Some Aspects of Sequential Monte Carlo

Abstract: The most general form of the sequential Monte Carlo method is described and history reviewed. A few success stories of the methodology, e.g., phylogenetic inference, estimating normalizing constant, and deconvolving wireless digital signals, are used as illustration.

1 General Sequential Monte Carlo Methodology

The Monte Carlo method has been used in major scientific researches since the World War II. Although it is known to most scientists as a "statistical sampling approach for solving numerical problems concerned with a complex system," the Monte Carlo method was never seriously considered by mainstream statisticians before 1980s. Its "rediscovery" by statisticians as one of the most versatile and powerful computational tools was, to a large extent, motivated by the need to integrate out nuisance parameters in likelihood-based statistical inferences, especially Bayesian inferences, for complex models [4]. Monte Carlo methods can be loosely classified into two types: one is based on the Markov chain theory (i.e., Markov chain Monte Carlo) and the other based on the importance sampling. The sequential Monte Carlo method is a combination of both: its central pillar, the sequential importance sampling (SIS), clearly belongs to the second type, whereas some recent modifications incorporate also MCMC steps [6].

The SIS-based methods have been invented independently in at least three research areas. The first invention dates back to the 1950s and was motivated by a polymer simulation problem. The other two were more recent: one was motivated by statistical missing data problems the other by a nonlinear filtering problem. The key idea of the SIS can be very generally and simply formulated and applied to different problems. The applications of the SIS and its variations now include computer vision, finance, genetic linkage analysis, medical diagnosis, target tracking, wireless signal processing, and statistical computing [4].

Suppose we can decompose $\boldsymbol{x}$ as $(x_1, \ldots, x_d)$ where each of the x_j may be multidimensional. Denote $\boldsymbol{x}_t = (x_1, \ldots, x_t)$ (thus, $\boldsymbol{x}_d \equiv \boldsymbol{x}$). The trial density can be constructed as

$$g(\boldsymbol{x}) = g_1(x_1)g_2(x_2 \mid x_1) \cdots g_d(x_d \mid \boldsymbol{x}_{d-1}), \tag{1}$$

by which we hope to obtain some guidance from the target density while building up the trial density. Corresponding to the decomposition of $\boldsymbol{x}$, we can rewrite the target density as

$$\pi(\boldsymbol{x}) = \pi(x_1)\pi(x_2 \mid x_1) \cdots \pi(x_d \mid \boldsymbol{x}_{d-1}), \tag{2}$$

implying that if we can let $g_i(x_i|\boldsymbol{x}_{i-1})$ be equal to $\pi(x_i|\boldsymbol{x}_{i-1})$, then the sample drawn from $g(\boldsymbol{x})$ would be a perfect one. The importance weight can be decomposed accordingly:

$$w(\boldsymbol{x}) = \frac{\pi(x_1)\pi(x_2 \mid x_1) \cdots \pi(x_d \mid \boldsymbol{x}_{d-1})}{g_1(x_1)g_2(x_2 \mid x_1) \cdots g_d(x_d \mid \boldsymbol{x}_{d-1})}, \tag{3}$$

suggesting that we can monitor the importance weight recursively by computing

$$w_t(\boldsymbol{x}_t) = w_{t-1}(\boldsymbol{x}_{t-1})\pi(x_t \mid \boldsymbol{x}_{t-1})/g_t(x_t \mid \boldsymbol{x}_{t-1}), \quad \text{for} \quad t = 1, \ldots, d.$$

However, both (2) and (3) are impossible to use because one needs, at least, to be able to compute $\pi(\boldsymbol{x}_t) = \int \pi(\boldsymbol{x}) dx_{t+1} \cdots dx_d$, which is often more difficult than the original problem.

Suppose, however, that we can find a sequence of "auxiliary distributions," $\pi_1(x_1), \pi_2(\boldsymbol{x}_2), \ldots,$ so that $\pi_t(\boldsymbol{x}_t)$ is a reasonable approximation to the marginal distribution $\pi(\boldsymbol{x}_t)$, for $t = 1, \ldots, d-1$, and $\pi_d = \pi$. We want to emphasize that the π_t are only required to be known up to a normalizing constant and they *only* serve as "guides" to our construction of $\boldsymbol{x}$. The SIS method can then be defined as the following recursive procedure (for $t = 2, \ldots, d$):

(A) Draw $X_t{=}x_t$ from $g_t(x_t|\boldsymbol{x}_{t-1})$, and let $\boldsymbol{x}_t = (\boldsymbol{x}_{t-1}, x_t)$.

(B) Compute

$$u_t = \frac{\pi_t(\boldsymbol{x}_t)}{\pi_{t-1}(\boldsymbol{x}_{t-1}) g_t(x_t \mid \boldsymbol{x}_{t-1})}, \tag{4}$$

and let $w_t = w_{t-1} u_t$. Here u_t is called an "incremental weight."

It is easy to show that $\boldsymbol{x}_t$ is properly weighted by w_t with respect to π_t, provided that $\boldsymbol{x}_{t-1}$ is properly weighted by w_{t-1} with respect to π_{t-1}. The auxiliary distribution sequence $\pi_1, \ldots, \pi_d$ can be used to help construct a more efficient trial distribution:

- We may choose g_t that is sufficiently close to π_t, e.g., let $g_t(x_t|\boldsymbol{x}_{t-1}) = \pi_t(x_t|\boldsymbol{x}_{t-1})$, when possible. Then, the incremental weight becomes $u_t = \pi_t(\boldsymbol{x}_t)/\pi_{t-1}(\boldsymbol{x}_{t-1})$.
- When some w_t are too small, we may want to reject the sample halfway and restart again. Since an outright rejection incurs bias, the rejection control technique can be applied [7].

The most "artistic" and flexible part of the SIS method is to find a good set of auxiliary distributions. This flexibility is key to the versatility of SIS.

A recent innovation that significantly improved the applicability and efficiency of the SIS is known as *resampling* in the statistics community, and *pruning and enrichment* in the polymer simulation community. Suppose we carried out m independent SIS processes in parallel. Then at any time t, we have a set of samples $\mathcal{S}_t = \{(\boldsymbol{x}_t^{(j)}, w_t^{(j)})\}_{j=1}^m$, in which the $\boldsymbol{x}_t^{(j)}$ are properly weighted by the $w_t^{(j)}$ with respect to π_t. By treating $\mathcal{S}_t$ as a discrete representation of π_t, we can generate another discrete representation as follows:

- For $j' = 1, \ldots, \tilde{m}$,
  - let $\tilde{\boldsymbol{x}}_t^{(j')}$ be a random draw from the set $\{\boldsymbol{x}_t^{(j)}\}_{j=1}^m$ with probability proportional to $\{a^{(j)}\}_{j=1}^m$;
  - let the new weight of $\tilde{\boldsymbol{x}}_t^{(j')}$ be $\tilde{w}_t^{(j')} = w_t^{(j)}/a^{(j)}$.
- Return the new representation $\tilde{\mathcal{S}}_t = \{(\tilde{\boldsymbol{x}}_t^{(j')}, \tilde{w}_t^{(j')})\}_{j'=1}^{\tilde{m}}$.

The new set $\tilde{\mathcal{S}}_t$ thus formed is also (approximately) proper with respect to π_t. It is not obvious, however, why resampling is useful. In fact, it does not help at all in a static importance sampling scheme. A few heuristics [5] are as follows: (a) resampling can prune away those hopelessly bad samples (by giving them a small $a^{(j)}$) and (b) resampling can produce multiple copies of those good samples (by giving them a big $a^{(j)}$) to help generate better future samples in the SIS setting.

Consequently, resampling can *steer* the sampling towards the right direction. In light of these arguments, one should choose $a^{(j)}$ as a monotone function of $w_t^{(j)}$. If we let $a^{(j)} = w_t^{(j)}$, then the foregoing scheme is exactly the same as the one described earlier [3, 5]. But having additional flexibility in choosing the sampling weights $a^{(j)}$ is rather intriguing and can potentially be very useful.

2 Sequential Monte Carlo in Action.

Normalizing Constant and Contingency Tables. The SIS approach can be used to estimate the normalizing constant. Suppose the target distribution is $\pi(\boldsymbol{x})$ and an auxiliary distribution sequence is found as $\Pi = \{\pi_t(\boldsymbol{x}_t),\ t = 1, \ldots, N\}$ (with $\pi_N \equiv \pi$). Each member in Π is known up to a normalizing constant. In other words, we know the unnormalized distribution function, $q_t(\boldsymbol{x}_t) = Z_t \pi_t(\boldsymbol{x}_t)$, for $\pi_t \in \Pi$. Suppose an SIS scheme with the trial distribution $\{g_t(x_t \mid \boldsymbol{x}_{t-1}),\ t = 1, \ldots, N\}$ is implemented. The unnormalized incremental weight u_t can be computed as

$$u_t = \frac{q_t(\boldsymbol{x}_t)}{q_{t-1}(\boldsymbol{x}_{t-1}) g_t(x_t \mid \boldsymbol{x}_{t-1})} = \frac{Z_t \pi_t(\boldsymbol{x}_t)}{Z_{t-1} \pi_{t-1}(\boldsymbol{x}_{t-1}) g_t(x_t \mid \boldsymbol{x}_{t-1})}.$$

The final unnormalized weight takes the form

$$w_N = \prod_{t=1}^{N} u_t = \frac{Z_N}{Z_1} \frac{\pi_N(\boldsymbol{x}_N)}{g_1(x_1) \cdots g_N(x_N \mid \boldsymbol{x}_{N-1})}. \tag{5}$$

Thus, the average of w_N is an unbiased estimate of $E(w_N) = Z_N / Z_1$.

We can use the above procedure to estimate the total number Z of distinctive tables that contain only 0's and 1's and have the fixed marginal row sums $r_1, \ldots, r_m$ and column sums $c_1, \ldots, c_n$. The reason is that the uniform distribution on space of all such tables is of the form $1/Z$. Briefly, the SIS begins by filling in columns of the $m \times n$ table from left to right sequentially. After the first $t-1$ columns are filled, the row sums are updated and the tth column is filled in by sampling c_t of its m possible positions to put in 1's. These c_t positions are sampled according to a distribution related to the updated row sums. For example, one can sample these positions with probability proportional to the product of the corresponding row sums. At the end, we can estimate Z by averaging the inverse of the generation probabilities of the obtained tables. This SIS method gives us a rather accurate estimate for fairly large tables. More details and other applications can be found in Y. Chen's Ph.D. thesis at Stanford.

Phylogenetic Inference. Evolutionary theory holds that stochastic mutational events may alter the genome of an individual and that these changes may be passed to its progeny. Thus, comparing homologous DNA regions (segments) of a random sample of individuals taken from a certain population can shed light on the evolutionary process of this population.

A useful simple demographic model assumes that a haploid population evolves in non-overlapping generations for an infinite time and is of constant size N throughout the history. Each individual in the population is a sufficiently small chromosomal region in which no recombination is allowed. Thus, each chromosomal segment seen in the dataset can only differ from its parental segment by mutations. For a given mutation rate θ, it is easy to write down the probability for generating a particular segment given the parental chromosomal segment. But it is much more difficult to

estimate θ from a random sample of size n drawn from the current population. Because of the nature of the model, all the individuals in the sample have to coalesce at some point in the history. However, the coalescence process $\mathcal{H}$ is completely missing (the estimation of θ would have been trivial had we known it). Given θ, one can "impute" $\mathcal{H}$ by SIS. The SIS needs to simulate a "backward process" with an uncertain period of time. The auxiliary sequences, however, correspond to a set of forward probabilities for evolution. More details can be found in [4].

Mixture Kalman filter and digital signal extraction. Many mobile communication channels can be modeled as Rayleigh flat-fading channels, which have the following form:

$$\begin{array}{rl} \text{State equations:} & \left\{ \begin{array}{rcl} \boldsymbol{x}_t & = & F\boldsymbol{x}_{t-1} + Ww_t \\ \alpha_t & = & G\boldsymbol{x}_t \\ s_t & \sim & p(\cdot \mid s_{t-1}), \end{array} \right. \\ \text{Observation equation:} & y_t = \alpha_t s_t + Vv_t, \end{array}$$

where s_t are the input digital signals (symbols), y_t are the received complex signals, and α_t are the unobserved (changing) fading coefficients. Both w_t and v_t are complex Gaussian with identity covariance matrices. It is important to note that given the input signals s_t, the system is linear in $\boldsymbol{x}_t$ and y_t. Therefore, we can design a special SIS algorithm, the mixture Kalman filter, which focuses solely on the s_t (with $\boldsymbol{x}_t$ integrated out). That is, given $\boldsymbol{s}_t = (s_1, \ldots, s_t)$, we observe that $p(x_t \mid \boldsymbol{y}_t)$ is Gaussian and its mean and variance can be computed recursively by a Kalman filtering technique. Thus, we can use SIS to simulate the discrete variable $s_1, \ldots, s_t$, conditional on which the system is easy to handle. Algorithmic details can be found in [1, 2].

References

[1] R. Chen and J. S. Liu. Mixture kalman filters. *J. Roy. Statist. Soc. B*, 62:493–508, 2000.

[2] R. Chen, X. D. Wang, and J. S. Liu. Adaptive joint detection and decoding in flat-fading channels via mixture kalman filtering. *IEEE Trans. Info. Theory*, 46(6):2079–94, 2000.

[3] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear non-gaussian bayesian state estimation. *IEE Proceedings on Radar and Signal Processing*, 140:107–113, 1993.

[4] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, 2001.

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

[6] J. S. Liu and R. Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

[7] J. S. Liu, R. Chen, and W. H. Wong. Rejection control and sequential importance sampling. *Journal of the American Statistical Association*, 93(443):1022–1031, 1998.