{(j)})\right]$ where the concavity can be seen in Tan (2004). The standard theory for M-estimation with a convex criterion function cannot apply due to the dependence of $\left\{x_{0:t}^{(j)}\right\}$. Here the argument for the expansion to hold follows Hjort and Pollard (1994, basic corollary) and Li et al. (2012, lemma 4). Then with Lemma 1, the proof is completed. □

With the expansion in Lemma 2, the consistency for the correction and selection holds in the following lemma.

Lemma 5.3. *Under the inductive hypothesis, it holds that*

$$\frac{\sum_{j=1}^{N} h(x_{1:t}^{(j)}) v_t^{(j)}}{\sum_{j=1}^{N} v_t^{(j)}} \xrightarrow{P} \int h(x_{1:t}) \pi_t(x_{0:t}) dx_{0:t}, \tag{5.14}$$

$$\frac{1}{N} \sum_{j=1}^{N} h(\widetilde{x}_{1:t}^{(j)}) \xrightarrow{P} \int h(x_{1:t}) \widetilde{\pi}_t^*(x_{0:t}) dx_{0:t}. \tag{5.15}$$

Proof. The Taylor expansion of $N^{-1} \sum_{j=1}^{N} h(x_{1:t}^{(j)}) v_t^{(j)}$ for $\widehat{\boldsymbol{\zeta}}_t$ around 0 gives that

$$\begin{aligned}
&\frac{1}{N} \sum_{j=1}^{N} h(x_{1:t}^{(j)}) v_t^{(j)} \\
&= \frac{1}{N} \sum_{j=1}^{N} \frac{h(x_{1:t}^{(j)}) \pi_t(x_{0:t}^{(j)})}{\widetilde{\pi}_{t-1}(\widetilde{x}_{0:t-1}^{(j)}) \left[q_{\boldsymbol{\alpha}}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)}) + \widehat{\boldsymbol{\zeta}}_t^T \boldsymbol{g}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)}) \right]} \\
&= \frac{1}{N} \sum_{j=1}^{N} \frac{h(x_{1:t}^{(j)}) \pi_t(x_{0:t}^{(j)})}{\widetilde{\pi}_{t-1}(\widetilde{x}_{0:t-1}^{(j)}) q_{\boldsymbol{\alpha}}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})} - \frac{1}{N} \sum_{j=1}^{N} \frac{h(x_{1:t}^{(j)}) \pi_t(x_{0:t}^{(j)}) \boldsymbol{g}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})^T}{\widetilde{\pi}_{t-1}(\widetilde{x}_{0:t-1}^{(j)}) q_{\boldsymbol{\alpha}}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})^2} \widehat{\boldsymbol{\zeta}}_t + o_p\left(\frac{1}{\sqrt{N}}\right).
\end{aligned}$$

Plugging in the expansion of $\widehat{\boldsymbol{\zeta}}_t$ of Lemma 3 and gives that

$$\begin{aligned}
\frac{1}{N} \sum_{j=1}^{N} h(x_{1:t}^{(j)}) v_t^{(j)} &= \frac{1}{N} \sum_{j=1}^{N} \frac{h(x_{1:t}^{(j)}) \pi_t(x_{0:t}^{(j)})}{\widetilde{\pi}_{t-1}(\widetilde{x}_{0:t-1}^{(j)}) q_{\boldsymbol{\alpha}}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})} \\
&\quad - \frac{1}{N} \sum_{j=1}^{N} \frac{h(x_{1:t}^{(j)}) \pi_t(x_{0:t}^{(j)}) \boldsymbol{g}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})^T}{\widetilde{\pi}_{t-1}(\widetilde{x}_{0:t-1}^{(j)}) q_{\boldsymbol{\alpha}}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})^2} Var\left[\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]^{-1} \frac{1}{N} \sum_{j=1}^{N} \frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}} + o_p\left(\frac{1}{\sqrt{N}}\right) \\
&= \frac{1}{N} \sum_{j=1}^{N} \frac{h(x_{1:t}^{(j)}) \pi_t(x_{0:t}^{(j)})}{\widetilde{\pi}_{t-1}(\widetilde{x}_{0:t-1}^{(j)}) q_{\boldsymbol{\alpha}}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})} - \boldsymbol{\tau}_{1t}^T \frac{1}{N} \sum_{i=1}^{N} \frac{\boldsymbol{g}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})}{q_{\boldsymbol{\alpha}}(x_t^{(j)} | \widetilde{x}_{0:t-1}^{(j)})} + o_p\left(\frac{1}{\sqrt{N}}\right),
\end{aligned} \tag{5.16}$$

where

$$\boldsymbol{\tau}_{1t} = Var\left[\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]^{-1} Cov\left[\frac{h \pi_t}{\widetilde{\pi}_{t-1} q_{\boldsymbol{\alpha}}}, \frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right],$$

and the second equation holds by Lemma 1. Then by Lemma 1 again, from 5.16 we have

$$\frac{1}{N} \sum_{j=1}^{N} h(x_{1:t}^{(j)}) v_t^{(j)} \xrightarrow{P} e_{t-1}^{-1} \int h(x_{1:t}) \pi_t(x_{0:t}) dx_{0:t}.$$

Similarly, $\frac{1}{N} \sum_{j=1}^{N} v_t^{(j)} \xrightarrow{P} e_{t-1}^{-1}$. Therefore (5.14) holds by Slutsky's theorem.

For (5.15), make the decomposition

$$\frac{1}{N}\sum_{j=1}^{N} h(\widetilde{x}_{1:t}^{(j)}) = \frac{1}{N}\sum_{j=1}^{N}\left\{h(\widetilde{x}_{1:t}^{(j)}) - E\left[h(\widetilde{x}_{1:t}^{(j)})|\left\{x_{0:t}^{(j)}\right\}\right]\right\} + \frac{1}{N}\sum_{j=1}^{N} E\left[h(\widetilde{x}_{1:t}^{(j)})|\left\{x_{0:t}^{(j)}\right\}\right].$$

Following the similar argument as Cappé et al. (2005, Theorem 9.2.9), with the fact that $\left\{\widetilde{x}_{0:t}^{(j)}\right\}$ conditional on $\left\{x_{0:t}^{(j)}\right\}$ are i.i.d from multinomial distribution with probability $\left\{\widetilde{w}_t^{(i)}\right\}$, it can be shown that

$$\frac{1}{N}\sum_{j=1}^{N}\left\{h(\widetilde{x}_{1:t}^{(j)}) - E\left[h(\widetilde{x}_{1:t}^{(j)})|\left\{x_{0:t}^{(j)}\right\}\right]\right\} \xrightarrow{P} 0.$$

Then since

$$\frac{1}{N}\sum_{j=1}^{N} E\left[h(\widetilde{x}_{1:t}^{(j)})|\left\{x_{0:t}^{(j)}\right\}\right] = \frac{\sum_{j=1}^{N} h(x_{1:t}^{(j)})\widetilde{v}_t^{(j)}}{\sum_{j=1}^{N}\widetilde{v}_t^{(j)}}$$

and along the same line as proof for (5.14), it holds that

$$\frac{1}{N}\sum_{j=1}^{N} E\left[h(\widetilde{x}_{1:t}^{(j)})|\left\{x_{0:t}^{(j)}\right\}\right] \xrightarrow{P} \int h(x_{1:t})\widetilde{\pi}_t^*(x_{0:t})dx_{0:t},$$

and therefore (5.15) holds. □

Then the CLT of the correction and selection steps can be established.

Proof of Theorem 2. For (5.3), we only need to show the weak convergence of the vector

$$\sqrt{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\begin{pmatrix}\left(h(x_{1:t}^{(j)}) - \int h(x_{1:t})\pi_t(x_{0:t})dx_{0:t}\right)v_t^{(j)} \\ v_t^{(j)}\end{pmatrix} - \begin{pmatrix}0 \\ e_{t-1}^{-1}\end{pmatrix}\right\} \tag{5.17}$$

and then (5.3) is the result of applying the delta method on the weak convergence of (5.17). Similar to the expansion in (5.16), we have the expansion

$$(5.16) = \sqrt{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\begin{pmatrix}(h-\mu_t)\frac{\pi_t}{\widetilde{\pi}_{t-1}q_{\boldsymbol{\alpha}}} - \boldsymbol{\tau}_{1t}^T\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}} \\ \frac{\pi_t}{\widetilde{\pi}_{t-1}q_{\boldsymbol{\alpha}}} - \boldsymbol{\tau}_{2t}^T\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\end{pmatrix} - \begin{pmatrix}0 \\ e_{t-1}^{-1}\end{pmatrix}\right\} + o_p(1), \tag{5.18}$$

where $\boldsymbol{\tau}_{2t} = Var\left[\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]^{-1} Cov\left[\frac{\pi_t}{\widetilde{\pi}_{t-1}q_{\boldsymbol{\alpha}}}, \frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\right]$.

The first term of (5.18) weakly converges to a multivariate normal distribution by the generalization of Lemma 1 to two dimension. See Chopin (2004) for the generalization

of the corresponding results to multivariate case. Therefore (5.17) also weakly converges to the same multivariate normal distribution and (5.3) holds by the delta method.

For (5.4), make the decomposition

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(h\left(\widetilde{x}_{1:t}^{(j)}\right)-E\left[h\left(\widetilde{x}_{1:t}^{(j)}\right)\mid\left\{x_{0:t}^{(j)}\right\}\right]\right)\\
&+\sqrt{N}\left(\frac{1}{N}\sum_{j=1}^{N}E\left[h\left(\widetilde{x}_{1:t}^{(j)}\right)\mid\left\{x_{0:t}^{(j)}\right\}\right]-\int h(x_{1:t})\widetilde{\pi}_t^*(x_{0:t})dx_{0:t}\right)\\
\equiv&A_N+B_N.
\end{aligned}$$

Following the similar argument in Cappé et al. (2005, Theorem 9.2.14), with the fact that

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N}Var\left[h\left(\widetilde{x}_{1:t}^{(j)}\right)\mid\left\{x_{0:t}^{(j)}\right\}\right]&=\frac{\sum_{j=1}^{N}h(x_{1:t}^{(j)})^2\widetilde{v}_t^{(j)}}{\sum_{j=1}^{N}\widetilde{v}_t^{(j)}}-\left(\frac{\sum_{j=1}^{N}h(x_{1:t}^{(j)})\widetilde{v}_t^{(j)}}{\sum_{j=1}^{N}\widetilde{v}_t^{(j)}}\right)^2\\
&\xrightarrow{P}Var_{\widetilde{\pi}_t^*}\left[h(x_{1:t})\right],
\end{aligned}$$

it can be shown that

$$E\left[exp\left(iuA_N\right)\mid\left\{x_{0:t}^{(j)}\right\}\right]\xrightarrow{P}exp\left(-\frac{u^2}{2}Var_{\widetilde{\pi}_t^*}\left[h(x_{1:t})\right]\right).$$

Since $E\left[h\left(\widetilde{x}_{1:t}\right)\mid\left\{x_{1:t}^{(j)}\right\}\right]=\sum_{j=1}^{N}h(x_{1:t}^{(j)})\widetilde{v}_t^{(j)}/\sum_{j=1}^{N}\widetilde{v}_t^{(j)}$, by (5.3), it holds that

$$\begin{aligned}
B_N&\xrightarrow{\mathfrak{L}}N(0,\sigma_{1,t}^2(\frac{\widetilde{\pi}_t^*(x_{0:t})(h-\widetilde{\mu}_t)}{\widetilde{\pi}_{t-1}^*(x_{0:t-1})q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})}-\boldsymbol{\gamma}_t^T\frac{\boldsymbol{g}}{q_\alpha})),\\
\text{where }\boldsymbol{\gamma}_t&=Var\left[\frac{\boldsymbol{g}}{q_\alpha}\right]^{-1}Cov\left[\frac{\widetilde{\pi}_t^*(h-\widetilde{\mu}_t)}{\widetilde{\pi}_{t-1}^*q_{\boldsymbol{\alpha}}},\frac{\boldsymbol{g}}{q_\alpha}\right].
\end{aligned}$$

Then follow the argument in Cappé et al. (2005, Theorem 9.2.14), (5.4) holds. □

For applying Owen and Zhou's regression approach in the estimation of the generalized SMC method, which results in $\widehat{\mu}_{n,Reg}$ in (5.2), its asymptotic variance is stated

below. Let

$$\sigma'^{\,2}_{1,n}(h) = \sigma'^{\,2}_{3,n-1}(E_{q_\alpha}[h|x_{0:n-1}]) + E_{\widetilde{\pi}^*_{n-1}}[Var_{q_{\boldsymbol{\alpha}}}[h|x_{0:n-1}]],$$
$$\sigma'^{\,2}_{2,n}(h) = \sigma'^{\,2}_{1,n}(\frac{\pi^*_n(x_{0:n})(h-\mu_n)}{\widetilde{\pi}^*_{n-1}(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})} - \boldsymbol{\beta}_n^T\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}),$$
$$\sigma'^{\,2}_{3,n}(h) = \sigma'^{\,2}_{1,n}(\frac{\widetilde{\pi}^*_n(x_{0:n})(h-\widetilde{\mu}_n)}{\widetilde{\pi}^*_{n-1}(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})}) + Var_{\widetilde{\pi}^*_n}[h].$$

Proposition 5.1. *For any n, $\sigma'^{\,2}_{2,n}(h)$ and $\sigma'^{\,2}_{3,n}(h)$ are finite and the following convergence holds:*

$$\sqrt{N}\left(\widehat{\mu}_{n,Reg} - \mu_n\right) \xrightarrow{\mathfrak{L}} N(0, \sigma'^{\,2}_{2,n}(h)), \tag{5.19}$$
$$\sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N} f(\widetilde{x}^{(i)}_{1:n}) - \widetilde{\mu}_n\right] \xrightarrow{\mathfrak{L}} N(0, \sigma'^{\,2}_{3,n}(h)). \tag{5.20}$$

Analytically,

$$\sigma'^{\,2}_{2,n}(h) = \sum_{t=1}^{n-1}\int \frac{\pi^*_n(x_{0:t})(\mu_t(x_{0:t}) - \mu_n)}{\widetilde{\pi}^*_{t-1}(x_{0:t-1})q_{\boldsymbol{\alpha}}(x_t|x_{0:t-1})}dx_{0:t}$$
$$+ \int \frac{\left[\pi^*_n(x_{0:n})(h(x_{1:n}) - \mu_n) - \boldsymbol{\beta}^T_{nn}\widetilde{\pi}^*_{n-1}(x_{0:n-1})\boldsymbol{g}(x_n|x_{0:n-1})\right]^2}{\widetilde{\pi}^*_{n-1}(x_{0:n-1})q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})}dx_{0:n}.$$

Proof. Consider the generalized SMC method with estimator $\sum_{j=1}^N h(x_n^{(j)})\omega_n^{(j)} / \sum_{j=1}^N \omega_n^{(j)}$. (5.20) is the CLT for the selection step. The theoretical results for the mutation, correction and selection steps, similar to Theorem 1, can be established by the extension of Theorem 1 to mixture proposal $q_{\boldsymbol{\alpha}}$ and the results for auxiliary particle filter in Johansen and Doucet (2008). Then Following the similar arguments as the expansion (5.18), at time n it holds that

$$\sqrt{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\begin{pmatrix}(h-\mu_n)\,\omega_n^{(j)} - \widehat{\boldsymbol{\tau}}^T_{1n}\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\\ \omega_n^{(j)} - \widehat{\boldsymbol{\tau}}^T_{2n}\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\end{pmatrix} - \begin{pmatrix}0\\ e_{n-1}^{-1}\end{pmatrix}\right\}$$
$$= \sqrt{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\begin{pmatrix}(h-\mu_t)\,\omega_n^{(j)} - \boldsymbol{\tau}^T_{1t}\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\\ \omega_n^{(j)} - \boldsymbol{\tau}^T_{2t}\frac{\boldsymbol{g}}{q_{\boldsymbol{\alpha}}}\end{pmatrix} - \begin{pmatrix}0\\ e_{t-1}^{-1}\end{pmatrix}\right\} + o_p(1) \tag{5.21}$$

by applying CLT of the mutation step on the function vector $\boldsymbol{g}(x_n|x_{0:n-1})/q_{\boldsymbol{\alpha}}(x_n|x_{0:n-1})$ and the consistency of mutation step on the statistics $\widehat{\boldsymbol{\tau}}_{1n}$ and $\widehat{\boldsymbol{\tau}}_{2n}$. Then (5.19) holds by the delta method, and the analytical expression of $\sigma'^{\,2}_{2,n}(h)$ can be obtained through

algebra. □

Bibliography

Battiti, R. and Masulli, F. (1990). BFGS optimization for faster and automated supervised learning. In *Proceedings of the International Neural Network Conference (INNC 90)-Paris-France*, pages 757–760.

Bauwens, L. and Lubrano, M. (2008). Bayesian inference on GARCH models using the Gibbs sampler. *The Econometrics Journal*, 1(1):23–46.

Binder, K. and Heermann, D. W. (2010). *Monte Carlo simulation in statistical physics: an introduction*. Springer.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

Briers, M., Doucet, A., and Maskell, S. (2010). Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61–89.

Caflisch, R. E. (1998). Monte carlo and quasi-monte carlo methods. *Acta numerica*, 1998:1–49.

Cappé, O., Godsill, S., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential monte carlo. *Proceedings of the IEEE*, 95(5):899–924.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Verlag.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE proceedings. Radar, sonar and navigation*, 146(1):2–7.

Cérou, F., Del Moral, P., Furon, T., and Guyader, A. (2012). Sequential Monte Carlo for rare event estimation. *Statistics and Computing*, 22(3):795–808.

Chan, H. and Lai, T. (2011). A sequential Monte Carlo approach to computing tail probabilities in stochastic models. *The Annals of Applied Probability*, 21(6):2315–2342.

Chen, R. (2005). Sequential Monte Carlo methods and their applications. *IMS Lecture Notes Series, Markov Chain Monte Carlo*, 7:147–182.

Chen, R. and Liu, J. (2000). Mixture kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):493–508.

Chen, Y., Diaconis, P., Holmes, S., and Liu, J. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–120.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411.

Chow, Y. and Teicher, H. (2003). *Probability theory: independence, interchangeability, martingales*. Springer Verlag.

Cochran, W. (1977). *Sampling Techniques*. New York: Wiley.

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

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Denny, M. (2001). Introduction to importance sampling in rare-event simulations. *European Journal of Physics*, 22:403.

Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 64–69. IEEE.

Doucet, A., Briers, M., and Sénécal, S. (2006). Efficient block sampling strategies for sequential Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 15(3):693–711.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208.

Doucet, A. and Johansen, A. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, pages 656–704.

Duffie, D. and Pan, J. (1997). An overview of value at risk. *The Journal of derivatives*, 4(3):7–49.

Dunkel, J. and Weber, S. (2007). Efficient Monte Carlo methods for convex risk measures in portfolio credit risk models. In *Simulation Conference, 2007 Winter*, pages 958–966. IEEE.

Dupuis, J. A. (1995). Bayesian estimation of movement and survival probabilities from capture-recapture data. *Biometrika*, 82(4):761–772.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. Oxford University Press.

Durrett, R. (1996). *Probability: theory and examples*. Duxbury Press.

Elinas, P., Sim, R., and Little, J. (2006). $\sigma SLAM$: stereo vision SLAM using the Rao-Blackwellised particle filter and a novel mixture proposal distribution. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 1564–1570. Ieee.

Emond, M., Raftery, A., and Steele, R. (2001). Easy computation of Bayes factors and normalizing constants for mixture models via mixture importance sampling. Technical report, Department of Statistics, Washiton University Seattle.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. *Statistical Science*, 10(3):254–272.

Evans, M. and Swartz, T. (1996). Bayesian integration using multivariate student importance sampling. *Computing Science and Statistics*, pages 456–461.

Fan, S., Chenney, S., Hu, B., Tsui, K., and Lai, Y. (2006). Optimizing control variate estimators for rendering. In *Computer Graphics Forum*, volume 25, pages 351–357. Eurographics Association.

Fearnhead, P. (1998). *Sequential Monte Carlo methods in filter theory*. PhD thesis, University of Oxford.

Fearnhead, P. (2008). Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Statistics and Computing*, 18(2):151–171.

Ford, E. and Gregory, P. (2007). Bayesian model selection and extrasolar planet detection. In *Statistical Challenges in Modern Astronomy IV*, volume 371, page 189.

Fox, D., Thrun, S., Burgard, W., and Dellaert, F. (2001). Particle filters for mobile robot localization. *Sequential Monte Carlo Methods in Practice*, pages 401–428.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339.

Geweke, J. (1994). Bayesian comparison of econometric models. *Federal Reserve bank of Minneapolis working paper*, 532.

Geweke, J. (1996). Monte carlo simulation and numerical integration. *Handbook of Computational Economics*, 1:731–800.

Geyer, C. (1994). On the asymptotics of constrained M-estimation. *The Annals of statistics*, 22(4):1993–2010.

Gilks, W. and Berzuini, C. (2001). Following a moving target-Monte Carlo inference for dynamic bayesian models. *Journal of The Royal Statistical Society Series B*, 63(1):127–146.

Giordani, P. and Kohn, R. (2010). Adaptive independent Metropolis-Hastings by fast estimation of mixtures of normals. *Journal of Computational and Graphical Statistics*, 19(2):243–259.

Givens, G. and Raftery, A. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association*, 91(433):132–141.

Glasserman, P., Heidelberger, P., and Shahabuddin, P. (2000). Variance reduction techniques for estimating value-at-risk. *Management Science*, 46(10):1349–1364.

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 107–113. IET.

Guo, D., Wang, X., and Chen, R. (2005). New sequential Monte Carlo methods for nonlinear dynamic systems. *Statistics and Computing*, 15(2):135–147.

Haberman, S. (1989). Concavity and estimation. *The Annals of Statistics*, 17(4):1631–1661.

Haug, A. J. (2012). *Bayesian Estimation and Tracking: A Practical Guide*. Wiley.

Hesterberg, T. (1988). *Advances in importance sampling*. PhD thesis, Stanford University.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194.

Hesterberg, T. (1996). Control variates and importance sampling for efficient bootstrap simulations. *Statistics and Computing*, 6(2):147–157.

Hjort, N. and Pollard, D. (1994). Asymptotics for minimisers of convex processes. Statistical Research Report, Department of Mathematics, University of Oslo.

Hoogerheide, L. and Van Dijk, H. (2010). Bayesian forecasting of value at risk and expected shortfall using adaptive importance sampling. *International Journal of Forecasting*, 26(2):231–247.

Johansen, A. and Doucet, A. (2008). A note on auxiliary particle filters. *Statistics & Probability Letters*, 78(12):1498–1504.

Jorion, P. (1997). *Value At Risk: The New Benchmark For Controlling Market Risk*. McGraw-Hill Chicago.

Kahn, H. (1949). Modification of the monte carlo methods. In *International Business Machine Corporation Proceedings, Seminar on Scientific Computation*, pages 20–27. International Business Machine Corporation.

Kahn, H. and Harris, T. (1949). Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series*, 12:27–32.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Kitagawa, G. (1996). Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of computational and graphical statistics*, 5(1):1–25.

Kong, A., McCullagh, P., Meng, X., Nicolae, D., and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):585–604.

Li, W., Tan, Z., and Chen, R. (2012). Two-stage imortance sampling with mixture proposals. Working paper (available upon request), Rutgers University.

Liang, F., Liu, C., and Carroll, R. J. (2007). Stochastic approximation in monte carlo computation. *Journal of the American Statistical Association*, 102(477):305–320.

Lin, M., Zhang, J., Cheng, Q., and Chen, R. (2005). Independent particle filters. *Journal of the American Statistical Association*, 100(472):1412–1421.

Liu, J. (2008). *Monte Carlo strategies in scientific computing*. Springer Verlag.

Liu, J. and Chen, R. (1995). Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90(430):567–576.

Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032.

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice*. Springer.

Liu, J. S., Chen, R., and Logvinenko, T. (2001). A theoretical framework for sequential importance sampling with resampling. In *Sequential Monte Carlo Methods in Practice*, pages 225–246. Springer.

McAllister, M. K. and Ianelli, J. N. (1997). Bayesian stock assessment using catch-age data and the sampling: importance resampling algorithm. *Canadian Journal of Fisheries and Aquatic Sciences*, 54(2):284–300.

Miguez, J. and Djuric, P. M. (2002). Blind equalization by sequential importance sampling. In *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, volume 1, pages I–845. IEEE.

Nocedal, J. and Wright, S. (1999). *Numerical optimization*. Springer verlag.

Oh, M. and Berger, J. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association*, pages 450–456.

O'Hagan, A. (1987). Monte carlo is fundamentally unsound. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 36(2/3):247–249.

Owen, A. and Zhou, Y. (2000). Safe and Effective Importance Sampling. *Journal of the American Statistical Association*, 95(449).

Owen, A. B. and Zhou, Y. (1999). Adaptive importance sampling by mixtures of products of beta distributions. Technical report, Department of Statisticsm, Stanford University.

Pitt, M. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Raftery, A. E., Givens, G. H., and Zeh, J. E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association*, 90(430):402–416.

Raghavan, N. and Cox, D. (1998). Adaptive mixture importance sampling. *Journal of Statistical Computation and Simulation*, 60(3):237–260.

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

Rosenbluth, M. N. and Rosenbluth, A. W. (1955). Monte carlo calculation of the average extension of molecular chains. *The Journal of Chemical Physics*, 23:356.

Rothenberg, T. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics*, 2:881–935.

Rubin, D. B. (1988). Using the sir algorithm to simulate posterior distributions. *Bayesian statistics*, 3:395–402.

Rubinstein, R. and Kroese, D. (2008). *Simulation and the Monte Carlo method*, volume 707. Wiley-Interscience.

Saha, S., Mandal, P., Boers, Y., Driessen, H., and Bagchi, A. (2009). Gaussian proposal density using moment matching in SMC methods. *Statistics and Computing*, 19(2):203–208.

Sandmann, G. and Koopman, S. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics*, 87(2):271–301.

Shephard, N. (2005). *Stochastic Volatility: Selected Readings*. Advanced Texts in Econometrics Series. Oxford University Press.

Singh, S., Kantas, N., Vo, B., Doucet, A., and Evans, R. (2007). Simulation-based optimal sensor scheduling with application to observer trajectory planning. *Automatica*, 43(5):817–830.

Singh, S., Vo, B., Doucet, A., and Evans, R. (2004). Variance reduction for Monte Carlo implementation of adaptive sensor management. In *Proceedings of the international conference on information fusion*, pages 901–908.

Smith, J. and Santos, A. (2006). Second-order filter distribution approximations for financial time series with extreme outliers. *Journal of Business and Economic Statistics*, 24(3):329–337.

Smith, P., Shafi, M., and Gao, H. (1997). Quick simulation: A review of importance sampling techniques in communications systems. *Selected Areas in Communications, IEEE Journal on*, 15(4):597–613.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *Journal of the American Statistical Association*, 99(468):1027–1036.

Thrun, S., Fox, D., Burgard, W., and Dellaert, F. (2001). Robust Monte Carlo localization for mobile robots. *Artificial intelligence*, 128(1):99–141.

Van der Vaart, A. (2000). *Asymptotic statistics*. Cambridge University Press.

Van Der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*. Springer Verlag.

Veach, E. and Guibas, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 419–428. ACM.

Wang, X., Chen, R., and Guo, D. (2002). Delayed-pilot sampling for mixture Kalman filter with application in fading channels. *Signal Processing, IEEE Transactions on*, 50(2):241–254.

West, M. (1993). Approximating posterior distributions by mixture. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2):409–422.