4F7-STATISTICAL SIGNAL ANALYSIS

Solutions to the Examples Paper

Question 1: Consider the following hidden Markov model,

$$X_{k+1} = aX_k + bW_{k+1},$$

$$Y_k = cX_k + dV_k, \qquad k = 0, 1, \ldots \tag{0.1}$$

where $\{V_k\}$ and $\{W_k\}$ are independent and identically distributed $\mathcal{N}(0,1)$ and X_0 is $\mathcal{N}(0,b^2)$. Give the expressions for the transition probability density function $f(x_k, x_{k+1})$ and the observation probability density function $g(x_k, y_k)$.

If $X_k = x_k$ then X_{k+1} is a Gaussian random variable with mean ax_k and variance b^2, or

$$f(x_k, x_{k+1}) = \frac{1}{\sqrt{2\pi}b} \exp\left\{-\frac{(x_{k+1} - ax_k)^2}{2b^2}\right\}.$$

A similar reasoning applied to the observation process yields

$$g(x_k, y_k) = \frac{1}{\sqrt{2\pi}d} \exp\left\{-\frac{(y_k - cx_k)^2}{2d^2}\right\}.$$

Question 2: Henceforth, let $a = 0$ and $c = 1$. Find the expression for $p(x_0, \ldots, x_n \mid y_0, \ldots, y_n)$.

$$
\begin{aligned}
&p(x_0, y_0, \ldots, x_n, y_n) \\
&= p(x_0)g(x_0, y_0)f(x_0, x_1)g(x_1, y_1)\cdots f(x_{n-1}, x_n)g(x_n, y_n).
\end{aligned}
$$

When $a = 0$ the state at time $k+1$ does not depend on the state at time k and

$$f(x_k, x_{k+1}) = \frac{1}{\sqrt{2\pi}b} \exp\left\{-\frac{(x_{k+1})^2}{2b^2}\right\}$$

which is of the same form as $p(x_0)$. The hidden chain is a sequence of independent and identically distributed random variables. Thus

$$
\begin{aligned}
&p(x_0, y_0, \ldots, x_n, y_n) \\
&= p(x_0)g(x_0, y_0)p(x_1)g(x_1, y_1)\cdots p(x_n)g(x_n, y_n)
\end{aligned}
$$

and

$$
\begin{aligned}
p(x_0, \ldots, x_n \mid y_0, \ldots, y_n) &= \frac{p(x_0)g(x_0, y_0)}{p(y_0)} \cdots \frac{p(x_n)g(x_n, y_n)}{p(y_n)} \\
&= p(x_0 \mid y_0) \cdots p(x_n \mid y_n).
\end{aligned}
$$

To derive $p(x_k \mid y_k)$, we note that (X_k, Y_k) is a zero mean Gaussian vector with covariance

$$\begin{bmatrix} b^2 & b^2 \\ b^2 & b^2 + d^2 \end{bmatrix}.$$

Thus given $Y_k = y_k$ the conditional probability density function of X_k, $p(x_k \mid y_k)$, is Gaussian with mean

$$y_k \frac{b^2}{b^2+d^2} \quad \text{and variance} \quad \frac{b^2 d^2}{b^2+d^2}.$$

As $d \to 0$, the conditional probability density function mean will tend to y_k and the variance tends to zero. If $d = \infty$ then the conditional probability density function is just the prior of X_k.

Question 3: Construct a self-normalising importance sampling of $p(x_0, \ldots, x_n \mid y_0, \ldots, y_n)$ and consequently an importance sampling estimate of $p(y_0, \ldots, y_n)$. Show that the estimate of $p(y_0, \ldots, y_n)$ is unbiased.

Let $q_n(x_0, \ldots, x_n) = q_0(x_0) q_1(x_0, x_1) \cdots q_n(x_{n-1}, x_n)$ be the proposal probability density function where $\int q_k(x_{k-1}, x_k) dx_k = 1$ for all k and $\int q_0(x_0) dx_0 = 1$. Let $X_{0:n}^i$, $i = 1, \ldots, N$, be independent samples from $q_n(x_{0:n})$ and let

$$w_n^i = \pi_n(X_{0:n}^i)/q_n(X_{0:n}^i)$$

where

$$\pi_n(x_0, \ldots, x_n) = p(x_0, y_0, \ldots, x_n, y_n).$$

The self-normalised importance sampling estimate of $\int h_n(x_{0:n})p(x_{0:n} \mid y_{0:n})dx_{0:n}$, for any function of interest $h_n(x_{0:n})$ we wish to integrate, can be found by first expressing the integral as

$$\begin{aligned}
&\int h_n(x_{0:n})p(x_{0:n} \mid y_{0:n})dx_{0:n} \\
&= \frac{\int h_n(x_{0:n})p(x_{0:n}, y_{0:n})dx_{0:n}}{\int p(x_{0:n}, y_{0:n})dx_{0:n}} \\
&= \frac{\int h_n(x_{0:n})\pi_n(x_{0:n})dx_{0:n}}{\int \pi_n(x_{0:n})dx_{0:n}} \\
&= \frac{\int h_n(x_{0:n})\left(\pi_n(x_{0:n})/q_n(x_{0:n})\right)q_n(x_{0:n})dx_{0:n}}{\int \left(\pi_n(x_{0:n})/q_n(x_{0:n})\right)q_n(x_{0:n})dx_{0:n}}.
\end{aligned}$$

Thus the importance sampling estimate using independent samples from $q_n(x_{0:n})$ is

$$\frac{\sum_{i=1}^N w_n^i h_n(X_{0:n}^i)}{\sum_{j=1}^N w_n^j}.$$

Note that the samples from $q_n(x_{0:n})$ are used to approximate the numerator and denominator separately.

Since

$$\begin{aligned}
p(y_0, \ldots, y_n) &= \int \pi_n(x_{0:n})dx_{0:n} \\
&= \int \frac{\pi_n(x_{0:n})}{q_n(x_{0:n})} q_n(x_{0:n})dx_{0:n},
\end{aligned}$$

an unbiased estimate of $p(y_0, \ldots, y_n)$ is

$$\frac{1}{N}\sum_{j=1}^N w_n^j.$$

This estimate is unbiased because

$$\mathbb{E}\left(w_n^j\right) = \mathbb{E}\left(\pi_n(X_{0:n}^j)/q_n(X_{0:n}^j)\right) = \int \frac{\pi_n(x_{0:n})}{q_n(x_{0:n})} q_n(x_{0:n}) dx_{0:n}.$$

Question 4: Find the variance σ^2/N of the self-normalising importance sampling of $p(x_0, \ldots, x_n \mid y_0, \ldots, y_n)$ and then the variance σ_0^2/N of the estimate that uses N independent samples from $p(x_0, \ldots, x_n \mid y_0, \ldots, y_n)$.

The variance of self-normalising importance sampling is approximately (see lecture notes)

$$\frac{1}{N}\mathbb{E}_{\pi_n^*}\left\{\left(h_n(X_{0:n}) - s_1\right)^2 w^*(X_{0:n})\right\}$$

where

$$\pi_n^*(x_{0:n}) = \frac{\pi_n(x_{0:n})}{\int \pi_n(x_{0:n}) dx_{0:n}}, \qquad w_n^*(x_{0:n}) = \frac{\pi_n^*(x_{0:n})}{q_n(x_{0:n})}.$$

and

$$s_1 = \mathbb{E}_{\pi_n^*}\left\{h_n(X_{0:n})\right\}.$$

The variance of the estimate using N independent samples from $p(x_{0:n} \mid y_{0:n})$ is

$$\frac{1}{N}\mathbb{E}_{\pi_n^*}\left\{\left(h_n(X_{0:n}) - s_1\right)^2\right\}.$$

(Note this is an exact calculation.) The difference between the two variances is due to the non-negative term inside the expectation,

$$\begin{aligned} w_n^*(x_{0:n}) &= \frac{\pi_n^*(x_{0:n})}{q_n(x_{0:n})} \\ &= \frac{p(x_0 \mid y_0)}{p(x_0)} \cdots \frac{p(x_n \mid y_n)}{p(x_n)} \end{aligned}$$

where the last line follows if we let

$$q_n(x_{0:n}) = p(x_0)p(x_1)\cdots p(x_n),$$

which is the probability density function of the hidden state.

Question 5: Find the number of samples N_1 such that $\sigma^2/N_1 = \sigma_0^2/N$. Discuss what happens as $d \to 0$.

Equating the two variances σ^2/N_1 and σ_0^2/N gives

$$\frac{1}{N_1}\mathbb{E}_{\pi_n^*}\left\{(h_n(X_{0:n}) - s_1)^2 \, w^*(X_{0:n})\right\} = \frac{1}{N}\mathbb{E}_{\pi_n^*}\left\{(h_n(X_{0:n}) - s_1)^2\right\}$$

whence

$$N_1 = N\frac{\mathbb{E}_{\pi_n^*}\left\{(h_n(X_{0:n}) - s_1)^2 \, w^*(X_{0:n})\right\}}{\mathbb{E}_{\pi_n^*}\left\{(h_n(X_{0:n}) - s_1)^2\right\}}.$$

Unlike the case of uniformly distributed observations Y_k in the lecture notes, a further simplification is not trivial.

From an earlier question we found that $p(x_k \mid y_k)$ is a Gaussian and its mass concentrates around y_k as $d \to 0$. As $d \to 0$, the ratio

$$\frac{p(x_k \mid y_k)}{p(x_k)}$$

becomes very large for all values of x_k in a neighbourhood around its mean. Since $w_n^*(x_{0:n})$ is a product of $n+1$ such ratios, it grows in size exponentially in n. (Note that an estimate using N samples from $p(x_0,\ldots,x_n \mid y_0,\ldots,y_n)$ directly does not suffer this problem of exponential variance growth.) Thus we expect many more samples N_1 are needed to match the quality of σ_0^2/N.

Caveat! This explanation of the behaviour of N_1/N as $d \to 0$ is a not a proof and indeed not a substitute for an actual verification via a more detailed analysis. (The question does not ask for such a detailed analysis.)

Question 6: Construct importance sampling estimates of $p(y_0),\ldots,p(y_n)$ and calculate the variance of the estimates.

Since $p(y_k) = \int p(x_k)g(x_k,y_k)dx_k$,

$$\frac{1}{N}\sum_{i=1}^{N} g(X_k^i, y_k)$$

is an unbiased estimate of $p(y_k)$ when X_k^i are independent samples from $p(x_k)$. Its variance is

$$\mathbb{E}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\left[g(X_k^i,y_k)-p(y_k)\right]\right)^2\right\} = \frac{C_k}{N}$$

where

$$C_k = \mathbb{E}\left\{g(X_k^i,y_k)^2\right\} - p(y_k)^2.$$

For use later on, the relative variance (which is by definition variannce/mean^2) is

$$
\begin{aligned}
\frac{C_k}{p(y_k)^2} &= \frac{\mathbb{E}\left\{g(X_k^i, y_k)^2\right\}}{p(y_k)^2} - 1 \\
&= \frac{\int p(y_k \mid x_k) p(y_k \mid x_k) p(x_k) dx_k}{p(y_k)^2} - 1 \\
&= \frac{\int p(y_k \mid x_k) p(x_k \mid y_k) dx_k}{p(y_k)} - 1.
\end{aligned}
$$

Question 7: Show that the product of the importance sampling estimates of $p(y_0), \ldots, p(y_n)$ is also an unbiased estimate of $p(y_0, \ldots, y_n)$. Compare the variance of this new estimate with that of the importance sampling estimate of $p(y_0, \ldots, y_n)$ from Question 3.

An estimate of $p(y_0, \ldots, y_n) = p(y_0) \cdots p(y_n)$ is thus

$$
\left(\frac{1}{N} \sum_{i=1}^{N} g(X_0^i, y_0)\right) \cdots \left(\frac{1}{N} \sum_{i=1}^{N} g(X_n^i, y_n)\right)
$$

and

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\frac{1}{N} \sum_{i=1}^{N} g(X_0^i, y_0)\right) \cdots \left(\frac{1}{N} \sum_{i=1}^{N} g(X_n^i, y_n)\right)\right\} \\
&= \mathbb{E}\left\{\left(\frac{1}{N} \sum_{i=1}^{N} g(X_0^i, y_0)\right)\right\} \cdots \mathbb{E}\left\{\left(\frac{1}{N} \sum_{i=1}^{N} g(X_n^i, y_n)\right)\right\}
\end{aligned}
$$

by independence of each of the products. It is thus unbiased.

The variance is

$$\mathbb{E}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} g(X_0^i, y_0)\right)^2 \cdots \left(\frac{1}{N}\sum_{i=1}^{N} g(X_n^i, y_n)\right)^2\right\} - p(y_0)^2 \cdots p(y_n)^2$$

$$= \mathbb{E}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} g(X_0^i, y_0)\right)^2\right\} \cdots \mathbb{E}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} g(X_n^i, y_n)\right)^2\right\} - p(y_0)^2 \cdots p(y_n)^2$$

where the last line uses their independence. Recall that

$$\mathbb{E}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} g(X_k^i, y_k)\right)^2\right\} - p(y_k)^2 = \frac{C_k}{N}$$

where C_k is given in the previous question. Thus the variance is

$$\left(\frac{C_0}{N} + p(y_0)^2\right) \ldots \left(\frac{C_n}{N} + p(y_n)^2\right) - p(y_0)^2 \cdots p(y_n)^2$$

and the relative variance is

$$\begin{aligned} &\left(\frac{C_0}{Np(y_0)^2} + 1\right) \ldots \left(\frac{C_n}{Np(y_n)^2} + 1\right) - 1 \\ &< \exp\left(\frac{1}{N}\sum_{i=0}^{n} \frac{C_i}{p(y_i)^2}\right) - 1 \end{aligned}$$

by using the bound $1 + c/N < \exp(c/N)$. Thus we expect the relative variance will not grow with data length $n+1$ if N is also increased linearly with the number of terms $n+1$, e.g. using $N = N_0 n$.

Now comparing with the estimate of $p(y_0, \ldots, y_n)$ from Question 3. Since

$$
\begin{aligned}
p(y_0, \ldots, y_n) &= \int \pi_n(x_{0:n}) dx_{0:n} \\
&= \int g(x_0, y_0) \cdots g(x_n, y_n) p(x_0) p(x_1) \cdots p(x_n) dx_{0:n}
\end{aligned}
$$

an unbiased estimate of $p(y_0, \ldots, y_n)$ is

$$N^{-1} \sum_{i=1}^{N} w_n^i$$

where $w_n^i = g(X_0^i, y_0) \cdots g(X_n^i, y_n)$ and $X_{0:n}^i$ are N independent samples from $p(x_0)p(x_1) \cdots p(x_n)$. Its variance is

$$
\begin{aligned}
&\frac{1}{N} \left(\mathbb{E}\left\{ g(X_0^i, y_0)^2 \cdots g(X_n^i, y_n)^2 \right\} - p(y_0)^2 \cdots p(y_n)^2 \right) \\
&= \frac{1}{N} \left(\mathbb{E}\left\{ g(X_0^i, y_0)^2 \right\} \cdots \mathbb{E}\left\{ g(X_n^i, y_n)^2 \right\} - p(y_0)^2 \cdots p(y_n)^2 \right)
\end{aligned}
$$

and relative variance is

$$\frac{1}{N} \left(\frac{\mathbb{E}\left\{ g(X_0^i, y_0)^2 \right\}}{p(y_0)^2} \cdots \frac{\mathbb{E}\left\{ g(X_n^i, y_n)^2 \right\}}{p(y_n)^2} - 1 \right).$$

Each term

$$\frac{\mathbb{E}\left\{ g(X_k^i, y_k)^2 \right\}}{p(y_k)^2} > 1$$

since $\mathbb{E}\left\{ g(X_0^i, y_0)^2 \right\} - p(y_0)^2 > 0$ and so the relative variance increases exponentially in n due to the product of $n+1$ terms larger than 1. So the number of samples N will also have to increase at the same rate to control growth of the relative variance with data length.

S.S. Singh, Department of Engineering, University of Cambridge, Cambridge, CB1 7AT, UK

Email address: `sss40@cam.ac.uk`