Lecture Notes: Importance Sampling

Last time we considered how a non-Gaussian distribution could be modelled. Figure 1 shows an example. Realizing that the distribution could be intractable, we abandoned the idea of writing an equation to describe it. Instead, we can approximate the distribution using a Monte Carlo technique:

$$\chi = \{x^{(m)}, w^{(m)}\}_{m=1}^{M} \tag{1}$$

where $x^{(m)}$ represents the state and $w^{(m)}$ represents the weight of a single sample (m).

In a filtering problem, the primary distribution of interest is the probability of state given a measurement, denoted as $p(x|y)$. This is the distribution we will model using χ. Looking at equation 1, is it easy to envision randomly selecting samples $x^{(m)}$ in the state space. Each sample is simply a guess at the actual state. However, it is not clear how to weight each guess, in other words how to calculate $w^{(m)}$. If we knew $p(x|y)$ then we would calculate $w^{(m)} = p(x^{(m)}|y)$; in other words, the weight is the probability of that state being the actual state given the measurement. But we don't know $p(x|y)$.

Importance sampling gives us a technique to work around this problem. We start by defining the expected value of $x|y$:

$$E_p[x] = \int x \cdot p(x|y) dx \tag{2}$$

It is the value of all possible states x across the probability of each of those states given the measurement. Importance sampling uses a simple but clever identity:

$$E_p[x] = \int x \frac{p(x|y)}{q(x|y)} q(x|y) dx \tag{3}$$

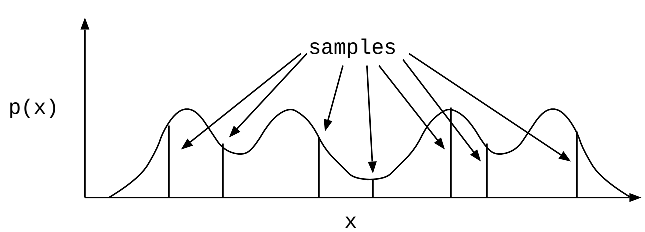

Figure 1: Monte Carlo approximation of a non-Gaussian distribution.

The distribution $q(x|y)$ is called a proposal distribution, also known as a sampling distribution. It must be known and tractable; for example, it could be a simple Gaussian. Therefore we can calculate its values. Let

$$w(x) = \frac{p(x|y)}{q(x|y)} \tag{4}$$

Then

$$E_p[x] = \int x \cdot w(x) q(x|y) dx = E_q[x \cdot w(x)] \tag{5}$$

In other words, we can calculate expected values (or other properties of the distribution, such as local maxima) on $p(x|y)$ using $q(x|y)$ if we can weight them according to $w(x)$. Combining this with the Monte Carlo principle, we obtain

$$E_p[x] \approx \sum_{m=1}^{M} w^{(m)} \cdot x^{(m)} \tag{6}$$

where the weights are calculated as

$$w^{(m)} = \frac{p(x^{(m)}|y)}{q(x^{(m)}|y)} \tag{7}$$

The astute reader may notice that equation 7 does not seem to address the original problem. Although we can calculate values from q we cannot calculate values from p, so how is this any better than where we started?

Sequential importance sampling puts the idea to use in an iterative framework. We define the weight of a sample at time t as

$$w_t^{(m)} = \frac{p(x_{0:t}^{(m)}|y_{0:t})}{q(x_{0:t}^{(m)}|y_{0:t})} \tag{8}$$

Recall the formula for recursive Bayesian estimation:

$$p(x_{0:t}|y_{0:t}) = \frac{p(x_t|x_{t-1})p(y_t|x_t)}{p(y_t|y_{0:t-1})} p(x_{0:t-1}|y_{0:t-1}) \tag{9}$$

Combining these two equations, we obtain:

$$w_t^{(m)} = \frac{p(x_t^{(m)}|x_{t-1}^{(m)})p(y_t|x_t^{(m)})p(x_{0:t-1}^{(m)}|y_{0:t-1})}{p(y_t|y_{0:t-1})q(x_{0:t}^{(m)}|y_{0:t})} \tag{10}$$

The term $p(x_t^{(m)}|x_{t-1}^{(m)})$ is known from the state transition equation, so it can be calculated. The term $p(y_t|x_t^{(m)})$ is known from the observation equation, so it can be calculated. The term $p(x_{0:t-1}^{(m)}|y_{0:t-1})$ is our previous estimate of state, and is known in an iterative framework. The only troublesome term is $p(y_t|y_{0:t-1})$, but its main purpose is to normalize the distribution. Therefore, we will abandon it, at the cost of having non-normalized but still proportional weights. Let

$$w_t^{(m)} \propto \tilde{w}_t^{(m)} = \frac{p(x_t^{(m)}|x_{t-1}^{(m)})p(y_t|x_t^{(m)})p(x_{0:t-1}^{(m)}|y_{0:t-1})}{q(x_{0:t}^{(m)}|y_{0:t})} \tag{11}$$

The denominator term can also be expanded iteratively as follows:

$$q(x_{0:t}^{(m)}|y_{0:t}) = q(x_{0:t-1}^{(m)}|y_{0:t-1}) \cdot q(x_t^{(m)}|x_{0:t-1}^{(m)}, y_{0:t}) \tag{12}$$

where the term $q(x_{0:t-1}^{(m)}|y_{0:t-1})$ represents the distribution at the previous time $t-1$, and the term $q(x_t^{(m)}|x_{0:t-1}^{(m)}, y_{0:t})$ represents the probability of transitioning to state $x^{(m)}$ at time t given the new measurement y_t. Equations 11-12 can be combined to produce:

$$\tilde{w}_t^{(m)} = \frac{p(x_t^{(m)}|x_{t-1}^{(m)})p(y_t|x_t^{(m)})}{q(x_t^{(m)}|x_{0:t-1}^{(m)}, y_{0:t})} \frac{p(x_{0:t-1}^{(m)}|y_{0:t-1})}{q(x_{0:t-1}^{(m)}|y_{0:t-1})} \tag{13}$$

The second fraction in that equation can be recognized as the weight at the previous iteration. Therefore:

$$\tilde{w}_t^{(m)} = \frac{p(x_t^{(m)}|x_{t-1}^{(m)})p(y_t|x_t^{(m)})}{q(x_t^{(m)}|x_{0:t-1}^{(m)}, y_{0:t})} w_{t-1}^{(m)} \tag{14}$$

After calculating the iteratively updated weights $\tilde{w}$, they must be normalized:

$$w_t^{(m)} = \frac{\tilde{w}_t^{(m)}}{\sum_{m=1}^{M} \tilde{w}_t^{(m)}} \tag{15}$$

Through this derivation, sequential importance sampling has provided a method to avoid making calculations that involve p. However, the astute reader will again notice a problem. Equation 14 contains the strange term $q(x_t^{(m)}|x_{0:t-1}^{(m)}, y_{0:t})$. How can this be calculated?

The final principle to making this work in a filtering framework is to select the q distribution. Recall that previously, all we said was that it needs to be tractable and known. It turns out that there are a few good choices for q that make filtering easy. One is to select it as the state transition equation $p(x_t^{(m)}|x_{t-1}^{(m)})$, also known as the prior importance function. Then equation 14 simplifies to

$$\tilde{w}_t^{(m)} = p(y_t|x_t^{(m)})\; w_{t-1}^{(m)} \tag{16}$$

Other functions can be selected that similarly simplify equation 14. Theoretically, the function should be selected such that it has good coverage of the original $p(x|y)$ distribution. This means that it should tend to follow the same shape, or at least have appreciable value across the same general range. However, in practice the q distribution is almost always chosen to simplify the weight update equation, making the computations easier.