Lecture Notes for STAT546: Computational Statistics

—Lecture 9: Monte Carlo

Faming Liang

Purdue University

September 11, 2024

Monte Carlo Dynamically Weighted Importance Sampling

The MCDWIS algorithm is a Monte Carlo version of the DWIS algorithm. As in DWIS, the state space of the Markov chain is augmented to a population, a collection of weighted samples $(\boldsymbol{x}, \boldsymbol{w}) = \{x_1, w_1; \ldots; x_n, w_n\}$. Given the current population $(\boldsymbol{x}_t, \boldsymbol{w}_t)$, one iteration of the MCDWIS consists of two steps:

1. *Monte Carlo Dynamic weighting (MCDW):* Update each individual state of the current population by a MCDW transition.
2. *Population control:* Split or replicate the individual states with large weights and discard the individual states with small weights.

The MCDW step allows for the use of Monte Carlo estimates in MCMC simulations. The bias induced thereby is counterbalanced by giving different weights to the new samples produced. Therefore, a Monte Carlo estimate of $Z(x)/Z(x')$ can be incorporated into the simulation, while leaving $\pi(x|D)$ invariant with respect to dynamic importance weights. Note that conventional MCMC algorithms do not allow for the use of Monte Carlo estimates in simulations. Otherwise, the detailed balance condition will be violated.

A MCDW sampler

Let $(\boldsymbol{x}_t, \boldsymbol{w}_t)$ denote the current population, let (x, w) denote an individual state of the population, and let (x', w') denote the individual state transmitted from (x, w) in one transition step.

1. Draw x^* from some proposal distribution $T(x, x^*)$.
2. Simulate auxiliary samples $D_1, \ldots, D_m$ from $f(D|x^*)$ using a MCMC algorithm, say, the MH algorithm. Estimate the normalizing constant ratio $R_t(x, x^*) = Z(x)/Z(x^*)$ by

$$\widehat{R}_t(x, x^*) = \frac{1}{m} \sum_{i=1}^{m} \frac{p(D_i, x)}{p(D_i, x^*)}, \tag{1}$$

   which is also known as the importance sampling (IS) estimator of $R_t(x, x^*)$.

3. Calculate the Monte Carlo dynamic weighting ratio

$$r_d = r_d(x, x^*, w) = w\widehat{R}_t(x, x^*)\frac{p(D, x^*)}{p(D, x)}\frac{T(x^*, x)}{T(x, x^*)}.$$

4. Choose $\theta_t = \theta_t(\boldsymbol{x}_t, \boldsymbol{w}_t) \geq 0$ and draw $U \sim unif(0, 1)$. Update (x, w) as (x', w')

$$(x', w') = \begin{cases} (x^*, r_d/a), & \text{if } U \leq a, \\ (x, w/(1-a)), & \text{otherwise,} \end{cases}$$

where $a = r_d/(r_d + \theta_t)$; θ_t is a function of $(\boldsymbol{x}_t, \boldsymbol{w}_t)$, but remains a constant for each individual state of the same population.

Remark 1:

$\widehat{R}_t(x, x^*)$ is an unbiased and consistent estimator of $R_t(x, x^*)$. Following from the central limit theorem, we have

$$\sqrt{m}\left(\widehat{R}_t(x, x^*) - R_t(x, x^*)\right) \to N\left(0, \sigma_t^2\right), \tag{2}$$

where σ_t^2 can be expressed as

$$\sigma_t^2 = \text{Var}\left(\frac{p(D_1, x)}{p(D_1, x^*)}\right) + 2\sum_{i=2}^{\infty} \text{Cov}\left(\frac{p(D_1, x)}{p(D_1, x^*)}, \frac{p(D_i, x)}{p(D_i, x^*)}\right).$$

Alternatively, we can write $\widehat{R}_t(x, x^*) = R_t(x, x^*)(1 + \epsilon_t)$, where $\epsilon_t \sim N(0,\ \sigma_t^2\ /\ [mR_t^2(x, x^*)])$.

Remark 2

This sampler is designed according to the scheme-R of DWIS. A similar sampler can also be deigned according to the scheme-W of DWIS. As discussed previously, the parameter θ_t can be specified as a function of the population $(\boldsymbol{x}_t, \boldsymbol{w}_t)$. For simplicity, we here concentrate only on the cases where $\theta_t = 0$ or 1.

Theorem 1

The Monte Carlo dynamic weighting sampler is $IWIW_p$; that is, if the joint distribution $g_t(x, w)$ for $(\boldsymbol{x}_t, \boldsymbol{w}_t)$ is correctly weighted with respect to $\pi(x|D)$, after one Monte Carlo dynamic weighting step, the new joint density $g_{t+1}(x', w')$ for $(\boldsymbol{x}_{t+1}, \boldsymbol{w}_{t+1})$ is also correctly weighted with respect to $\pi(x|D)$.

For the case $\theta_t > 0$,

$$\int_0^\infty w' g_{t+1}(x', w') dw'$$

$$= \int_{\mathcal{X}} \int_0^\infty \int_{-\infty}^\infty [r_d(x, x', w) + \theta_t] g_t(x, w) T(x, x') \varphi(\epsilon_t) \frac{r_d(x, x', w)}{r_d(x, x', w) + \theta_t} d$$

$$+ \int_{\mathcal{X}} \int_0^\infty \int_{-\infty}^\infty \frac{w[r_d(x', z, w) + \theta_t]}{\theta_t} g_t(x', w) T(z|x') \varphi(\epsilon_t) \frac{\theta_t}{r_d(x', z, w) + \theta}$$

$$= \int_{\mathcal{X}} \int_0^\infty \int_{-\infty}^\infty wR(x, x')(1 + \epsilon_t) \frac{p(D, x')\pi(x')}{p(D, x)\pi(x)} T(x', x) g_t(x, w) \varphi(\epsilon_t) d\epsilon_t d$$

$$+ \int_{\mathcal{X}} \int_0^\infty \int_{-\infty}^\infty w g_t(x', w) T(x', z) \varphi(\epsilon_t) d\epsilon_t dw dz$$

$$
\begin{aligned}
&= \int_{\mathcal{X}} \int_0^{\infty} wR(x, x') \frac{p(D, x')\pi(x')}{p(D, x)\pi(x)} T(x', x) g_t(x, w) dw dx \\
&+ \int_{\mathcal{X}} \int_0^{\infty} w g_t(x', w) T(x', z) dw dz \\
&= \int_{\mathcal{X}} \frac{\pi(x'|D)}{\pi(x|D)} T(x', x) \Big(\int_0^{\infty} w g_t(x, w) dw \Big) dx + \int_{\mathcal{X}} c_{t,x'} \pi(x'|D) T(x', z) d \\
&= \pi(x'|D) \int_{\mathcal{X}} c_{t,x} T(x', x) dx + c_{t,x'} \pi(x'|D) \\
&= \pi(x'|D) \int_{\mathcal{X}} c_{t,x} T(x', x) dx + c_{t,x'} \pi(x'|D) \\
&= 2 c_{t,x'} \pi(x'|D),
\end{aligned}
$$

where $\varphi(\cdot)$ denotes the density of ϵ_t.

For the case $\theta_t = 0$, only the term (I) remains. Thus,

$$\int_0^\infty w' g_{t+1}(x', w') dw' = c_{t,x'} \pi(x'|D).$$

By defining $c_{t+1,x'} = 2c_{t,x'}$ for the case $\theta_t > 0$ and $c_{t+1,x'} = c_{t,x'}$ for the case $\theta_t = 0$, it is easy to see that the condition (**??**) still holds for the new population. Hence, $g_{t+1}(x', w')$ is still correctly weighted with respect to $\pi(x|D)$.

Monte Carlo Dynamically Weighted Importance Sampling

Similar to the DWIS, we compose the following MCDWIS algorithm, which alters between the MCDW and population control steps. Since both the MCDW step and the population control step are IWIW$_p$, the MCDWIS algorithm is also IWIW$_p$. Let W_c denote a dynamic weighting move switching parameter, which switches the value of θ_t between 0 and 1 depending on the value of $W_{up,t-1}$. One iteration of the MCDWIS algorithm can be described as follows.

- *MCDW*: Apply the Monte Carlo dynamic weighting move to the population $(\boldsymbol{x}_{t-1}, \boldsymbol{w}_{t-1})$. If $W_{up,t-1} \leq W_c$, then set $\theta_t = 1$. Otherwise, set $\theta_t = 0$. The new population is denoted by $(\boldsymbol{x}'_t, \boldsymbol{w}'_t)$.
- *Population Control*: Apply APEPCS to $(\boldsymbol{x}'_t, \boldsymbol{w}'_t)$. The new population is denoted by $(\boldsymbol{x}_t, \boldsymbol{w}_t)$.

To analyze the weight behavior of the MCDWIS, we introduce the following lemma.

Lemma 2

Let $f(D|x) = p(D, x)/Z(x)$ denote the likelihood function of D, let $\pi(x)$ denote the prior distribution of x, and let $T(\cdot, \cdot)$ denote a proposal distribution of x. Define
$p(x, x'|D) = p(D, x)\pi(x)T(x, x')$, *and*
$r(x, x') = \widehat{R}(x, x')p(x', x|D)\ /p(x, x'|D)$ *to be a Monte Carlo MH ratio, where $\widehat{R}(x, x')$ denotes an unbiased estimator of $Z(x)/Z(x')$. Then*

$$e_0 = E \log r(x, x') \leq 0,$$

where the expectation is taken with respect to the joint density $\varphi(\widehat{R}) \times p(x, x'|D)/Z(x)$.

Proof.

By Jensen's inequality,

$$e_0 = E \log \left[\widehat{R}(x, x') \frac{p(x', x|D)}{p(x, x'|D)} \right] \leq \log E \left[\widehat{R}(x, x') \frac{p(x', x|D)}{p(x, x'|D)} \right] = 0,$$

where the equality holds when $p(x', x|D) = p(x, x'|D)$, and $\varphi(\cdot)$ is a Dirac measure with $\varphi(\widehat{R} = R) = 1$ and 0 otherwise. □

Given this lemma, it is easy to see that, theoretically, MCDWIS shares the same weight behavior with scheme-R of DWIS; that is, the following theorem holds for MCDWIS.

Theorem 3

MCDWIS has almost surely finite moments of any order.

Sequentially dynamically weighed importance sampling

Although DWIS has significantly improved the mixing of the MH algorithm, it may still have a hard time in simulating from a system for which the attraction basin of the global minimum energy solution is very narrow. One approach to alleviate this difficulty is to use DWIS in conjunction with a complexity ladder by simulating from a sequence of systems with gradually flattened energy landscapes. The resulting approach is called sequentially dynamically weighted importance sampling (SDWIS).
Suppose that one wants to simulate from a distribution $f(x)$, and a sequence of trial distributions $f_1, \cdots, f_k$ has been constructed with $f_k \equiv f$. For example, $f_{k-1}(x_{k-1})$ can be set as a marginal distribution of $f_k(x_k)$. See §**??** for discussions on how to construct such a complexity ladder of distributions.

1.

- *Sample*: Sample $x_{1,1}, \ldots, x_{1,N_1}$ from $f_1(\cdot)$ using a MCMC algorithm, and set $w_{1,i} = 1$ for $i = 1, \ldots, N_1$. The samples form a population $(\boldsymbol{x}_1, \boldsymbol{w}_1) = (x_{1,1}, w_{1,1}; \ldots; x_{1,N_1}, w_{1,N_1})$.
- *DWIS*: Generate $(\boldsymbol{x}'_1, \boldsymbol{w}'_1)$ from $(\boldsymbol{x}_1, \boldsymbol{w}_1)$ using DWIS with $f_1(x)$ working as the target distribution.

2.

- *Extrapolation*: Generate $x_{2,i}$ from $x'_{1,i}$ with the extrapolation operator $T_{12}(x'_{1,i}, x_{2,i})$, and set

$$w_{2,i} = w'_{1,i} \frac{f_2(x_{2,i})}{f_1(x'_{1,i}) T_{12}(x'_{1,i}, x_{2,i})},$$

for each $i = 1, 2, \cdots, N'_1$.

- *DWIS*: Generate $(\boldsymbol{x}_2', \boldsymbol{w}_2')$ from $(\boldsymbol{w}_2, \boldsymbol{w}_2)$ using DWIS with $f_2(x)$ working as the target distribution.

.....

k.

- *Extrapolation*: Generate $x_{k,i}$ from $x'_{k-1,i}$ with the extrapolation operator $T_{k-1,k}(x'_{k-1,i}, x_{k,i})$, and set

$$w_{k,i} = w'_{k-1,i} \frac{f_k(x_{k,i})}{f_{k-1}(x'_{k-1,i}) T_{k-1,k}(x'_{k-1,i}, x_{k,i})},$$

  for $i = 1, 2, \cdots, N'_{k-1}$.
- *DWIS*: Generate $(\boldsymbol{x}_k', \boldsymbol{w}_k')$ from $(\boldsymbol{x}_k, \boldsymbol{w}_k)$ using DWIS with $f_k(x)$ working as the target distribution.

The advantage of SDWIS over conventional sequential importance sampling (SIS) algorithms is apparent: The DWIS steps will remove the bad seed samples at the early stages of sampling and force the good seed samples to produce more offspring. More importantly, SDWIS overcomes the sample degeneracy problem suffered by conventional sequential importance sampling or particle filter algorithms by including the DWIS step which maintains the diversity of the population. Together with a radial basis function network, SDWIS has been successfully applied to modeling of the sea surface temperatures by Ryu, Liang and Mallick (2009). In a numerical example, Ryu, Liang and Mallick (2009) showed that SDWIS can be more efficient than the standard SIS and the partial rejection control SIS (Liu, Chen and Wong, 1998) algorithms, and SDWIS indeed avoids the sample degeneracy problem.

The framework of SDWIS is so general that it has included several other algorithms as special cases. If only the population control scheme is performed at the DWIS step, SDWIS is reduced to the pruned-enriched Rosenbluth method (Grassberger, 1997). If only some MH or Gibbs steps are performed at the DWIS step (with $f_k(x)$ being chosen as a power function of $f(x)$), SDWIS is reduced to annealed importance sampling (Neal, 2001). Note that the MH or Gibbs step will not alter the correctly weightedness of a population, as they are IWIW. With this observation, MacEachern, Clyde and Liu (1998) proposed to improve the performance of sequential importance sampling by mixing with some MCMC steps.