LINKÖPING UNIVERSITY

Sequential Monte Carlo methods

Lecture 3 – Monte Carlo and importance sampling

Johan Alenlöv, Linköping University

2021-08-16

Outline – Lecture 3

Aim: Motivate and introduce the Monte Carlo idea and derive importance sampling.

Outline:

1. Why do we need Monte Carlo?
2. The Monte Carlo idea
3. Importance sampling
4. Ex. joint filtering using importance sampling

Why do we need Monte Carlo methods?

Probabilistic modelling often produce intractable optimization and/or integration problems.

Recall the nonlinear filtering problem or consider the computation of a point estimate via **expectation**, e.g. the conditional mean

$$\widehat{x}_{t \mid t} = \mathbb{E}[X_t \mid y_{1:t}] = \int x_t p(x_t \mid y_{1:t}) \mathrm{d}x_t.$$

Why do we need Monte Carlo methods?

Probabilistic modelling often produce intractable optimization and/or integration problems.

Recall the nonlinear filtering problem or consider the computation of a point estimate via **expectation**, e.g. the conditional mean

$$\widehat{x}_{t \mid t} = \mathbb{E}[X_t \mid y_{1:t}] = \int x_t p(x_t \mid y_{1:t}) \mathrm{d}x_t.$$

Monte Carlo methods are **computational solutions** where the distributions of interest are approximated by a large number of N random samples sometimes called particles.

Common test functions

Hence, Monte Carlo methods can be used to solve integrals like

$$\mathbb{E}[\varphi(X) \mid y_{1:t}] = \int \varphi(x) p(x \mid y_{1:T}) \mathrm{d}x$$

Common test functions $\varphi(x)$ include:

- Conditional mean $\varphi(x) = x$ (previous slide)
- Indicator function $\varphi(x) = I(x > \vartheta)$ for some threshold ϑ, which provides an estimate of tail probabilities (modelling e.g. extreme events).
- Covariances and other higher order moments.
- ...

The Monte Carlo idea (I/II)

Let $X \sim \pi(x)$, where we refer to $\pi(x)$ as the **target density**.

(Very) restrictive assumption: Assume that we have N samples $\{x^i\}_{i=1}^N$ from the target density $\pi(x)$, making up an **empirical approximation**

$$\widehat{\pi}^N(x) = \sum_{i=1}^N \frac{1}{N} \delta_{x^i}(x).$$

The Monte Carlo idea (I/II)

Let $X \sim \pi(x)$, where we refer to $\pi(x)$ as the **target density**.

(Very) restrictive assumption: Assume that we have N samples $\{x^i\}_{i=1}^N$ from the target density $\pi(x)$, making up an **empirical approximation**

$$\widehat{\pi}^N(x) = \sum_{i=1}^N \frac{1}{N} \delta_{x^i}(x).$$

Allows for the following approximation of the integral,

$$\mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x \approx \int \varphi(x) \sum_{i=1}^N \frac{1}{N} \delta_{x^i}(x)\mathrm{d}x = \frac{1}{N} \sum_{i=1}^N \varphi(x^i)$$

$$" \int + \delta \to \sum "$$

The Monte Carlo idea (II/II)

The integral

$$I(\varphi) = \mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x$$

is approximated by

$$\widehat{I}_N(\varphi) = \frac{1}{N}\sum_{i=1}^{N} \varphi(x^i).$$

The Monte Carlo idea (II/II)

The integral

$$I(\varphi) = \mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x$$

is approximated by

$$\widehat{I}_N(\varphi) = \frac{1}{N}\sum_{i=1}^{N} \varphi(x^i).$$

The strong law of large numbers tells us that

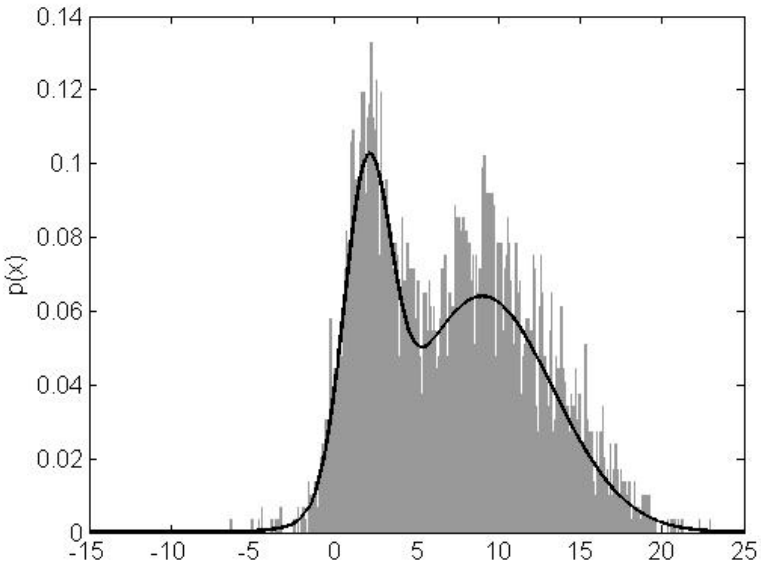
$$\widehat{I}_N(\varphi) \xrightarrow{\text{a.s.}} I(\varphi), \qquad N \to \infty,$$

and the central limit theorem states that

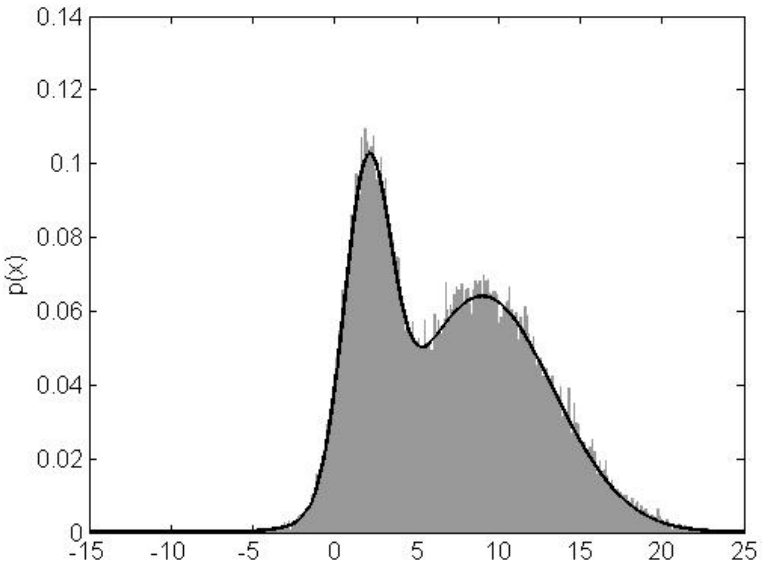
$$\frac{\sqrt{N}\left(\widehat{I}_N(\varphi) - I(\varphi)\right)}{\sigma_\varphi} \xrightarrow{\text{d}} \mathcal{N}(0,1), \qquad N \to \infty.$$

The Monte Carlo idea – toy illustration

$$\pi(x) = 0.3\mathcal{N}(x \mid 2, 2) + 0.7\mathcal{N}(x \mid 9, 19)$$



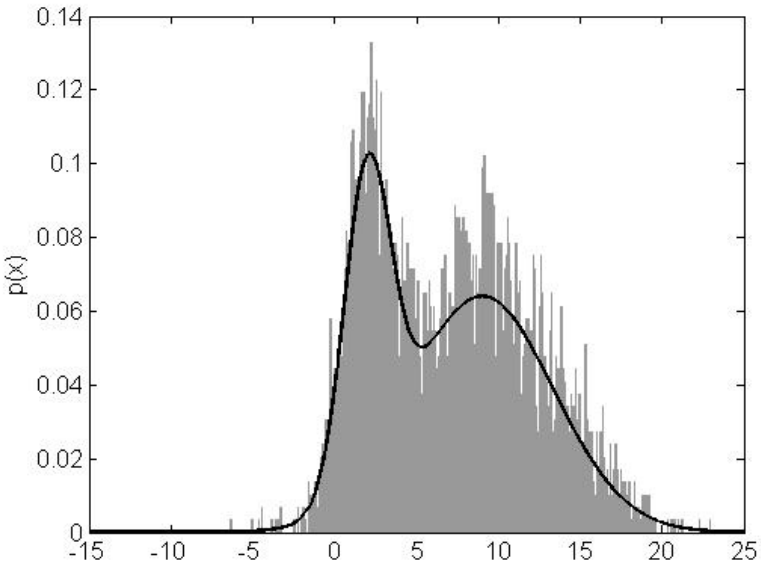
5 000 samples



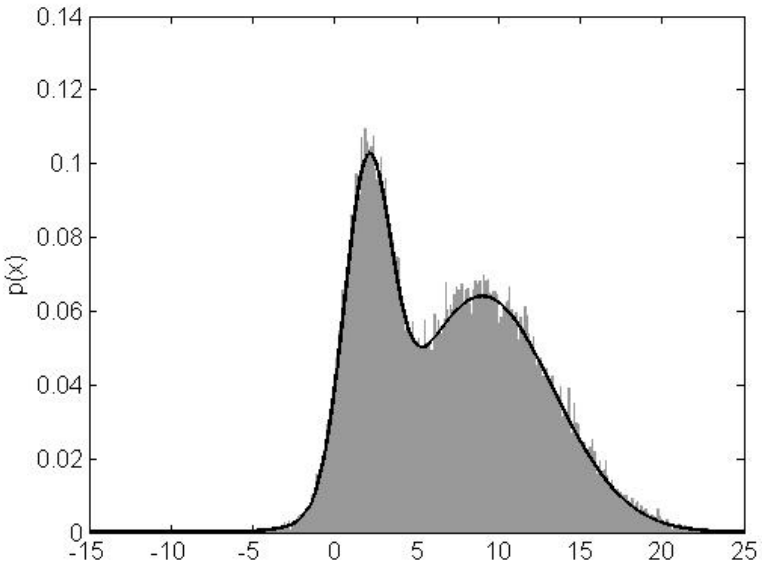
50 000 samples

The Monte Carlo idea – toy illustration

$$\pi(x) = 0.3\mathcal{N}(x \mid 2, 2) + 0.7\mathcal{N}(x \mid 9, 19)$$



5 000 samples



50 000 samples

Obvious problem: In general we are **not** able to directly sample from the density we are interested in.

Importance sampling

Importance sampling

Importance sampling can be used to evaluate integrals of the form

$$I(\varphi) = \mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x \qquad (x \sim \pi(x))$$

without requiring exact samples from the target.

Importance sampling

Importance sampling can be used to evaluate integrals of the form

$$I(\varphi) = \mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x \qquad (x \sim \pi(x))$$

without requiring exact samples from the target.

Note that

$$\mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x = \int \varphi(x)\frac{\pi(x)}{q(x)}q(x)\mathrm{d}x$$

where q is known as the **proposal distribution**.

Importance sampling – proposal distribution

The **proposal distribution**[1] is chosen by the user:

1. It should be simple to sample from and
2. we require $q(x) > 0$ for all x where $\pi(x) > 0$

Idea: Chose the **proposal** density $q(x)$ such that it is easy to generate samples from it and somehow **compensate** for the mismatch between the target and the proposal.

[1] a.k.a. importance distribution or instrumental distribution.

Importance sampling

We have that

$$\mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x = \int \varphi(x)\frac{\pi(x)}{q(x)}q(x)\mathrm{d}x$$

Importance sampling

We have that

$$\mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x = \int \varphi(x)\frac{\pi(x)}{q(x)}q(x)\mathrm{d}x$$

We define the **weight function**

$$\omega(x) = \frac{\pi(x)}{q(x)}$$

Importance sampling

We have that

$$\mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\pi(x)\mathrm{d}x = \int \varphi(x)\frac{\pi(x)}{q(x)}q(x)\mathrm{d}x$$

We define the **weight function**

$$\omega(x) = \frac{\pi(x)}{q(x)}$$

Using this we get that

$$\mathbb{E}_\pi[\varphi(X)] = \int \varphi(x)\omega(x)q(x)\mathrm{d}x = \mathbb{E}_q[\varphi(x')\omega(x')] \qquad (x' \sim q(x'))$$

Importance sampling

Algorithm 1 Importance sampling

1. Sample $x^i \sim q(x)$.
2. Compute the weights $w^i = \omega(x^i) = \pi(x^i)/q(x^i)$.

Each step is performed for $i = 1, \dots N$.

We perform an MC estimator as

$$\widehat{I}_N(\varphi) = \frac{1}{N} \sum_{i=1}^{N} w^i \varphi(x^i).$$

Point-wise evaluation of the target

It is often the case that the target density $\pi(x)$ can only be evaluated up to an unknown normalization constant Z,

$$\pi(x) = \frac{\widetilde{\pi}(x)}{Z}$$

where $\widetilde{\pi}(x)$ can be evaluated for any x, but the constant Z is unknown.

Point-wise evaluation of the target

It is often the case that the target density $\pi(x)$ can only be evaluated up to an unknown normalization constant Z,

$$\pi(x) = \frac{\widetilde{\pi}(x)}{Z}$$

where $\widetilde{\pi}(x)$ can be evaluated for any x, but the constant Z is unknown.

Ex. (nonlinear joint filtering problem) The target density is given by $\pi(x) = p(x_{0:t} \mid y_{1:t})$ and we have

$$\underbrace{p(x_{0:t} \mid y_{1:t})}_{\pi(x)} = \frac{\overbrace{p(x_{0:t}, y_{1:t})}^{\widetilde{\pi}(x)}}{\underbrace{p(y_{1:t})}_{Z}},$$

where we can evaluate $\widetilde{\pi}(x) = p(x_{0:t}, y_{1:t})$ point-wise, but $Z = p(y_{1:t})$ is intractable in general.

Self-normalized importance sampling

Self-normalized importance sampling

Inserting

$$\pi(x) = \frac{\widetilde{\pi}(x)}{Z}$$

into the importance sampling integral results in

$$I(\varphi) = \mathbb{E}[\varphi(X)] = \int \varphi(x) \frac{\widetilde{\pi}(x)}{Z q(x)} q(x) \mathrm{d}x = \frac{1}{Z} \int \varphi(z) \underbrace{\frac{\widetilde{\pi}(x)}{q(x)}}_{=\omega(x)} q(x) \mathrm{d}x$$

Self-normalized importance sampling

Inserting

$$\pi(x) = \frac{\widetilde{\pi}(x)}{Z}$$

into the importance sampling integral results in

$$I(\varphi) = \mathbb{E}[\varphi(X)] = \int \varphi(x) \frac{\widetilde{\pi}(x)}{Z q(x)} q(x) \mathrm{d}x = \frac{1}{Z} \int \varphi(z) \underbrace{\frac{\widetilde{\pi}(x)}{q(x)}}_{=\omega(x)} q(x) \mathrm{d}x$$

Hence, the importance sampling estimator is

$$\widetilde{I}^N(\varphi) = \frac{1}{NZ} \sum_{i=1}^{N} \widetilde{w}^i \varphi(x^i),$$

where $\widetilde{w}^i = \omega(x^i)$.

Self-normalized importance sampling

Inserting

$$\pi(x) = \frac{\widetilde{\pi}(x)}{Z}$$

into the importance sampling integral results in

$$I(\varphi) = \mathbb{E}[\varphi(X)] = \int \varphi(x) \frac{\widetilde{\pi}(x)}{Z q(x)} q(x) \mathrm{d}x = \frac{1}{Z} \int \varphi(z) \underbrace{\frac{\widetilde{\pi}(x)}{q(x)}}_{=\omega(x)} q(x) \mathrm{d}x$$

Hence, the importance sampling estimator is

$$\widetilde{I}^N(\varphi) = \frac{1}{NZ} \sum_{i=1}^{N} \widetilde{w}^i \varphi(x^i),$$

where $\widetilde{w}^i = \omega(x^i)$.

The normalization constant Z is still problematic.

Self-normalized importance sampling

The normalization constant is given by the following integral

$$Z = \int \widetilde{\pi}(x)\mathrm{d}x,$$

Self-normalized importance sampling

The normalization constant is given by the following integral

$$Z = \int \widetilde{\pi}(x)\mathrm{d}x,$$

which we can approximate using our samples $\{x^i\}_{i=1}^N$ from $q(x)$. The result is

$$Z = \int \frac{\widetilde{\pi}(x)}{q(x)} q(x)\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} \widetilde{w}^i.$$

Self-normalized importance sampling

The normalization constant is given by the following integral

$$Z = \int \widetilde{\pi}(x)\mathrm{d}x,$$

which we can approximate using our samples $\{x^i\}_{i=1}^N$ from $q(x)$. The result is

$$Z = \int \frac{\widetilde{\pi}(x)}{q(x)} q(x)\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} \widetilde{w}^i.$$

The **self-normalized** importance sampling estimate is obtained by inserting this into $\widetilde{I}^N(\varphi)$,

$$\widetilde{I}^N(\varphi) = \sum_{i=1}^{N} w^i \varphi(x^i), \qquad w^i = \frac{\widetilde{w}^i}{\sum_{j=1}^{N} \widetilde{w}^j}$$

Self-normalized importance sampling

Algorithm 2 Self-noramlized Importance sampler

1. Sample $x^i \sim q(x)$.
2. Compute the weights $\widetilde{w}^i = \widetilde{\pi}(x^i)/q(x^i)$.
3. Normalize the weights $w^i = \widetilde{w}^i / \sum_{j=1}^N \widetilde{w}^j$.

Each step is carried out for $i = 1, \ldots, N$.

The convergence of the resulting approximation $\widehat{\pi}^N(x) = \sum_{i=1}^N w^i \delta_{x^i}(x)$ is since long well established.

Self-normalized importance sampling

Algorithm 3 Self-noramlized Importance sampler

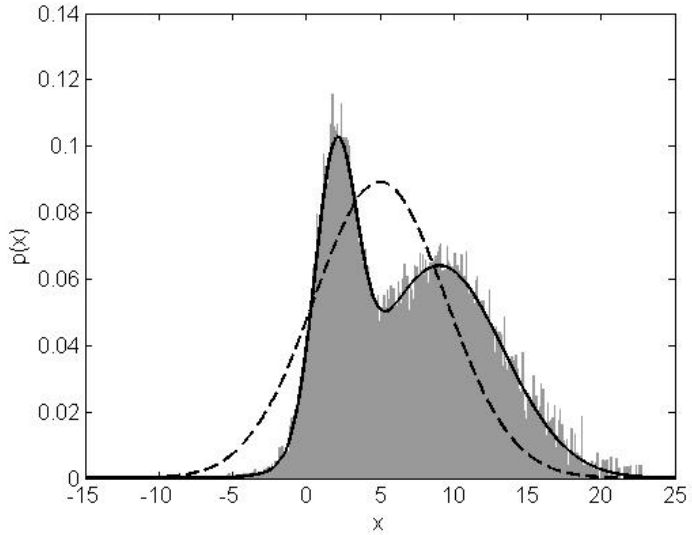
1. Sample $x^i \sim q(x)$.
2. Compute the weights $\widetilde{w}^i = \widetilde{\pi}(x^i)/q(x^i)$.
3. Normalize the weights $w^i = \widetilde{w}^i / \sum_{j=1}^N \widetilde{w}^j$.

Each step is carried out for $i = 1, \ldots, N$.

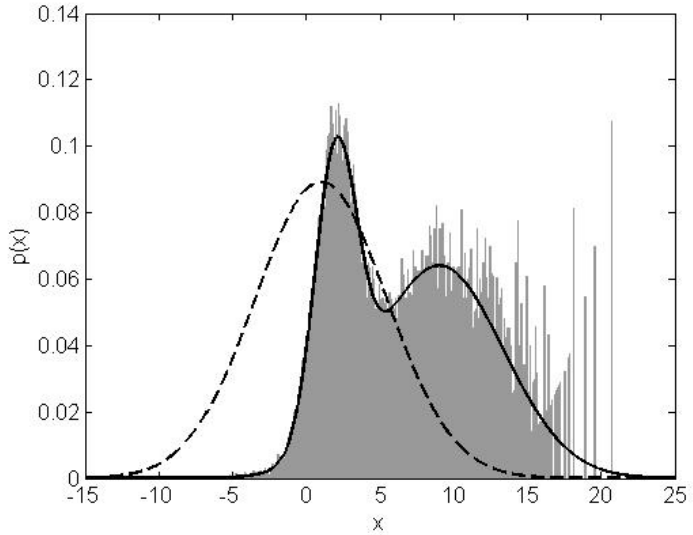
The convergence of the resulting approximation $\widehat{\pi}^N(x) = \sum_{i=1}^N w^i \delta_{x^i}(x)$ is since long well established.

The fact that we are sampling from a user-chosen proposal distribution $q(x)$ is corrected for by the weights, which **accounts for the discrepancy** between the proposal $q(x)$ and the target $\pi(x)$.

The importance of a good proposal density



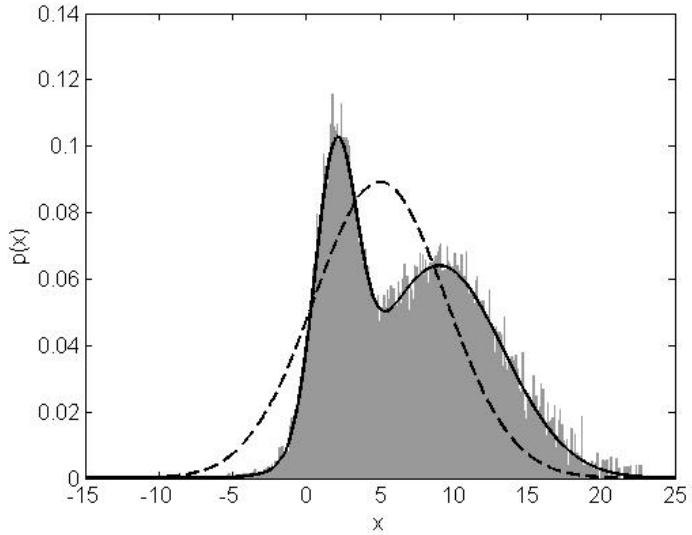
$q_1(x) = \mathcal{N}(x \mid 5, 20)$ (dashed)



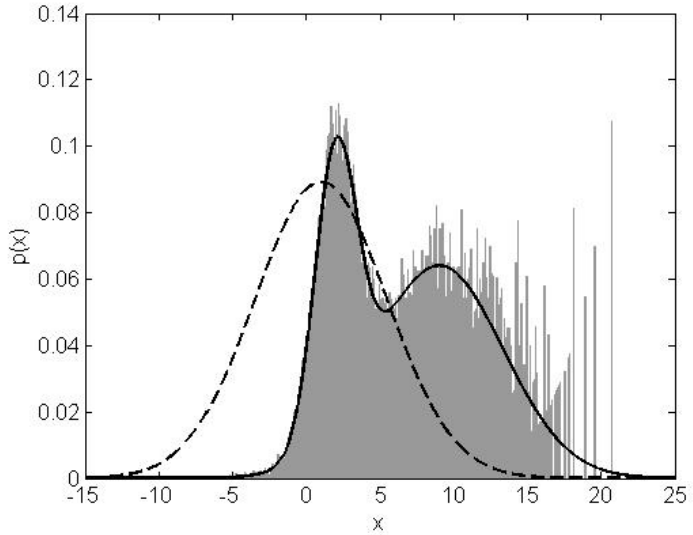
$q_2(x) = \mathcal{N}(x \mid 1, 20)$ (dashed)

50 000 samples were used in both simulations.

The importance of a good proposal density



$q_1(x) = \mathcal{N}(x \mid 5, 20)$ (dashed)



$q_2(x) = \mathcal{N}(x \mid 1, 20)$ (dashed)

50 000 samples were used in both simulations.

Lesson learned: It is important to be careful in selecting the proposal distribution.

Ex) Importance sampling of the joint filtering PDF

Problem statement: Use importance sampling to compute the joint filtering PDF $p(x_{1:t} \,|\, y_{1:t})$ for $(x = x_{1:t}, \pi(x) = p(x_{1:t} \,|\, y_{1:t}))$

$$X_t \,|\, (X_{t-1} = x_{t-1}) \sim p(x_t \,|\, x_{t-1}), \qquad X_t = f(X_{t-1}) + V_t,$$
$$Y_t \,|\, (X_t = x_t) \sim p(y_t \,|\, x_t), \qquad Y_t = g(X_t) + E_t,$$
$$X_0 \sim p(x_0). \qquad X_0 \sim p(x_0).$$

Ex) Importance sampling of the joint filtering PDF

Problem statement: Use importance sampling to compute the joint filtering PDF $p(x_{1:t} \,|\, y_{1:t})$ for $(x = x_{1:t}, \pi(x) = p(x_{1:t} \,|\, y_{1:t}))$

$$X_t \,|\, (X_{t-1} = x_{t-1}) \sim p(x_t \,|\, x_{t-1}), \qquad X_t = f(X_{t-1}) + V_t,$$
$$Y_t \,|\, (X_t = x_t) \sim p(y_t \,|\, x_t), \qquad Y_t = g(X_t) + E_t,$$
$$X_0 \sim p(x_0). \qquad X_0 \sim p(x_0).$$

Key challenge: Nontrivial to design proposal distributions for high-dimensional problems. Here the dimension of the space $\mathcal{X}^t$ grows with t! $(x_t \in \mathcal{X})$.

Ex) Importance sampling of the joint filtering pdf

Idea: Reuse computations over time by exploiting the sequential structure of the SSM via a proposal distribution that factorizes as

$$q(x_{0:t} \mid y_{1:t}) = q(x_0) \prod_{s=1}^{t} q(x_s \mid x_{0:s-1}, y_{1:s}) = q(x_0) \prod_{s=1}^{t} q(x_s \mid x_{s-1}, y_s)$$

Ex) Importance sampling of the joint filtering pdf

Idea: Reuse computations over time by exploiting the sequential structure of the SSM via a proposal distribution that factorizes as

$$q(x_{0:t} \mid y_{1:t}) = q(x_0) \prod_{s=1}^{t} q(x_s \mid x_{0:s-1}, y_{1:s}) = q(x_0) \prod_{s=1}^{t} q(x_s \mid x_{s-1}, y_s)$$

Next we derive the weight function

$$\begin{aligned} \omega_t(x_{0:t}) &= \frac{\widetilde{\pi}(x_{0:t})}{q(x_{0:t})} = \frac{p(x_{0:t}, y_{1:t})}{q(x_{0:t} \mid y_{1:t})} = \dots \\ &= \frac{p(y_t \mid x_t) p(x_t \mid x_{t-1})}{q(x_t \mid x_{t-1}, y_t)} \underbrace{\frac{p(x_{0:t-1}, y_{1:t-1})}{q(x_{0:t-1} \mid y_{1:t-1})}}_{\omega_{t-1}(x_{0:t-1})} \end{aligned}$$

Ex) Importance sampling of the joint filtering pdf

Idea: Reuse computations over time by exploiting the sequential structure of the SSM via a proposal distribution that factorizes as

$$q(x_{0:t} \mid y_{1:t}) = q(x_0) \prod_{s=1}^{t} q(x_s \mid x_{0:s-1}, y_{1:s}) = q(x_0) \prod_{s=1}^{t} q(x_s \mid x_{s-1}, y_s)$$

Next we derive the weight function

$$\begin{aligned} \omega_t(x_{0:t}) &= \frac{\widetilde{\pi}(x_{0:t})}{q(x_{0:t})} = \frac{p(x_{0:t}, y_{1:t})}{q(x_{0:t} \mid y_{1:t})} = \dots \\ &= \frac{p(y_t \mid x_t) p(x_t \mid x_{t-1})}{q(x_t \mid x_{t-1}, y_t)} \underbrace{\frac{p(x_{0:t-1}, y_{1:t-1})}{q(x_{0:t-1} \mid y_{1:t-1})}}_{\omega_{t-1}(x_{0:t-1})} \end{aligned}$$

Hence, the weights can also be computed sequentially

$$\widetilde{w}_t = \frac{p(y_t \mid x_t) p(x_t \mid x_{t-1})}{q(x_t \mid x_{t-1}, y_t)} \widetilde{w}_{t-1}$$

Ex) Importance sampling of the joint filtering pdf

Sequential importance sampling: New samples are proposed sequentially and weights are computed sequentially.

Show stopper: It can be shown that the variance of the weights will grow unboundedly (**weight degeneracy**).

Practical consequence of weight degeneracy: after some time there will only be one weight with non-zero value (more in lecture 5).

Ex) Importance sampling of the joint filtering pdf

Sequential importance sampling: New samples are proposed sequentially and weights are computed sequentially.

Show stopper: It can be shown that the variance of the weights will grow unboundedly (**weight degeneracy**).

Practical consequence of weight degeneracy: after some time there will only be one weight with non-zero value (more in lecture 5).

Next lecture we will derive a working importance sampler by directly target the (marginal) filtering density $p(x_t \mid y_{1:t})$.

Note that the dimension of x_t is fixed, whereas the dimension of $p(x_{0:t} \mid y_{1:t})$ grows with t.

A few concepts to summarize lecture 3

Monte Carlo method: Computational method making use of random sampling to obtain numerical solutions.

Target density: The probability density function that we are interested in.

Empirical approximation: An approximation of a distribution made up of weighted samples.

Importance sampling: A general technique for estimating properties of some target distribution when we only have access to samples from a distribution that is different from the target distribution.

Proposal distribution: A user-chosen distribution that it should be simple to sample from.

Sequential importance sampling: An importance sampler where the proposal distribution is defined sequentially and where the weights can be evaluated sequentially.