Advanced Statistical Computing

Fall 2016

Steve Qin
Outline

• Collapsing, predictive updating
• Sequential Monte Carlo
Collapsing and grouping

• Want to sample from $\mathbf{x} = (x_1, x_2, \ldots, x_d)$

• Regular Gibbs sampler:
  – Sample $x_1^{(t+1)}$ from $\pi(x_1^{(t+1)} \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_d^{(t)})$,
  – Sample $x_2^{(t+1)}$ from $\pi(x_2^{(t+1)} \mid x_1^{(t)}, x_3^{(t)}, \ldots, x_d^{(t)})$,
  – …
  – Sample $x_d^{(t+1)}$ from $\pi(x_d^{(t+1)} \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_{d-1}^{(t)})$,

• Alternatively:
  – Grouping: $\mathbf{x}_{d-1}' = (x_{d-1}, x_d)$.
  – Collapsing, i.e., integrate out $x_d$: $\mathbf{x}^- = (x_1, x_2, \ldots, x_{d-1})$
The three-schemes

standard

grouping

collapsing
Some theory

• Hilbert space $L_2(\pi)$ of functions $h()$.
• Define $\langle h, g \rangle = E_\pi \{h(x)g(x)\}$, thus $\|h\| = \text{var}_\pi(h)$.
• Define forward operator $F$ as
  $$Fh(x) = \int K(x, y)h(y)dy = E_\pi \{h(x^{(t+1)}) | x^{(t)} = x\};$$
  $$\|F\| = \sup_h \|Fh(x)\|$$ for all functions with $E(h^2) = 1$.
• The convergence of Markov chains is tied to the norms of the corresponding forward operators.
Three-scheme theorem

- Standard $F_s$: $x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_d$;
- Grouping $F_g$: $x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow \{x_{d-1}, x_d\}$;
- Collapsing $F_c$: $x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_{d-1}$.

**Theorem** The norms of the three forward operators are ordered as

$$
\|F_c\| \leq \|F_g\| \leq \|F_s\|
$$
Examples

- Murray’s data
- Bivariate Gaussian with mean 0 and unknown covariance matrix $\Sigma$

\[
\begin{array}{cccccccc}
1 & 1 & -1 & -1 & 2 & 2 & -2 & -2 \\
1 & -1 & 1 & -1 & * & * & * & * \\
\end{array}
\]

\[
\begin{array}{cccccccc}
* & * & * & * & 2 & 2 & -2 & -2 \\
\end{array}
\]

standard

$\Sigma | y_{obs}, y_{mis},$

$y_{mis} | y_{obs}, \Sigma.$

collapsing

$y_{mis,i} | y_{obs}, y_{mis,-[i]}.$
Remarks

• Avoid introducing unnecessary parameters into a Gibbs sampler,
• Do as much analytical work as possible,
• However, introducing some clever auxiliary variables can greatly improve computation efficiency.
Sequential Monte Carlo

• We wish to evaluate an integral
  \[ \theta = \int h(x)\pi(x)dx = E_\pi[h(X)]. \]
  assume \( h(x) \geq 0. \)

• Riemann sum (on grid points) as approximation.

• Alternatively, use Monte Carlo. Select random samples uniformly on its support.
An example

\[ f(x, y) = 0.5e^{-90(x-0.5)^2-45(y+0.1)^4} + e^{-45(x+0.4)^2-60(y-0.5)^2} \]

(a) \hspace{5cm} (b)

• Both grid-point method and vanilla Monte Carlo methods wasted resources on “boring” desert area.
The basic idea

• Marshall (1956) suggested that one should focus on the region(s) of “importance” so as to save computational resources—*importance sampling*.

• Essential in high-dimensional models.
The algorithm

- To evaluate $\mu = E_\pi[h(X)] = \int h(x)\pi(x)dx$.
  - Draw $x^{(1)},\ldots,x^{(m)}$ from a trial distribution $g()$.
  - Calculate the importance weight $w^{(j)} = \pi(x^{(j)})/g(x^{(j)})$, for $j = 1,\ldots,m$.
  - Approximate $\mu$ by $\hat{\mu} = \frac{w^{(1)}h(x^{(1)}) + \cdots + w^{(m)}h(x^{(m)})}{w^{(1)} + \cdots + w^{(m)}}$.

- Remark: $\hat{\mu}$ is better than the unbiased estimator $\tilde{\mu} = \frac{1}{m}\{w^{(1)}h(x^{(1)}) + \cdots + w^{(m)}h(x^{(m)})\}$. why?
An example (cont.)

- Use proposal function

\[ g(x, y) \propto 0.5e^{-90(x-0.5)^2-10(y+0.1)^2} + e^{-45(x+0.4)^2-60(y-0.5)^2}, \]

with \((x,y) \in [-1,1] \times [-1,1]\), a truncated mixture of bivariate Gaussian

\[
0.46\mathcal{N}\left[\left( \begin{array}{c} 0.5 \\ -0.1 \end{array} \right), \left( \begin{array}{cc} 0.625 & 0 \\ 0 & 0.25 \end{array} \right)\right] + 0.54\mathcal{N}\left[\left( \begin{array}{c} -0.4 \\ 0.5 \end{array} \right), \left( \begin{array}{cc} 0.625 & 0 \\ 0 & 0.25 \end{array} \right)\right]
\]

Vanilla Monte Carlo

\[ \hat{\mu} = 0.1307 \]
\[ \text{std}(\hat{\mu}) = 0.009 \]

Importance Sampling

\[ \hat{\mu} = 0.1259 \]
\[ \text{std}(\hat{\mu}) = 0.0005 \]
Rao-Blackwellization

• Basic principle in Monte Carlo:
  carry out analytical computation as much as possible.
Rao-Blackwellization

- Estimating $E[h(x)]$.

  draw independent samples: $x^{(1)}, x^{(2)}, \ldots$

  $$\hat{I} = \frac{1}{m} \left\{ h(x^{(1)}) + \cdots + h(x^{(m)}) \right\}.$$  

- If $X = (x_1, x_2)$ and $E[h(x)|x_2]$ can be carried out analytically.

  $$\tilde{I} = \frac{1}{m} \left\{ E[h(x) \mid x_2^{(1)}] + \cdots + E[h(x) \mid x_2^{(m)}] \right\}.$$
Rao-Blackwellization

- Both $\hat{I}$ and $\tilde{I}$ are unbiased.

\[ E_\pi h(x) = E_\pi [E\{h(x) \mid x_2\}] . \]

- But $\tilde{I}$ should be preferred because

\[
\text{var}(\hat{I}) = \frac{\text{var}\{h(x)\}}{m} \geq \frac{\text{var}\{E[h(x) \mid x_2]\}}{m} = \text{var}(\tilde{I}) .
\]

Due to

\[
\text{var}\{h(x)\} = \text{var}\{E[h(x) \mid x_2]\} + E\{\text{var}[h(x) \mid x_2]\} ,
\]
Rao-Blackwellization

• Conditioning an inferior estimator on the value of sufficient statistics leads to the optimal estimator.
Sequential importance sampling

• For high dimensional problem, how to design trial distribution is challenging.
• Suppose the target density of $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ can be decomposed as

$$
\pi(\mathbf{x}) = \pi(x_1)\pi(x_2 \mid x_1) \cdots \pi(x_d \mid x_1, \ldots, x_{d-1})
$$

then constructed trial density as

$$
g(\mathbf{x}) = g_1(x_1)g_2(x_2 \mid x_1) \cdots g_d(x_d \mid x_1, \ldots, x_{d-1})
$$
Sequential importance sampling

\[ w(x) = \frac{\pi(x_1)\pi(x_2 \mid x_1) \cdots \pi(x_d \mid x_1, \ldots, x_{d-1})}{g_1(x_1)g_2(x_2 \mid x_1) \cdots g_d(x_d \mid x_1, \ldots, x_{d-1})} \]

Suggest a recursive way of computing and monitoring importance weight. Denote

\[ x_t = (x_1, x_2, \ldots, x_t) \]

then we have

\[ w_t(x_t) = w_{t-1}(x_{t-1}) \frac{\pi(x_t \mid x_{t-1})}{g_t(x_t \mid x_{t-1})} \]
Sequential importance sampling

• Advantages of the recursion scheme
  – Can stop generating further components of \( x \) if the partial weight is too small.
  – Can take advantage of \( \pi(x_t \mid x_{t-1}) \) in designing \( g_t(x_t \mid x_{t-1}) \)

• However, the scheme is impractical since requires the knowledge of marginal distribution \( \pi(x_t) \).
Sequential importance sampling

• Add another layer of complexity:
• Introduce a sequence of “auxiliary distributions” \( \pi_1(x_1)\pi_2(x_2)\pi_d(x) \) such that \( \pi_t(x_t) \) is a reasonable approximation of the marginal distribution \( \pi(x_t) \), for \( t = 1, \ldots, d - 1 \) and \( \pi_d = \pi \).
• Note the \( \pi_d \) are only required to be known up to a normalizing constant.
The SIS procedure

For $t = 2, \ldots, d$,

- Draw $X_t = x_t$ from $g_t (x_t \mid x_{t-1})$, and let $x_t = (x_{t-1}, x_t)$
- Compute $u_t = \pi_t (x_t) \pi_{t-1} (x_{t-1}) g_t (x_t \mid x_{t-1})$
  and let $w_t = w_{t-1} u_t$
- $u_t$: incremental weight.
- The key idea is to breaks a difficult task into manageable pieces.
- If $w_t$ is getting too small, reject.
References

• Hammersley and Morton (1954).
• Rosenbluth and Rosenbluth (1955).
• Liu JS (2001)
• Doucet et al. (2001).
Examples of SIS

• Growing a polymer
  – Self avoid walk
• Sequential imputation for statistical missing data problem.
• More and details of these examples, see Liu 2001.
Future topics

• Multigrid Monte Carlo (MGMC), density-scaling Monte Carlo, hybrid Monte Carlo (HMC), evolutionary Monte Carlo, exchange Monte Carlo.

• Cluster method, data augmentation. Parameter expansion, multicanonical sampling, umbrella sampling, simulated tempering, multi-try Metropolis, particle filtering, …