Bayesian Theory and Computation

Lecture 9: Sequential Monte Carlo

Cheng Zhang

School of Mathematical Sciences, Peking University

April 01, 2022

- It is nontrivial to design a good trial distribution for doing importance sampling in high-dimensional problems.
- One of the most useful strategies in these problems is to build up the importance distribution sequentially.
- Suppose we can decompose $x = (x_1, \ldots, x_d)$ where each of the x_j may be multidimensional. Then, the importance distribution can be constructed as follows

$$q(x) = q_1(x_1)q_2(x_2|x_1)\cdots q_d(x_d|x_1, \ldots, x_{d-1})$$

- We could, of course, decompose the target density accordingly

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

- The importance weight is

$$w(x) = \frac{\pi(x_1)\pi(x_2|x_1)\cdots\pi(x_d|x_1,\ldots,x_{d-1})}{q_1(x_1)q_2(x_2|x_1)\cdots q_d(x_d|x_1,\ldots,x_{d-1})}$$

- This suggests a recursive way of computing and monitoring the importance weight

$$w_t(x_{\le t}) = w_{t-1}(x_{\le t-1})\frac{\pi(x_t|x_1,\ldots,x_{t-1})}{q_t(x_t|x_1,\ldots,x_{t-1})}$$

- Ideally, this approach provides guidance in designing $q_t(x_t|x_{<t})$ and allows automatic termination of the generating process by monitoring the partial weight $w_{\le t}$
- However, the conditional distribution $\pi(x_t|x_{<t})$ is unknown. Estimation could be as difficult as, or even harder than, the original problem.

- One remedy for this issue is to introduce a sequence of auxiliary distributions

$$\pi_1(x_1),\ \pi_2(x_1, x_2),\ \ldots,\ \pi_d(x_1, x_2, \ldots, x_d)$$

  where $\pi_t(x_{\leq t})$ is a reasonable approximation to the marginal distribution $\pi(x_{\leq t})$, $\forall 1 \leq t \leq d-1$ and $\pi_d = \pi$.
- Note that π_t can be unnormalized and they only serve as "guides" to our construction of the whole sample $x = (x_1, \ldots, x_d)$.
- The sequential importance sampling (SIS) method can then be defined as follows

- Let N be the number of particles used for importance sampling.
- At time 1, assume we have approximate $\pi_1(x_1)$ and Z_1 using importance distribution $q_1(x_1)$; that is

$$\hat{\pi}_1(dx_1) = \sum_{i=1}^{N} W_1^{(i)} \delta_{X_1^{(i)}}(dx) \text{ where } W_1^{(i)} \propto w_1(X_1^{(i)})$$

$$\hat{Z}_1 = \frac{1}{N} \sum_{i=1}^{N} w_1(X_1^{(i)})$$

with

$$w_1(x_1) = \frac{\pi_1(x_1)}{q_1(x_1)}$$

- At time 2, we want to approximate $\pi_2(x_1, x_2)$ and Z_2 using an importance distribution $q_2(x_1, x_2)$.
- Since $q_2(x_1, x_2) = q_1(x_1)q_2(x_2|x_1)$, we can reuse the samples $X_1^{(i)}$ from $q_1(x_1)$ to build the IS approximation of $\pi_1(x_1)$. This only makes sense if $\pi_2(x_1) \approx \pi_1(x_1)$.
- Now we sample $X_2^{(i)}|X_1^{(i)} \sim q_2(x_2|X_1^{(i)})$ to obtain $(X_1^{(i)}, X_2^{(i)}) \sim q_2(x_1, x_2)$, and the importance weights are

$$w_2(x_1, x_2) = \frac{\pi_2(x_1, x_2)}{q_2(x_1, x_2)} = \frac{\pi_1(x_1)}{q_1(x_1)} \frac{\pi_2(x_1, x_2)}{\pi_1(x_1)q_2(x_2|x_1)}$$

$$= w_1(x_1) \frac{\pi_2(x_1, x_2)}{\pi_1(x_1)q_2(x_2|x_1)}$$

- For the normalized weights

$$W_2^{(i)} \propto W_1^{(i)} \frac{\pi_2(X_1^{(i)}, X_2^{(i)})}{\pi_1(X_1^{(i)})q_2(X_2^{(i)}|X_1^{(i)})}$$

- Generally speaking, at time t, since

$$\begin{aligned} q_t(x_{\leq t}) &= q_1(x_1)q_2(x_2|x_1)\cdots q_t(x_t|x_{\leq t-1}) \\ &= q_{t-1}(x_{\leq t-1})q_t(x_t|x_{\leq t-1}) \end{aligned}$$

  we can sample $X_t^{(i)} \sim q_t(x_t|X_{\leq t-1}^{(i)})$ to obtain samples from $q_t(x_{\leq t})$.
- The importance weights are updated according to

$$w_t(x_{\leq t}) = \frac{\pi_t(x_{\leq t})}{q_t(x_{\leq t})} = w_{t-1}(x_{\leq t-1})\frac{\pi_t(x_{\leq t})}{\pi_{t-1}(x_{\leq t-1})q_t(x_t|x_{\leq t-1})}$$

- At time $t = 1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1(X_1^{(i)}) = \frac{\pi_1(X_1^{(i)})}{q_1(X_1^{(i)})}$.
- At time $t \geq 2$
  - sample $X_t^{(i)} \sim q_t(\cdot | X_{\leq t-1}^{(i)})$
  - compute $w_t(X_{\leq t}^{(i)}) = w_{t-1}(X_{\leq t-1}^{(i)}) \frac{\pi_t(X_{\leq t}^{(i)})}{\pi_{t-1}(X_{\leq t-1}^{(i)}) q_t(X_t^{(i)} | X_{\leq t-1}^{(i)})}$.
- At any time t, we have

$$X_{\leq t}^{(i)} \sim q_t(X_{\leq t}), \; w_n(X_{\leq t}^{(i)}) = \frac{\pi_t(X_{\leq t}^{(i)})}{q_t(X_{\leq t}^{(i)})}$$

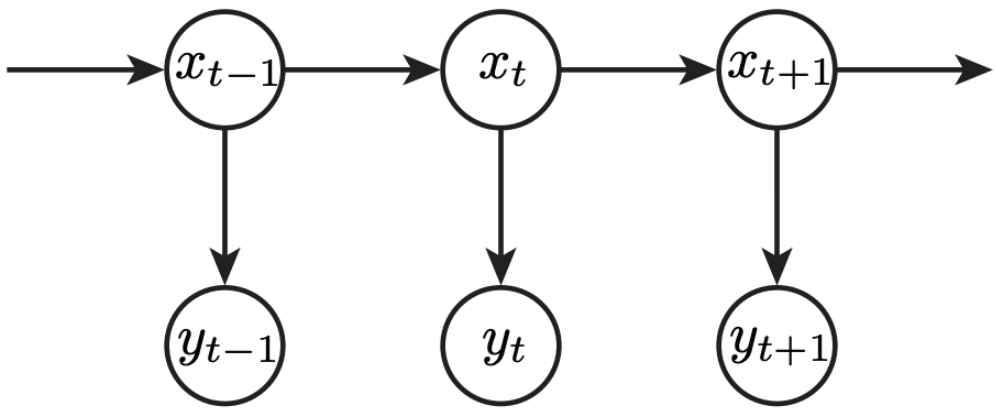
thus we can obtain easily an IS approximation of $\pi_d(x_1, \ldots, x_d)$ and of Z_d.

The Markovian, nonlinear, non-Gaussian state space model

- Unobserved signal or states $\{x_t | t \in \mathbb{N}\}$
- Observations or output $\{y_t | t \in \mathbb{N}^+\}$ or $\{y_t | t \in \mathbb{N}\}$

with the following probabilities

- $p(x_0)$ – initial distribution
- $p(x_t|x_{t-1}),\ t \geq 1$ – transition probability
- $p(y_t|x_t),\ t \geq 0$ – emission/observation probability



- Assume we receive $y_{\leq n}$, we are interested in the posterior distribution of the unobserved signal

$$p(x_{\leq n}|y_{\leq n}) = \frac{p(x_{\leq n}, y_{\leq n})}{p(y_{\leq n})}$$

and estimating $p(y_{\leq n})$ where

$$\pi(x_{\leq n}) = p(x_{\leq n}, y_{\leq n}) = p(x_0) \prod_{t=1}^{n} p(x_t|x_{t-1}) \prod_{t=0}^{n} p(y_t|x_t)$$

$$Z_n = p(y_{\leq n}) = \int \cdots \int p(x_0) \prod_{t=1}^{n} p(x_t|x_{t-1}) \prod_{t=0}^{n} p(y_t|x_t) dx_{0:n}$$

How About MCMC or Importance Sampling?

MCMC might not be well-suited for recursive estimation problems

- It can be hard to design a good proposal q
- What happens if we get a new data point y_{n+1}?
  - We cannot (directly) reuse the samples $\{x_{\leq n}^{(i)}\}$
  - We have to run a new MCMC simulations for $p(x_{\leq n+1}|y_{\leq n+1})$

Importance sampling has similar difficulties

- Designing a good importance distribution can be hard
- When seeing new data y_{n+1}, we cannot reuse the samples and weights for time n

$$\{x_{\leq n}^{(i)}, w_n^{(i)}\}_{i=1}^N$$

to sample from $p(x_{\leq n+1}|y_{\leq n+1})$

- Assume that the importance distribution can be factored as

$$\begin{aligned} q(x_{\leq n}|y_{\leq n}) &= q(x_{\leq n-1}|y_{\leq n-1}) \cdot q(x_n|x_{\leq n-1}, y_{\leq n}) \\ &= q(x_0|y_0) \prod_{t=1}^{n} q(x_t|x_{\leq t-1}, y_{\leq t}) \end{aligned}$$

- The importance weight can then be evaluated recursively

$$W_t^{(i)} \propto W_{t-1}^{(i)} \frac{p(X_t^{(i)}|X_{t-1}^{(i)})p(y_t|X_t^{(i)})}{q(X_t^{(i)}|X_{\leq t-1}^{(i)}, y_{\leq t})} \tag{1}$$

- Given the past trajectories $\{X_{\leq t-1}^{(i)}|1 \leq i \leq N\}$, we can
  - simulate $X_t^{(i)} \sim q(X_t^{(i)}|X_{\leq t-1}^{(i)}, y_{\leq t})$
  - update the weight $W_t^{(i)}$ for $X_{\leq t}^{(i)}$ based on $w_{t-1}^{(i)}$ using eq. (1)

- Bootstrap: we can select

$$q_0(x_0|y_0) = p(x_0), \quad q_t(x_t|x_{\leq t-1}, y_{\leq t}) = p(x_t|x_{t-1}), 1 \leq t \leq n$$

- At time $t = 1$, sample $X_0^{(i)} \sim p(x_0)$ and set $w_0^{(i)} = p(y_0|X_0^{(i)})$
- At time $t \leq 2$
  - sample $X_t^{(i)} \sim p(\cdot|X_{\leq t-1}^{(i)})$
  - compute $w_t^{(i)} = w_{t-1}^{(i)} p(y_t|X_t^{(i)})$.
- At time $t = n$, we have

$$X_{\leq n}^{(i)} \sim p(x_0) \prod_{t=1}^{n} p(x_t|x_{t-1}), \; w_n^{(i)} = \prod_{t=0}^{n} p(y_t|X_t^{(i)})$$

thus we can obtain easily an IS approximation of $p(x_{\leq n}|y_{\leq n})$ and of $p(y_{\leq n})$.

- The optimal zero-variance density at time n is simply given by

$$q_n(x_{\leq n}) = \tilde{\pi}_n(x_{\leq n}) = \frac{\pi_n(x_{\leq n})}{Z_n}$$

- As we have

$$\tilde{\pi}_n(x_{\leq n}) = \tilde{\pi}_n(x_1)\tilde{\pi}_n(x_2|x_1)\cdots\tilde{\pi}_n(x_n|x_{\leq n-1}),$$

so $\pi_t(x_t|x_{\leq t-1}) \propto \tilde{\pi}_n(x_t|x_{\leq t-1})$. This means that we have

$$q_t^{\text{opt}}(x_t|x_{\leq t-1}) = \tilde{\pi}_n(x_t|x_{\leq t-1})$$

- Obviously this result does depend on n so it is only useful if we are only interested in a specific target $\pi_n(x_{\leq n})$ and in such scenarios we need to typically approximate $\tilde{\pi}_n(x_t|x_{\leq t-1})$ which would be difficult in practice.

Locally Optimal Importance Distribution

- One sensible strategy consists of selecting $q_t(x_t|x_{\leq t-1})$ at time t so as to minimize the variance of the importance weights.
- We have for the importance weight

$$\begin{aligned} w_t(x_{\leq t}) &= \frac{\pi_t(x_{\leq t})}{q_{t-1}(x_{\leq t-1})q_t(x_t|x_{\leq t-1})} \\ &= \frac{Z_t\tilde{\pi}_t(x_{\leq t-1})}{q_{t-1}(x_{\leq t-1})} \cdot \frac{\tilde{\pi}_t(x_t|x_{\leq t-1})}{q_t(x_t|x_{\leq t-1})} \end{aligned}$$

- It follows directly that we have

$$q_t^{\text{opt}}(x_t|x_{\leq t-1}) = \tilde{\pi}_t(x_t|x_{\leq t-1})$$

Locally Optimal Importance Distribution

- The weight updating formula is

$$w_t(x_{\leq t}) = w_{t-1}(x_{\leq t-1}) \frac{\pi_t(x_{\leq t})}{\pi_{t-1}(x_{\leq t-1})\tilde{\pi}_t(x_t|x_{\leq t-1})}$$
$$= w_{t-1}(x_{\leq t-1}) \frac{\pi_t(x_{\leq t-1})}{\pi_{t-1}(x_{\leq t-1})}$$

- This locally optimal importance density will be used again and again.
- It is often impossible to sample directly from $\tilde{\pi}_t(x_t|x_{\leq t-1})$ and/or compute $\pi_t(x_{\leq t-1}) = \int \pi_t(x_{\leq t})dx_t$.
- In such cases, it is necessary to approximate $\tilde{\pi}_t(x_t|x_{\leq t-1})$ and $\pi_t(x_{\leq t-1})$.

- In the case of state space models, we have

$$q_t^{\text{opt}}(x_t|x_{\leq t-1}) = p(x_t|x_{\leq t-1}, y_{\leq t}) = p(x_t|x_{t-1}, y_t)$$
$$= \frac{p(x_t|x_{t-1})p(y_t|x_t)}{p(y_t|x_{t-1})}$$

- In this case,

$$w_t(x_{\leq t}) = w_{t-1}(x_{\leq t-1}) \frac{p(x_{\leq t}, y_{\leq t})}{p(x_{\leq t-1}, y_{\leq t-1})p(x_t|x_{t-1}, y_t)}$$
$$= w_{t-1}(x_{\leq t-1})p(y_t|x_{t-1})$$

Example: Linear Gaussian State Space Model

- Consider the simple model

$$X_t = \alpha X_{t-1} + \xi_t$$
$$Y_t = X_t + \sigma \eta_t$$

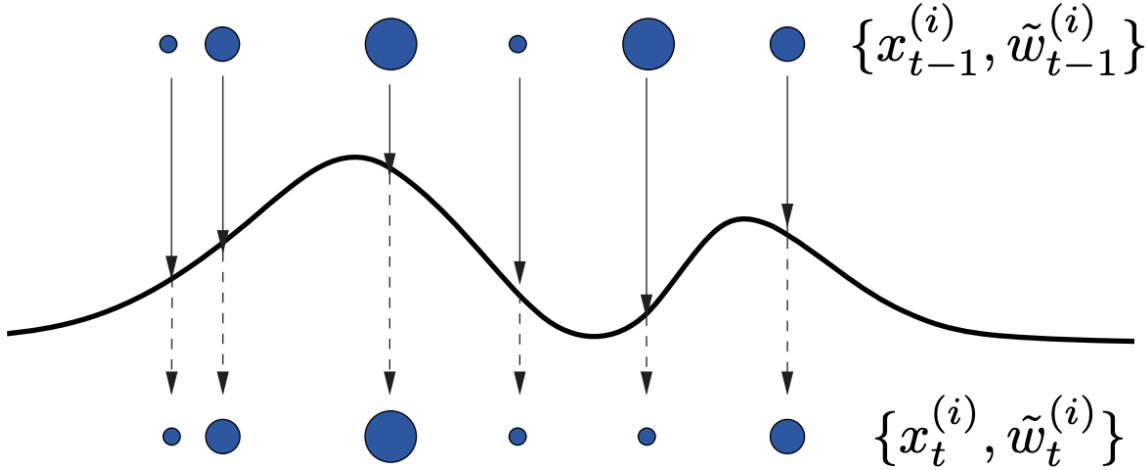
where $X_0 \sim \mathcal{N}(0,1), \xi_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

- Recall that previously we simply used

$$q_t(x_t|x_{\leq t-1}) = p(x_t|x_{t-1}) = \mathcal{N}(\alpha x_{t-1}, 1)$$

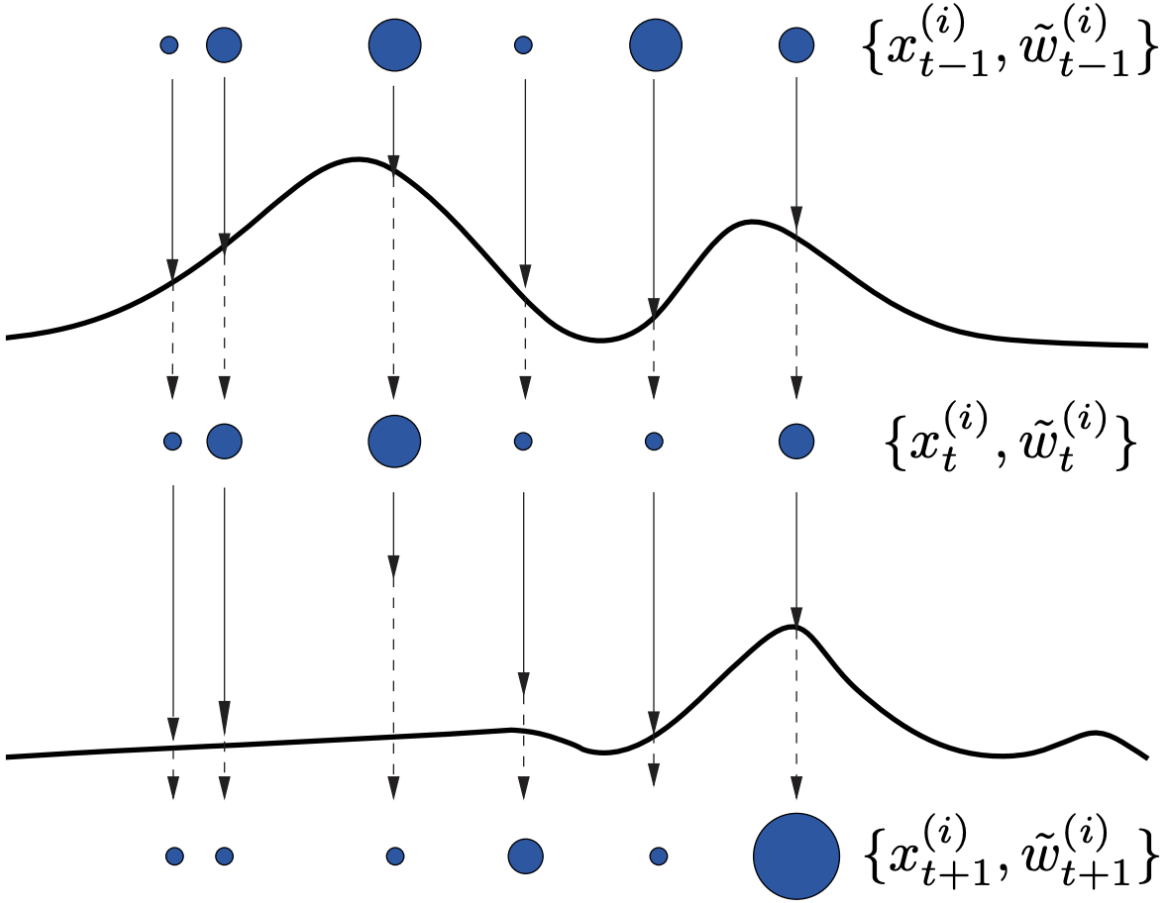
- The locally optimal importance distribution actually is

$$\begin{aligned} q_t^{\text{opt}}(x_t|x_{t-1}) &= p(x_t|x_{t-1}, y_t) \\ &= \mathcal{N}\left(\frac{\sigma^2}{\sigma^2+1}\left(\alpha x_{t-1} + \frac{y_t}{\sigma^2}\right), \frac{\sigma^2}{\sigma^2+1}\right) \end{aligned}$$



Adapted from (Doucet *et al.* , 2001)

Problem solved?



Adapted from (Doucet *et al.* , 2001)

Weights become highly degenerated after few steps

- SIS is an attractive idea: sequential and parallizable, only requires designing low-dimensional proposal distributions.
- SIS can only work for moderate size problems.
- Is there a way to partially fix this problem?

- As time t increases, the variance of the unnormalized weights $\{w_n^{(i)}\}$ tend to increase and all the mass is concentrated on a few particles.
- The key idea to eliminate this weight degeneracy is to get rid of particles with low importance weights and multiply particles with high importance weights.
- The intuition is that if a particle at time t has a low weight then typically it will still have a low weight at time $t+1$ (counterexamples can be easily found though).
- In general, you may want to focus your computation efforts on the more "promising" parts of the space.

- Introduce a resampling each time step (or occasionally)
- At time t, IS provides the following approximation of $\pi_t(x_{\leq t})$

$$\hat{\pi}_t(dx_{\leq t}) = \sum_{i=1}^{N} W_t^{(i)} \delta_{X_{\leq t}^{(i)}}(dx_{\leq t})$$

- The simplest resampling scheme consists of sampling N times $\tilde{X}_{\leq t}^{(i)} \sim \hat{\pi}_t(dx_{\leq t})$ to build the new approximation

$$\tilde{\pi}_t(dx_{\leq t}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\tilde{X}_{\leq t}^{(i)}}(dx_{\leq t}) = \frac{1}{N} \sum_{i=1}^{N} N_t^{(i)} \delta_{X_{\leq t}^{(i)}}(dx_{\leq t})$$

- The weights of the new particles are $\tilde{w}_t^{(i)} = \frac{1}{N}$.

- Note that we can rewrite

$$\tilde{\pi}_t(dx_{\leq t}) = \sum_{i=1}^{N} \frac{N_t^{(i)}}{N} \delta_{X_{\leq t}^{(i)}}(dx_{\leq t})$$

where $(N_t^{(1)}, \ldots, N_t^{(N)}) \sim \text{Multinomial}(N; W_t^{(1)}, \ldots, W_t^{(N)})$.

$$\mathbb{E}N_t^{(i)} = NW_t^{(i)}, \quad \mathbb{V}\text{ar}N_t^{(i)} = NW_t^{(i)}(1 - W_t^{(i)})$$

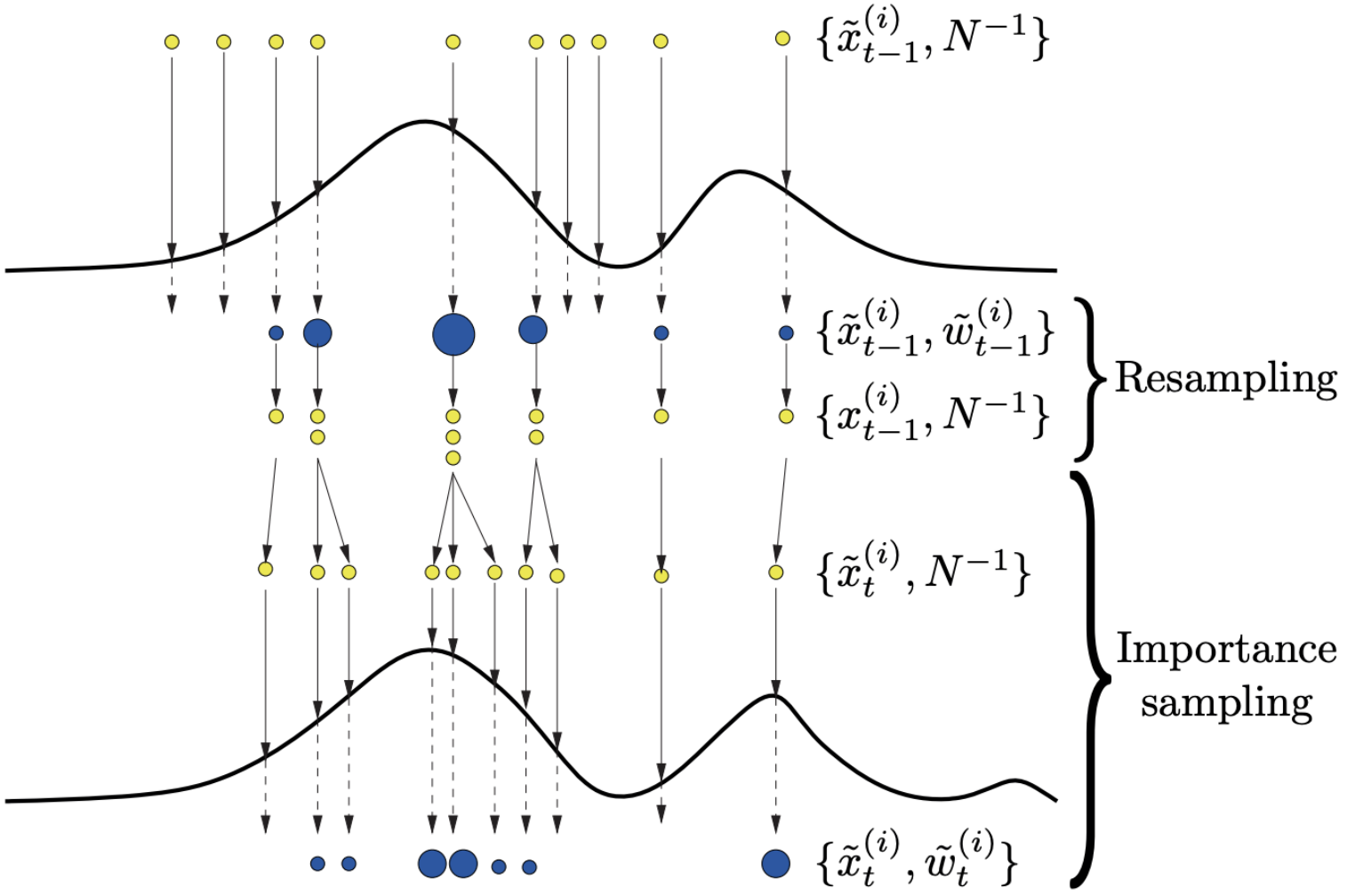
- It follows that the resampling step is an unbiased operation

$$\mathbb{E}(\tilde{\pi}_t(dx_{\leq t})|\hat{\pi}_t(dx_{\leq t})) = \hat{\pi}_t(dx_{\leq t})$$

but clearly it introduces some errors "locally" in time. That is for any test function ϕ, we have

$$\mathbb{V}\text{ar}_{\tilde{\pi}_t}\phi(X_{\leq t}) \geq \mathbb{V}\text{ar}_{\hat{\pi}_t}\phi(X_{\leq t})$$

Sequential Importance Resampling



$i = 1, \ldots, N$ and $N = 10$, figure modified from (Doucet *et al.* , 2001)

- Resampling at each time step is harmful. We should resample only when necessary.
- To measure the variation of the weights, we can use the Effective Sample Size (ESS) or the Coefficient of Variation CV.

$$\text{ESS} = \left(\sum_{i=1}^{N}(W_t^{(i)})^2\right)^{-1}, \quad \text{CV} = \left(\frac{1}{N}\sum_{i=1}^{N}(NW_t^{(i)} - 1)^2\right)^{\frac{1}{2}}$$

- We have $\text{ESS} = N$ and $\text{CV} = 0$ if $W_t^{(i)} = \frac{1}{N}, \forall i$.
- We have $\text{ESS} = 1$ and $\text{CV} = \sqrt{N-1}$ if $W_t^{(i)} = 1$ and $W_t^{(j)} = 0, \forall j \neq i$.
- Dynamic Resampling: if the variation of the weights is too high, then resample the particles.

- The above SMC strategy performs remarkably well in terms of estimation of the marginals $p(x_n|y_{\leq n})$. This is what is only necessary in many applications thankfully.
- However, the joint distribution $p(x_{\leq n}|y_{\leq n})$ is often poorly estimated when n is large.
- This is known as the path degeneracy problem, which is a product of the weight degeneracy problem.
- Since it is necessary to resample the particles, looking backward in time, many of the particles will be exactly the same. Therefore, the approximation of the joint distribution is in terms of a large number of similar paths, which cannot be expected to work well.
- Resampling only partially solve our problems.

- **Theorem**. For any bounded function ψ and any $p > 1$

$$\mathbb{E}\left(\left|\int \psi_t(x_{\leq t})(\hat{\pi}_t(x_{\leq t}) - \pi_t(x_{\leq t}))dx_{\leq t}\right|^p\right)^{1/p} \leq \frac{C_t\|\psi\|_\infty}{\sqrt{N}}$$

- It looks like a nice result but it is rather useless as C_t increases polynomially/exponentially with time.
- To achieve a fixed precision, this would require to use a time-increasing number of particles N.
- For more details, see Del Moral 2004.

- We now describe an SMC technique that is designed to simulate from a sequence of probability densities on a common *state-space*.
- Let us assume that it is of interest to sample from a single probability π, which is 'complex'.
- The approach here is to introduce a sequence of densities. The sequence starts at a very simple distribution and then moves towards π with related distributions interpolating between π and this initial distribution.
- The method is termed SMC samplers (Chopin 2002, Del Moral et al. 2006).

Sequence of Interpolating Distributions

- As an example, consider

$$\pi_t(x) \propto \pi(x)^{\phi_t}, \quad x \in \mathbb{R}^d$$

  with $0 < \phi_1 < \cdots < \phi_p = 1$.
- The idea is to start with a very simple density $\phi_1 = 0$ and then move gradually towards π.
- When $\phi_1 \approx 0$ the target density is 'flat' and should be easy to sample from. Then, by appropriately constructing the densities, it is possible to use the SMC algorithm to interpolate between π_1 and π.
- This idea has been successfully used in many different contexts, such as for rare events estimation, maximum likelihood estimation, as well as approximate Bayesian computation.

- Recall that SMC methods sample from a sequence of densities of increasing dimension.
- Our sequence of densities are on a common space.
- Consider the following idea. Perform IS w.r.t. π_1 via proposal η_1. Then to move to the next density, use a Markov kernel K_2.
- In this scenario, the importance weight is

$$\frac{\pi_2(x_2)}{\eta_2(x_2)}$$

  where $\eta_2(x_2) = \int \eta_1(x_1) K_2(x_1, x_2) dx_1$.
- In most scenarios of interest, one cannot compute this importance weight.

- It turns out that one approach (Del Moral 2006) to circumvent this problem is to introduce a sequence of densities

$$\tilde{\pi}_t(x_{\leq t}) = \pi_t(x_t) \prod_{j=2}^{t} L_j(x_j, x_{j-1})$$

  and use SMC methods on this sequence.
- The $\{L_j\}$ are artificial backward Markov kernels and up-to some minimal technical requirements are essentially arbitrary.
- The algorithm is thus nothing more than SIS. The incremental weights are of the form

$$\frac{\tilde{\pi}_t(x_{\leq t})}{\tilde{\pi}_{t-1}(x_{\leq t-1}) K_t(x_{t-1}, x_t)} = \frac{\pi_t(x_t) L_t(x_t, x_{t-1})}{\pi_{t-1}(x_{t-1}) K_t(x_{t-1}, x_t)}$$

- **Proposition**. The sequence of kernels $\{L_j^{\text{opt}}\}(j = 1, \ldots, t)$ minimizing the variance of the unormalized importance weight $w_t(x_{\leq t})$ is given for any j and t by

$$L_j^{\text{opt}}(x_j, x_{j-1}) = \frac{\eta_{j-1}(x_{j-1})K_j(x_{j-1}, x_j)}{\eta_j(x_j)}$$

and in this case

$$w_t(x_{\leq t}) = \frac{\pi_t(x_t)}{\eta_t(x_t)}$$

- When K_t is an MCMC kernel of invariant distribution π_t, we can use a generic approximation

$$L_t(x_t, x_{t-1}) = \frac{\pi_t(x_{t-1})K_t(x_{t-1}, x_t)}{\pi_t(x_t)}$$

- In this case, we have unnormalized incremental weight

$$\tilde{w}_t(x_{t-1}, x_t) = \frac{\pi_t(x_{t-1})}{\pi_{t-1}(x_{t-1})}$$

- In our previous example, the importance weight can be constructed as

$$W_t^{(i)} \propto W_{t-1}^{(i)} \frac{\pi_t(x_{t-1}^{(i)})}{\pi_{t-1}(x_{t-1}^{(i)})} = W_{t-1}^{(i)} \pi(x_{t-1}^{(i)})^{\phi_t - \phi_{t-1}}$$

- This algorithm can work very well in practice; see e.g. (Neal 2001; Chopin 2002; Del Moral et al, 2006).

References

- Doucet, Arnaud, De Freitas, Nando, & Gordon, Neil. (2001). An introduction to sequential Monte Carlo methods. Springer.
- Chopin, N. (2002). A sequential particle filter for static models. Biometrika, 89, 530-552
- Del Moral, P. (2004). Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with applications. Springer: New York.
- Del Moral, P. Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo Samplers. J. R. Statist. Soc. B, 68, 411-436.
- Neal, R. M. (2001). Annealed Importance Sampling. Statist. Comp., 11, 125-139.