Sequential Monte Carlo methods
Lecture 3 – Monte Carlo and importance sampling

Thomas Schöns, Uppsala University
2017-08-24

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

\[ \hat{x}_{t|t} = \mathbb{E}[X_t | y_{1:t}] = \int x_t p(x_t | y_{1:t}) dx_t. \]

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

Common test functions

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

\[ \mathbb{E}[\varphi(X) | y_{1:t}] = \int \varphi(x) p(x | y_{1:T}) dx \]

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

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

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
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

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

Allows for the following approximation of the integral,

$$E_{\pi}[\phi(X)] = \int \phi(x)\pi(x)dx \approx \int \phi(x)\hat{\pi}^N(x)dx = \frac{1}{N} \sum_{i=1}^N \phi(x^i).$$

The strong law of large numbers tells us that $\hat{I}_N(\phi) \xrightarrow{a.s.} I(\phi)$, $N \to \infty$, and the central limit theorem states that

$$\sqrt{N} \left( \hat{I}_N(\phi) - I(\phi) \right) \xrightarrow{\text{d}} N(0,1), \quad N \to \infty.$$
Importance sampling – proposal distribution

The proposal distribution 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\text{a.k.a. importance distribution or instrumental distribution.}

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 \( Z \),

\[
\pi(x) = \frac{\tilde{\pi}(x)}{Z}
\]

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

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

\[
\frac{p(x_{0:t} | y_{1:t})}{\pi(x)} = \frac{\frac{\tilde{\pi}(x)}{Z}}{\frac{p(x_{0:t}, y_{1:t})}{p(y_{1:t})}},
\]

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

Self-normalized importance sampling

Insert

\[
\pi(x) = \frac{\tilde{\pi}(x)}{Z}
\]

into the importance sampling integral results in

\[
I(\varphi) = \mathbb{E}[\varphi(X)] = \int \varphi(x) \frac{\tilde{\pi}(x)}{Zq(x)} q(x) dx = \frac{1}{Z} \int \varphi(z) \frac{\tilde{\pi}(x)}{q(x)} q(x) dx
\]

Hence, the importance sampling estimator is

\[
\tilde{I}^N(\varphi) = \frac{1}{NZ} \sum_{i=1}^{N} \tilde{w}^i \varphi(x^i),
\]

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

The normalization constant \( Z \) is still problematic.
The normalization constant is given by the following integral
\[ Z = \int \tilde{\pi}(x) dx, \]
which we can approximate using our samples \( \{x^i\}_{i=1}^N \) from \( q(x) \). The result is
\[ Z = \int \frac{\tilde{\pi}(x)}{q(x)} q(x) dx \approx \frac{1}{N} \sum_{i=1}^N \tilde{w}^i \]

The self-normalized importance sampling estimate is obtained by inserting this into \( \tilde{I}^N(\varphi) \),
\[ \tilde{I}^N(\varphi) = \sum_{i=1}^N w^i \varphi(x^i), \quad w^i = \frac{\tilde{w}^i}{\sum_{j=1}^N \tilde{w}^j} \]

Algorithm 1 Importance sampler
1. Sample \( x^i \sim q(x) \).
2. Compute the weights \( \tilde{w}^i = \frac{\tilde{\pi}(x^i)}{q(x^i)} \).
3. Normalize the weights \( w^i = \frac{\tilde{w}^i}{\sum_{j=1}^N \tilde{w}^j} \).

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

The convergence of the resulting approximation \( \hat{\pi}^N(x) = \sum_{i=1}^N w^i \delta_{x^i}(x) \) 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 \( \pi(x) \).

Ex) Importance sampling of the joint filtering PDF

Problem statement: Use importance sampling to compute the joint filtering PDF \( p(x_{1:t} \mid y_{1:t}) \) for \( (x = x_{1:t}, \pi(x) = p(x_{1:t} \mid y_{1:t})) \)
\[ X_{t+1} \mid (X_t = x_t) \sim p(x_{t+1} \mid x_t), \quad X_{t+1} = f(X_t) + V_t, \]
\[ Y_t \mid (X_t = x_t) \sim p(y_t \mid x_t), \quad Y_t = g(X_t) + E_t, \]
\[ X_0 \sim p(x_0), \quad X_0 \sim p(x_0). \]

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

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

Next we derive the weight function

\[
\omega_t(x_{0:t}) = \frac{\tilde{\pi}(x_{0:t})}{q(x_{0:t})} = \frac{p(x_{0:t}, y_{1:t})}{q(x_{0:t})} = \ldots
\]

\[
= \frac{p(y_t \mid x_t) p(x_t \mid x_{t-1})}{q(x_t \mid x_{t-1}, y_t)} \frac{p(x_{0:t-1}, y_{1:t-1})}{q(x_{0:t-1} \mid y_{1:t-1})}
\]

\[
\omega_{t-1}(x_{0:t-1})
\]

Hence, the weights can also be computed sequentially

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

**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.

**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 \(t\).