region boundaries. This form of clipping still does not prevent the occurrence of negative concentrations in this example, however. Negative concentration estimates are not avoided by either scaling or clipping of the sigma points. As a solution to this problem, the use of constrained optimization for the sigma points is proposed (Vachhani et al., 2006; Teixeira, Tôrres, Aguirre, and Bernstein, 2008). If one is willing to perform online optimization, however, MHE with a short horizon is likely to provide more accurate estimates at similar computational cost compared to approaches based on optimizing the locations of the sigma points.

Finally, Figure 4.6 presents typical results of applying constrained MHE to this example. For this simulation we choose $N = 10$ and the smoothing update for the arrival cost approximation. Note that MHE recovers well from the poor initial prior. Comparable performance is obtained if the filtering update is used instead of the smoothing update to approximate the arrival cost. The MHE estimates are also insensitive to the choice of horizon length N for this example. □

The EKF, UKF, and all one-step recursive estimation methods, suffer from the "short horizon syndrome" by *design*. One can try to reduce the harmful effects of a short horizon through tuning various other parameters in the estimator, but the basic problem remains. Large initial state errors lead to inaccurate estimation and potential estimator divergence. The one-step recursions such as the EKF and UKF can be viewed as one extreme in the choice between speed and accuracy in that only a single measurement is considered at each sample. That is similar to an MHE problem in which the user chooses $N = 1$. Situations in which $N = 1$ lead to poor MHE performance often lead to unreliable EKF and UKF performance as well.

4.7 Particle Filtering

Particle filtering is a different approach to the state estimation problem in which statistical sampling is used to approximate the evolution of the conditional density of the state given measurements (Handschin and Mayne, 1969). This method also handles nonlinear dynamic models and can address nonnormally distributed random disturbances to the state and measurement.

Sampled density. Consider a smooth probability density, $p(x)$. In particle filtering we find it convenient to represent this smooth density

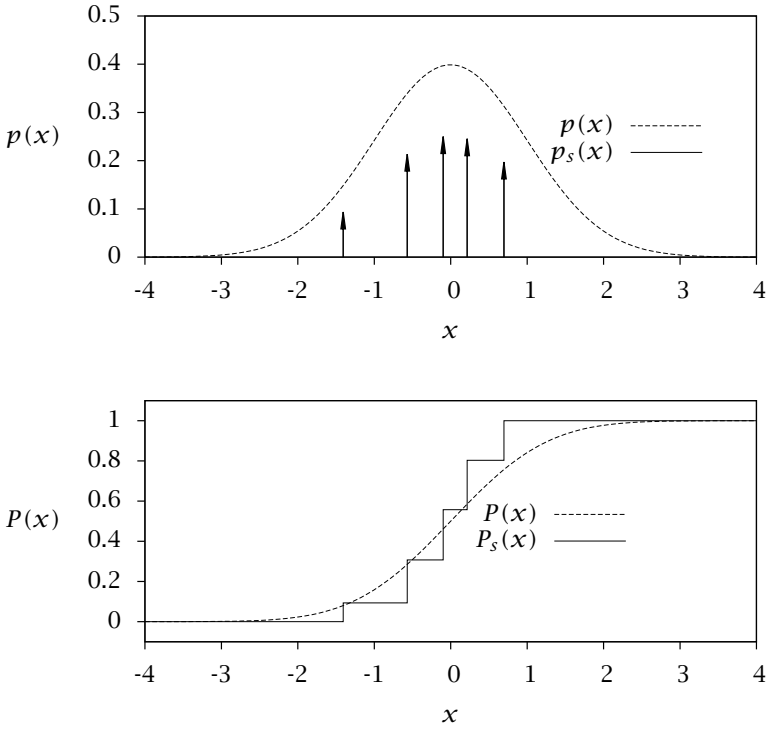

Figure 4.7: Top: exact density $p(x)$ and a sampled density $p_s(x)$ with five samples for $\xi \sim N(0,1)$. Bottom: corresponding exact $P(x)$ and sampled $P_s(x)$ cumulative distributions.

as a weighted, sampled density, $p_s(x)$

$$p(x) \approx p_s(x) := \sum_{i=1}^{s} w_i \delta(x - x_i)$$

in which $x_i, i = 1, \ldots s$ are the samples, w_i are the weights. As an example, the top of Figure 4.7 displays a normally distributed scalar random variable represented by a sampled density with five samples. The sampled density is a series of impulses at the sample locations x_i. In this example, the weights w_i are the values of $p(x_i)$, normalized to sum to unity. It may seem strange to represent a well-behaved function like $p(x)$ with such a "rough" function like $p_s(x)$, but we will

see the advantages shortly. Sometimes we may wish to study convergence of a sampled density to the original density as the number of samples becomes large. To define convergence of this representation of the probability distribution, we refer to the corresponding cumulative distribution rather than the density. From integration, the sampled cumulative distribution is

$$P_s(x) = \sum_{i \in \mathbb{I}_x} w_i \qquad \mathbb{I}_x = \{i | x_i \leq x\}$$

The bottom of Figure 4.7 shows the corresponding cumulative sampled distribution for the sampled density with five samples. The cumulative sampled distribution is a staircase function with steps of size w_i at the sample locations x_i. We can then measure convergence of $P_s(x)$ to $P(x)$ as $s \to \infty$ in any convenient function norm. We delay further discussion of convergence until Section 4.7.2 in which we present some of the methods for choosing the samples and the weights.

In the sequel, we mostly drop the subscript s on sampled densities and cumulative distributions when it is clear from context that we are referring to this type of representation of a probability distribution. We can conveniently calculate the expectation of any function of a random variable having a sampled density by direct integration to obtain

$$\begin{aligned} \mathcal{E}(f(\xi)) &= \int p_s(x) f(x) dx \\ &= \int \sum_i w_i \delta(x - x_i) f(x) dx \\ &= \sum_i w_i f(x_i) \end{aligned}$$

For example, we often wish to evaluate the mean of the sampled density, which is

$$\mathcal{E}(\xi) = \sum_i w_i x_i$$

The convenience of integrating the sampled density is one of its main attractive features. If we create a new function (not necessarily a density) by multiplication of $p(x)$ by another function $g(x)$

$$\overline{p}(x) = g(x) p(x)$$

we can easily obtain the sampled function $\overline{p}$. We simply adjust the

weights and leave the samples where they are

$$\begin{aligned}\overline{p}(x) &= g(x)p(x)\\ &= \sum_i g(x)w_i\delta(x-x_i)\\ &= \sum_i w_i g(x_i)\delta(x-x_i)\\ \overline{p}(x) &= \sum_i \overline{w}_i\delta(x-x_i) \qquad \overline{w}_i = w_i g(x_i)\end{aligned} \tag{4.36}$$

4.7.1 The Sampled Density of a Transformed Random Variable

Given the random variable ξ, assume we have a sampled density for its density $p_\xi(x)$

$$p_\xi(x) = \sum_{i=1}^{s} w_i\delta(x-x_i)$$

Define a new random variable η by an invertible, possibly nonlinear transformation

$$\eta = f(\xi) \qquad \xi = f^{-1}(\eta)$$

We wish to find a sampled density for the random variable η, $p_\eta(y)$. Denote the sampled density for η as

$$p_\eta(y) = \sum_{i=1}^{s} \overline{w}_i\delta(y-y_i)$$

We wish to find formulas for $\overline{w}_i$ and y_i in terms of w_i, x_i and f. We proceed as in the development of equation (A.30) in Appendix A. We wish to have an equivalence for every function $g(x)$

$$\begin{aligned}\mathcal{E}_{p_\xi}(g(\xi)) &= \mathcal{E}_{p_\eta}(g(f^{-1}(\eta)))\\ \int p_\xi(x)g(x)dx &= \int p_\eta(y)g(f^{-1}(y))dy \qquad \text{for all } g(\cdot)\end{aligned}$$

Using the sampled densities on both sides of the equation

$$\sum_i w_i g(x_i) = \sum_i \overline{w}_i g(f^{-1}(y_i))$$

One solution to this equation that holds for every g is the simple choice

$$y_i = f(x_i) \qquad \overline{w}_i = w_i \tag{4.37}$$

We see that for the transformed sampled density, we transform the samples and use the weights of the original density.

Example 4.28: Sampled density of the lognormal

The random variable η is distributed as a lognormal if its logarithm is distributed as a normal. Let $\xi \sim N(0, P)$ and consider the transformation

$$\eta = e^{\xi} \qquad \xi = \log(\eta) \quad \eta > 0$$

Represent p_ξ as a sampled density, use (4.37) to find a sampled density of p_η, and plot histograms of the two sampled densities. Compare the sampled density of p_η to the lognormal density. The two densities are given by

$$p_\xi(x) = \frac{1}{\sqrt{2\pi P}}\, e^{-x^2/2P}$$
$$p_\eta(y) = \frac{1}{y\sqrt{2\pi P}}\, e^{-\log^2(y)/2P}$$

Solution

First we take samples x_i from $N(0, 1)$ for ξ. Figure 4.8 shows the histogram of the sampled density for 5000 samples. Next we compute $y_i = e_i^x$ to generate the samples of η. The histogram of this sampled density is shown in Figure 4.9. Notice the good agreement between the sampled density and the lognormal density, which is shown as the continuous curve in Figure 4.9. □

Noninvertible transformations. Next consider η to be a noninvertible transformation of ξ

$$\eta = f(\xi) \qquad f \text{ not invertible}$$

Let ξ's sampled density be given by $\{x_i, w_i\}$. The sampled density $\{f(x_i), w_i\}$ remains a valid sampled density for η, which we show next

$$\xi \sim \{x_i, w_i\} \qquad \eta \sim \{f(x_i), w_i\}$$

We wish to show that

$$\mathcal{E}_{p_\eta}(g(\eta)) = \mathcal{E}_{p_\xi}(g(f(\xi))) \qquad \text{for all } g(\cdot)$$

Taking the expectations

$$\int p_\eta(y) g(y) dy = \int p_\xi(x) g(f(x)) dx$$

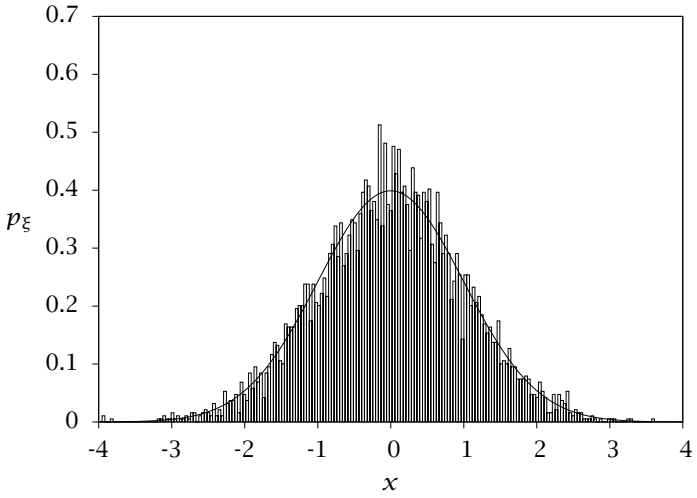

Figure 4.8: Sampled and exact probability densities for $\xi \sim N(0,1)$; 5000 samples.

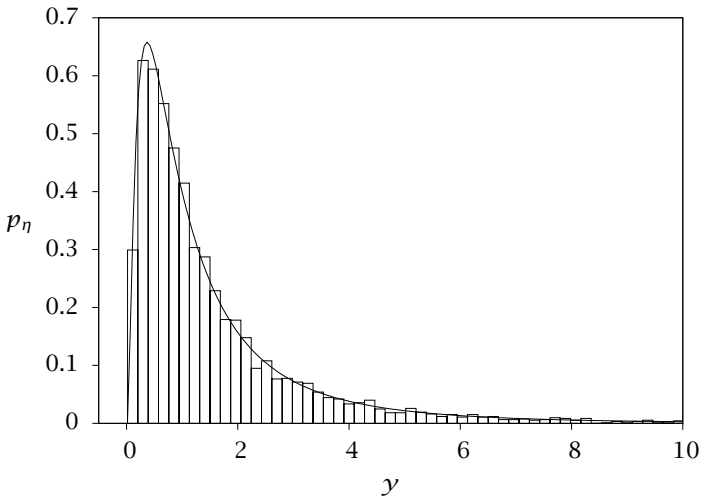

Figure 4.9: Sampled and exact probability densities for nonlinear transformation $\eta = e^{\xi}$; 5000 samples. The exact density of η is the lognormal, shown as the continuous curve.

Letting η's sampled density be $\{y_i, \overline{w}_i\}$, and using ξ's sampled density give

$$\sum_{i=1}^{s} \overline{w}_i g(y_i) = \sum_{i=1}^{s} w_i g(f(x_i))$$

and setting $\overline{w}_i = w_i, y_i = f(x_i), i = 1, \ldots, s$ achieves equality for all $g(\cdot)$, and we have established the result. The difference between the noninvertible and invertible cases is that we do not have a method to obtain samples of ξ from samples of η in the noninvertible case. We can transform the sampled density in only one direction, from p_ξ to p_η.

4.7.2 Sampling and Importance Sampling

Consider a random variable ξ with a smooth probability density $p(x)$. Assume one is able to draw samples x_i of ξ with probability

$$p_{\text{sa}}(x_i) = p(x_i) \tag{4.38}$$

in which $p_{\text{sa}}(x_i)$ denotes the probability of drawing a sample with value x_i. In this case, if one draws s samples, a sampled density for ξ is given by

$$p_s(x) = \sum_i w_i \delta(x - x_i) \qquad w_i = 1/s, \quad i = 1, \ldots, s \tag{4.39}$$

and the weights are all equal to $1/s$.

Convergence of sampled densities. It is instructive to examine how a typical sampled density converges with sample size to the density from which the samples are drawn. Consider a set of s samples. When drawing multiple samples of a density, we assume the samples are mutually independent

$$p_{\text{sa}}(x_1, x_2, \ldots, x_s) = p_{\text{sa}}(x_1) p_{\text{sa}}(x_2) \cdots p_{\text{sa}}(x_s)$$

We denote the cumulative distribution of the sampled density as

$$P_s(x; s) = \sum_{i \in \mathbb{I}_x} w_i \qquad \mathbb{I}_x = \{i | x_i \leq x\}$$

in which the second argument s is included to indicate P_s's dependence on the sample size. The value of P_s is itself a random variable because it is determined by the sample values x_i and weights w_i. We consider

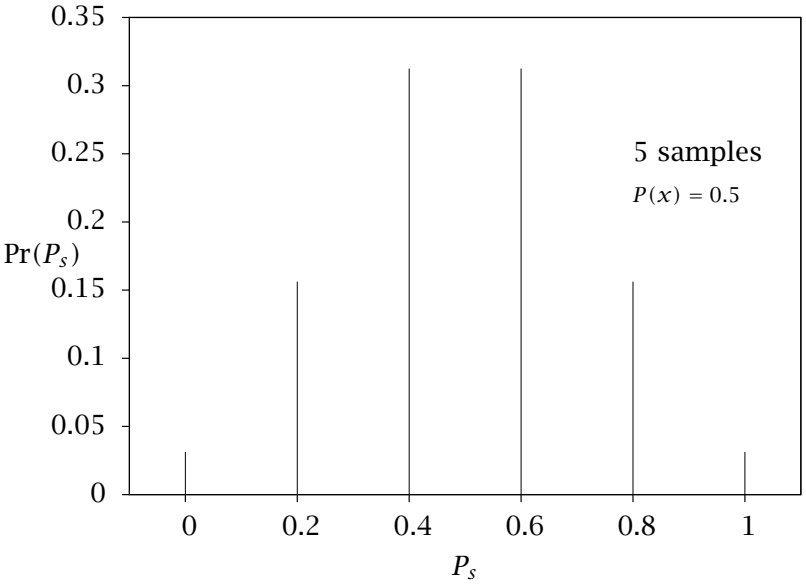

Figure 4.10: Probability density $\Pr(P_s(x;s))$ for x corresponding to $P(x) = 0.5$ and $s = 5$ samples. The distribution is centered at correct value, $P(x)$, but the variance is large.

the case with equal sample weights $w_i = 1/s$ and study the P_s values as a function of s and scalar x. They take values in the range

$$P_s \in \left\{0, \frac{1}{s}, \ldots, \frac{s-1}{s}, 1\right\} \qquad s \geq 1 \quad -\infty < x < \infty$$

Given the sampling process we can readily evaluate the probability of P_s over this set

$$\Pr(P_s(x;s)) = \begin{cases} \binom{s}{i} P(x)^i (1-P(x))^{s-i}, & P_s = \frac{i}{s}, \quad i = 0, \ldots, s \\ 0, & \text{otherwise} \end{cases}$$
$$-\infty < x < \infty \quad (4.40)$$

These probabilities are calculated as follows. For P_s to take on value zero, for example, all of the samples x_i must be greater than x. The probability that any x_i is greater than x is $1-P(x)$. Because the samples are mutually independent, the probability that all s samples are greater than x is $(1-P(x))^s$, which is the $i = 0$ entry of (4.40). Similarly, for P_s

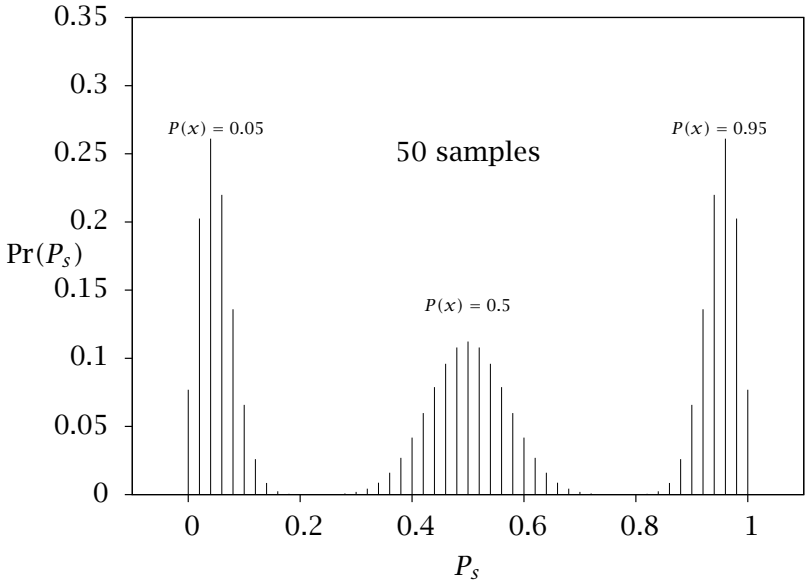

Figure 4.11: Probability density $\Pr(P_s(x;s))$ for three different x corresponding to $P(x) = 0.05, 0.5, 0.95$ and $s = 50$ samples. The three distributions are centered at the correct values, $P(x)$, and the variance is much reduced compared to Figure 4.10.

to have value i/s, i samples must be less than x and $s-i$ samples must be greater than x. This probability is given by $\binom{s}{i}P(x)^i(1-P(x)^{s-i})$, in which $P(x)^i(1-P(x)^{s-i})$ is the probability of having a sample with i values less than x and $s-i$ values greater than x, and $\binom{s}{i}$ accounts for the number of ways such a sample can be drawn from a set of s samples. Figure 4.10 shows the distribution of P_s for a sample size $s = 5$ at the mean, $P(x) = 0.5$. Notice the maximum probability occurs near the value $P_s = P(x)$ but the probability distribution is fairly wide with only 5 samples. The number of samples is increased to 50 in Figure 4.11, and three different x values are shown, at which $P(x) = 0.05, 0.5, 0.95$. The peak for each P_s distribution is near the value $P(x)$, and the distribution is much narrower for 50 samples. The sampled density $P_s(x;s)$ becomes arbitrarily sharply distributed with value $P(x)$

as the sample size s increases.

$$\lim_{s\to\infty} \Pr(P_s(x;s)) = \begin{cases} 1 & P_s = P(x) \\ 0 & \text{otherwise} \end{cases} \qquad -\infty < x < \infty$$

The convergence is often not uniform in x. Achieving a given variance in $P_s(x;s)$ generally requires larger sample sizes for x values in the tails of the density $p(x)$ compared to the sample sizes required to achieve this variance for x values in regions of high density. The nonuniform convergence is perhaps displayed more clearly in Figures 4.12 and 4.13. We have chosen the beta distribution for $P(x)$ and show the spread in the probability of P_s for three x values, corresponding to $P(x) = \{0.1, 0.5, 0.9\}$. Given $s = 25$ samples in Figure 4.12, we see a rather broad probability distribution for the sampled distribution $P_s(x)$. Turning up the number of samples to $s = 250$ gives the tighter probability distribution shown in Figure 4.13.

Finally, we present a classic sampling error distribution result due to Kolmogorov. The measure of sampling error is defined to be

$$D_s = \sup_x |P_s(x;s) - P(x)|$$

and we have the following result on the distribution of D_s for large sample sizes.

Theorem 4.29 (Kolmogoroff (1933))**.** [6] *Suppose that $P(x)$ is continuous. Then for every fixed $z \geq 0$ as $s \to \infty$*

$$\Pr\left(D_s \leq z s^{-1/2}\right) \to L(z) \tag{4.41}$$

in which $L(z)$ is the cumulative distribution function given for $z > 0$ by

$$L(z) = \sqrt{2\pi} z^{-1} \sum_{\nu=1}^{\infty} e^{-(2\nu-1)^2\pi^2/8z^2} \tag{4.42}$$

and $L(z) = 0$ for $z \leq 0$.

One of the significant features of results such as this one is that the limiting distribution is independent of the details of the sampled distribution $P(x)$ itself. Feller (1948) provides a proof of this theorem and discussion of this and other famous sampling error distribution results due to Smirnov (1939).

[6] Kolmogorov's theorem on sampling error was published in an Italian journal with the spelling Kolmogoroff.

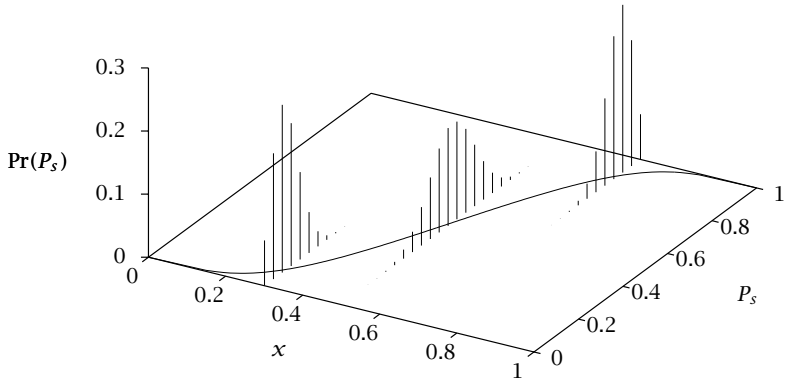

Figure 4.12: Probability density $\Pr(P_s(x;s))$ for $s = 25$ samples at three different x.

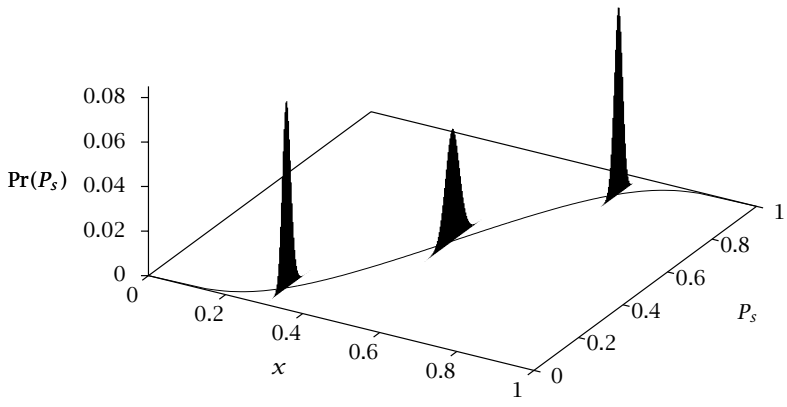

Figure 4.13: Probability density $\Pr(P_s(x;s))$ for $s = 250$ samples. Note the variance is much reduced compared to Figure 4.12.

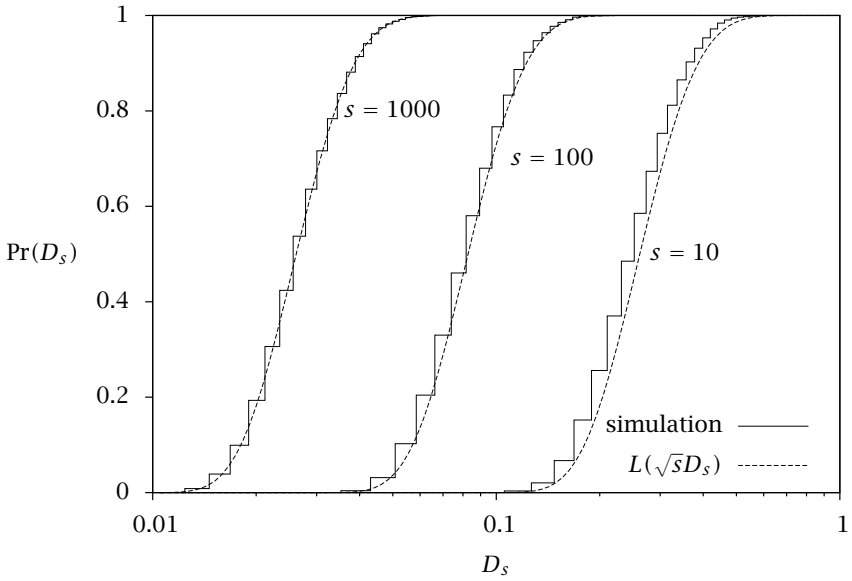

Figure 4.14: Cumulative distribution for the sampling error $\Pr(D_s)$ for three different sample sizes, $s = 10, 100, 1000$. Distribution from simulation using 5000 realizations (solid) and Kolmogorov limiting distribution (dashed).

Example 4.30: Sampling error distribution for many samples

Plot the actual and limiting distributions for D_s for $s = 10, 100, 1000$ when sampling a normal distribution with unit variance. How close is the limiting sampling error distribution to the actual sampling error distribution for these three sample sizes?

Solution

Figure 4.14 displays the result using 5000 realizations of the sampling process to approximate the actual distribution of D_s. Notice that for the small sample size, we can see a slight difference between the Kolmogorov limiting distribution and the one obtained from simulation. This difference is not noticeable for samples sizes greater than $s = 100$. From the argument scaling given in (4.41) we see that the mean of the sampling error decreases by a factor of $\sqrt{10}$ for each factor of 10 increase in sample size (on the log scale, the distribution of D_s is trans-

lated to the left by $\sqrt{10}$). Exercise 4.20 discusses this example further. □

Unbiasedness of sampled densities. A sampled density is *unbiased* if it possesses the following property

$$\mathcal{E}_{\text{sa}}(P_s(x;s)) = P(x) \qquad 1 \le s, \quad -\infty < x < \infty$$

in which the expectation is taken over the probability density of P_s considered as a random variable as discussed previously. As we discuss subsequently, some sampling procedures are unbiased for all s, while others are only asymptotically unbiased as s becomes large. A convenient test for unbiasedness is the following

$$\mathcal{E}_{\text{sa}}\left(\int p_s(x)g(x)dx\right) = \int p(x)g(x)dx \qquad \text{for all } g(\cdot) \tag{4.43}$$

In other words, the *expectation over the sampling process* of integrals of any function g with the sampled density should be equal to the integral of g with the exact density. If the sampling process has the probability given by (4.38), we can verify (4.43) as follows

$$\begin{aligned}
\mathcal{E}_{\text{sa}}\left(\int p_s(x)g(x)dx\right) &= \mathcal{E}_{\text{sa}}\left(\sum_i w_i g(x_i)\right) \\
&= \int p_{\text{sa}}(x_1,\ldots,x_s)\sum_i w_i g(x_i)dx_1\cdots dx_s \\
&= \int p_{\text{sa}}(x_1)\cdots p_{\text{sa}}(x_s)\sum_i w_i g(x_i)dx_1\cdots dx_s \\
&= \int p(x_1)\cdots p(x_s)\sum_i w_i g(x_i)dx_1\cdots dx_s \\
&= \frac{1}{s}\sum_i \int p(x_i)g(x_i)dx_i \prod_{j\neq i}\int p(x_j)dx_j \\
&= \frac{1}{s}\sum_i \int p(x_i)g(x_i)dx_i \\
\mathcal{E}_{\text{sa}}\left(\int p_s(x)g(x)dx\right) &= \int p(x)g(x)dx
\end{aligned}$$

Example 4.31: Sampling independent random variables

Consider two independent random variables ξ, η, whose probability density satisfies

$$p_{\xi,\eta}(x,y) = p_\xi(x)p_\eta(y)$$

and assume we have samples of the two marginals

$$\xi \sim \{x_i, w_{xi}\} \qquad w_{xi} = 1/s_x, \quad i = 1, \ldots, s_x$$
$$\eta \sim \{y_j, w_{yj}\} \qquad w_{yj} = 1/s_y, \quad j = 1, \ldots, s_y$$

We have many valid options for creating samples of the joint density. Here are three useful ones.

(a) Show the following is a valid sample of the joint density

$$\{(x_i, y_j), w_{ij}\} \qquad w_{ij} = 1/(s_x s_y), \quad i = 1, \ldots, s_x, \quad j = 1, \ldots, s_y$$

Notice we have $s_x s_y$ total samples of the joint density.

(b) If $s_x = s_y = s$, show the following is a valid sample of the joint density

$$\{(x_i, y_i), w_i\} \qquad w_i = 1/s, \quad i = 1, \ldots, s$$

Notice we have s total samples of the joint density unlike the previous case in which we would have s^2 samples.

(c) If we have available (or select) only a single sample of ξ's marginal, $s_x = 1$ and $s_y = s$ samples of η's marginal, show the following is a valid sample of the joint density

$$\{(x_1, y_i), w_{yi}\} \qquad w_{yi} = 1/s, \quad i = 1, \ldots, s$$

Here we have generated again s samples of the joint density, but we have allowed unequal numbers of samples of the two marginals.

Solution

Because the two random variables are independent, the probability of drawing a sample with values (x_i, y_j) is given by

$$p_{sa}(x_i, y_j) = p_{sa}(x_i) p_{\text{sa}}(y_j) = p_\xi(x_i) p_\eta(y_j) = p_{\xi,\eta}(x_i, y_j)$$

Denote the samples as $z_k = (x_{i(k)}, y_{j(k)})$. We have for all three choices

$$p_{\text{sa}}(z_k) = p_{\xi,\eta}(z_k) \qquad k = 1, \ldots, s \tag{4.44}$$

(a) For this case,

$$i(k) = \text{mod}(k-1, s_x) + 1 \qquad j(k) = \text{ceil}(k/s_x)$$
$$w_k = \frac{1}{s_x s_y}, \quad k = 1, \ldots, s_x s_y$$

in which $\text{ceil}(x)$ is the smallest integer not less than x.

(b) For this case

$$i(k) = k \qquad j(k) = k \qquad w_k = 1/s, \quad k = 1, \ldots, s$$

(c) For this case

$$i(k) = 1 \qquad j(k) = k \qquad w_k = 1/s, \quad k = 1, \ldots, s$$

Because all three cases satisfy (4.44) and the weights are equal to each other in each case, these are all valid samples of the joint density. □

If we arrange the s_x ξ samples and s_y η samples in a rectangle, the first option takes all the points in the rectangle, the second option takes the diagonal (for a square), and the third option takes one edge of the rectangle. See Exercise 4.19 for taking a single point in the rectangle. In fact, any set of points in the rectangle is a valid sample of the joint density.

Example 4.32: Sampling a conditional density

The following result proves useful in the later discussion of particle filtering. Consider conditional densities satisfying the following property

$$p(a, b, c|d, e) = p(a|b, d)p(b, c|e) \tag{4.45}$$

We wish to draw samples of $p(a, b, c|d, e)$ and we proceed as follows. We draw samples of $p(b, c|e)$. Call these samples $(b_i, c_i), i = 1, \ldots, s$. Next we draw for each $i = 1, \ldots, s$, one sample of $p(a|b_i, d)$. Call these samples a_i. We assemble the s samples (a_i, b_i, c_i) and claim they are samples of the desired density $p(a, b, c|d, e)$ with uniform weights $w_i = 1/s$. Prove or disprove this claim.

Solution

The claim is true, and to prove it we need to establish that the probability of drawing a sample with value (a_i, b_i, c_i) is equal to the desired density $p(a_i, b_i, c_i|d, e)$. We proceed as follows. From the definition of conditional density, we know

$$p_{\text{sa}}(a_i, b_i, c_i|d, e) = p_{\text{sa}}(a_i|b_i, c_i, d, e)p_{\text{sa}}(b_i, c_i|d, e)$$

For the selection of a_i described previously, we know

$$p_{\text{sa}}(a_i|b_i, c_i, d, e) = p(a_i|b_i, d)$$

The values of c_i and e are irrelevant to the sampling procedure generating the a_i. For the (b_i, c_i) samples, the sampling procedure gives

$$p_{\text{sa}}(b_i, c_i | d, e) = p(b_i, c_i | e)$$

and the value of d is irrelevant to the procedure for generating the (b_i, c_i) samples. Combining these results, we have for the (a_i, b_i, c_i) samples

$$p_{\text{sa}}(a_i, b_i, c_i | d, e) = p(a_i | b_i, d) p(b_i, c_i | e)$$

Equation (4.45) then gives

$$p_{\text{sa}}(a_i, b_i, c_i | d, e) = p(a_i, b_i, c_i | d, e)$$

We conclude the sampling procedure is selecting (a_i, b_i, c_i) samples with the desired probability, and as shown in (4.39), the weights are all equal to $1/s$ under this kind of sampling. □

Importance sampling. Consider next the case in which we have a smooth density $p(x)$ that is easy to *evaluate* but difficult to *sample* with probability given by (4.38). This situation is not unusual. In fact, it arises frequently in applications for the following reason. Many good algorithms are available for generating samples of the uniform density. One simple method to sample an arbitrary density for a scalar random variable is the following. First compute $P(x)$ from $p(x)$ by integration. Let u_i be the samples of the uniform density on the interval $[0, 1]$. Then samples of x_i of density $p(x)$ are given by

$$x_i = P^{-1}(u_i) \qquad u_i = P(x_i)$$

Figures 4.15 and 4.16 give a graphical display of this procedure. We briefly verify that the samples x_i have the claimed density. We show that if μ is a uniform random variable and ξ is defined by the invertible transformation given previously, $\mu = P(\xi)$, then ξ has density $p_\xi(x) = dP(x)/dx$. From (A.30) we have

$$p_\xi(x) = p_\mu(P(x)) \left| \frac{dP(x)}{dx} \right|$$

Since μ is uniformly distributed, $p_\mu = 1$, and $dP(x)/dx \geq 0$, we have

$$p_\xi(x) = \frac{dP(x)}{dx}$$

and the samples have the desired density. But notice this procedure for generating samples of $p(x)$ uses $P(x)$, which requires integration,

as well as evaluating $P^{-1}(x)$, which generally requires solving nonlinear equations. Importance sampling is a method for sampling $p(x)$ without performing integration or solving nonlinear equations.

The following idea motivates importance sampling. Consider the random variable ξ to be distributed with density p. Consider a new random variable η to be distributed with density q

$$\xi \sim p(x) \qquad \eta \sim q(x)$$

The density $q(x)$, known as the importance function, is any density that can be readily sampled according to (4.38) and has the same support as p. Examples of such q are uniforms for bounded intervals, lognormals and exponentials for semi-infinite intervals, and normals for infinite intervals. For any function $g(x)$, we have

$$\begin{aligned}
\mathcal{E}_p(g(\xi)) &= \int g(x)p(x)dx \\
&= \int \left[g(x)\frac{p(x)}{q(x)} \right] q(x)dx \\
\mathcal{E}_p(g(\xi)) &= \mathcal{E}_q\left(g(\eta)\frac{p(\eta)}{q(\eta)} \right) \qquad \text{for all } g(\cdot)
\end{aligned}$$

When we can sample p directly, we use for the sampled density

$$p_s = \left\{ x_i, \quad w_i = \frac{1}{s} \right\} \qquad p_{sa}(x_i) = p(x_i)$$

So when we cannot conveniently sample p but can sample q, we use instead

$$\overline{p}_s = \left\{ x_i, \quad w_i = \frac{1}{s}\frac{p(x_i)}{q(x_i)} \right\} \qquad p_{is}(x_i) = q(x_i)$$

Given s samples x_i from $q(x)$, denote the sampled density of q as q_s, and we have defined the importance-sampled density $\overline{p}_s(x)$ as

$$\overline{p}_s(x) = q_s(x)\frac{p(x)}{q(x)}$$

We next show that $\overline{p}_s(x)$ converges to $p(x)$ as sample size increases (Smith and Gelfand, 1992). Using the fact that q_s converges to q gives

$$\lim_{s\to\infty} \overline{p}_s(x) = \lim_{s\to\infty} q_s(x)\frac{p(x)}{q(x)} = p(x)$$

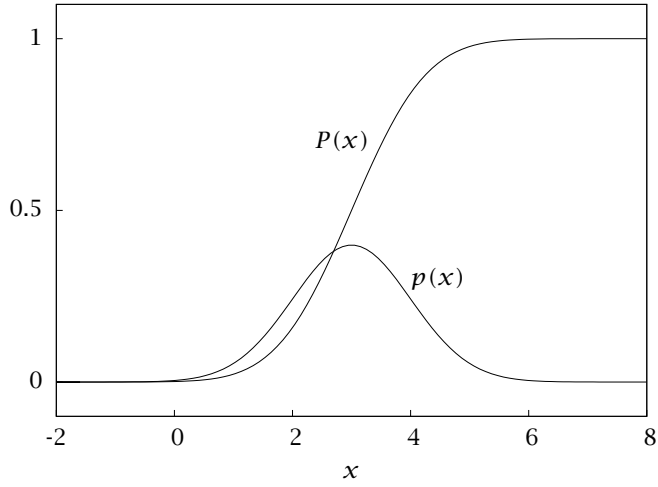

Figure 4.15: Probability density $p(x)$ to be sampled and the corresponding cumulative distribution $P(x)$.

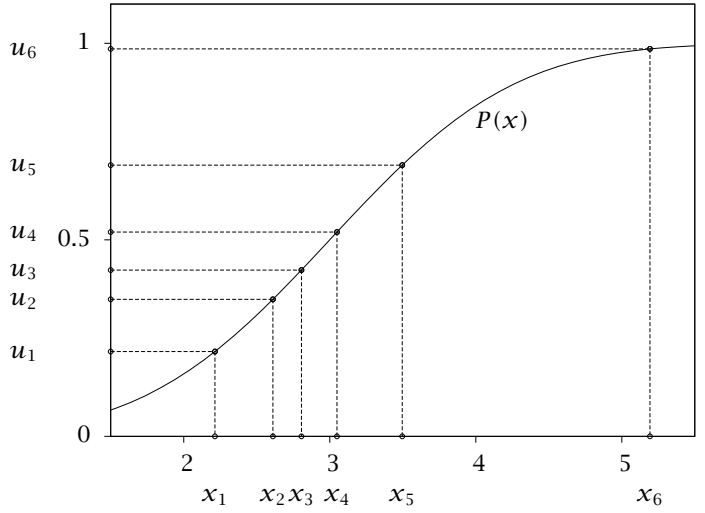

Figure 4.16: Six samples of the uniform density on $[0,1]$, u_i, and the corresponding samples of $p(x)$, x_i. The samples satisfy $x_i = P^{-1}(u_i)$.

The weighted sample of p is also unbiased for all sample sizes, which we can verify as follows

$$\begin{aligned}
\mathcal{E}_{\text{is}}\left(\overline{p}_s(x)\right) &= \mathcal{E}_{\text{is}}\left(\sum_i w_i \delta(x - x_i)\right) \\
&= \int p_{\text{is}}(x_1, \ldots, x_s) \sum_i w_i \delta(x - x_i) dx_1 \cdots dx_s \\
&= \int q(x_1) \cdots q(x_s) \sum_i w_i \delta(x - x_i) dx_1 \cdots dx_s \\
&= \sum_i \int q(x_i) w_i \delta(x - x_i) dx_i \prod_{j \neq i} \int q(x_j) dx_j \\
&= \sum_i \int q(x_i) \frac{1}{s} \frac{p(x_i)}{q(x_i)} \delta(x - x_i) dx_i \\
&= \frac{1}{s} \sum_i p(x) \\
\mathcal{E}_{\text{is}}\left(\overline{p}_s(x)\right) &= p(x)
\end{aligned}$$

Notice this result holds for all $s \geq 1$.

Using the same development, we can represent any function $h(x)$ (not necessarily a density) having the same support as $q(x)$ as a sampled function

$$\begin{aligned}
\overline{h}_s(x) &= \sum_{i=1}^{s} w_i \delta(x - x_i) \qquad w_i = \frac{1}{s} \frac{h(x_i)}{q(x_i)} \\
\lim_{s \to \infty} \overline{h}_s(x) &= h(x)
\end{aligned} \tag{4.46}$$

The next example demonstrates using importance sampling to generate samples of a multimodal density.

Example 4.33: Importance sampling of a multimodal density

Given the following bimodal distribution

$$p(x) = \frac{1}{2\sqrt{2\pi P_1}} e^{-(1/2)(x-m_1)^2/P_1} + \frac{1}{2\sqrt{2\pi P_2}} e^{-(1/2)(x-m_2)^2/P_2}$$
$$m_1 = -4 \quad m_2 = 4 \quad P_1 = P_2 = 1$$

generate samples using the following unimodal importance function

$$q(x) = \frac{1}{\sqrt{2\pi P}} e^{-(1/2)(x-m)^2/P} \qquad m = 0 \quad P = 4$$

Solution

Figure 4.17 shows the exact and sampled density of the importance function $q(x)$ using 5000 samples. The weighted density for $p(x)$ is shown in Figure 4.18. We obtain a good representation of the bimodal distribution with 5000 samples. Notice also that one should use a broad density for $q(x)$ to obtain sufficient samples in regions where $p(x)$ has significant probability. Using $q(x)$ with variance of $P = 1$ instead of $P = 4$ would require many more samples to obtain an accurate representation of $p(x)$. Of course we cannot choose $q(x)$ too broad or we sample the region of interest too sparsely. Choosing an appropriate importance function for an unknown $p(x)$ is naturally a significant challenge in many applications. □

Importance sampling when the density cannot be evaluated. In many applications we have a density $p(x)$ that is difficult to evaluate directly, but it can be expressed as

$$p(x) = \frac{h(x)}{\int h(x)dx} \qquad p(x) \propto h(x)$$

in which $h(x)$ is readily evaluated. We wish to avoid the task of integration of h to find the normalizing constant. Importance sampling can still be used to sample p in this case, but, as we discuss next, we lose the unbiased property of the sampled density for finite sample sizes. In this case, define the candidate sampled density as

$$\overline{p}_s(x) = \frac{q_s(x)}{d(s)}\frac{h(x)}{q(x)} \qquad d(s) = \frac{1}{s}\sum_j \frac{h(x_j)}{q(x_j)} \tag{4.47}$$

in which the samples are again chosen from the importance function $q(x)$. Summarizing, the candidate sampled density is

$$\overline{p}_s(x) = \sum_i w_i \delta(x - x_i)$$

$$p_{\text{is}}(x_i) = q(x_i) \qquad w_i = \frac{h(x_i)/q(x_i)}{\sum_j h(x_j)/q(x_j)} \quad i = 1, \ldots, s \tag{4.48}$$

Notice the weights are normalized in the case when we do not know the normalizing constant to convert from $h(x)$ to $p(x)$. We next show that $\overline{p}_s(x)$ converges to $p(x)$ as sample size increases (Smith and Gelfand,

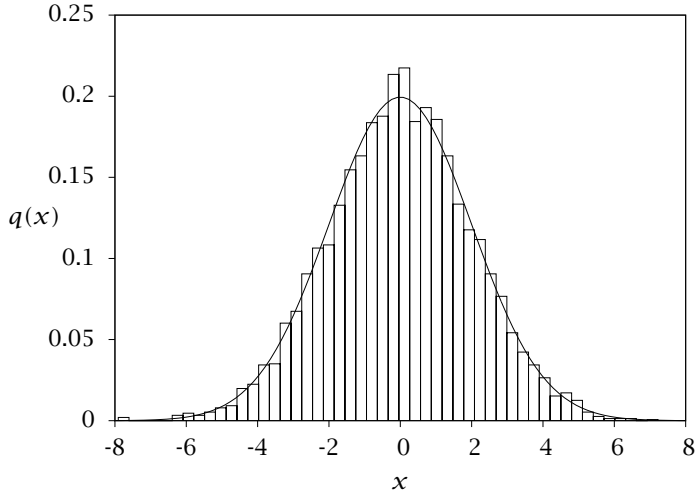

Figure 4.17: Importance function $q(x)$ and its histogram based on 5000 samples.

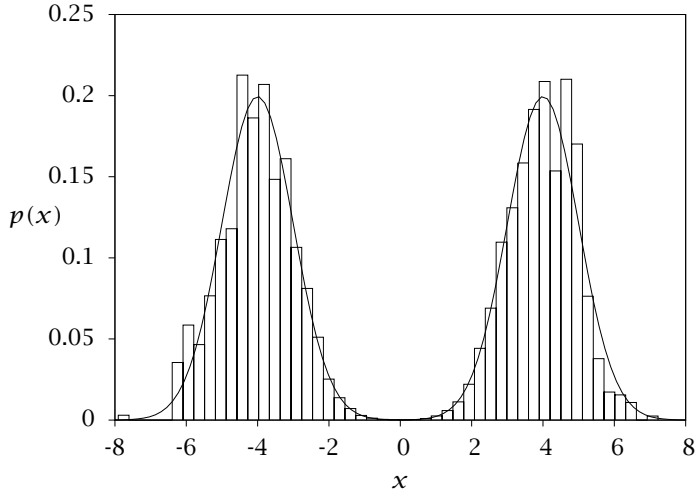

Figure 4.18: Exact density $p(x)$ and its histogram based on 5000 importance samples.

1992). First we express $d(s)$ as

$$\begin{aligned} d(s) &= \frac{1}{s}\sum_j \frac{h(x_j)}{q(x_j)} \\ &= \int_{-\infty}^{\infty} \frac{1}{s}\sum_j \frac{h(x_j)}{q(x_j)}\delta(x - x_j)dx \\ &= \int_{-\infty}^{\infty} \frac{1}{s}\sum_j \frac{h(x)}{q(x)}\delta(x - x_j)dx \\ d(s) &= \int_{-\infty}^{\infty} h_s(x)dx \end{aligned}$$

Exchanging the order of limit and integral and using (4.46) give

$$\lim_{s\to\infty} d(s) = \int_{-\infty}^{\infty} \lim_{s\to\infty} h_s(x)dx = \int_{-\infty}^{\infty} h(x)dx$$

Next we take the limit in (4.47) to obtain

$$\begin{aligned} \lim_{s\to\infty} \overline{p}_s(x) &= \lim_{s\to\infty} \frac{q_s(x)}{d(s)}\frac{h(x)}{q(x)} \\ &= \frac{\lim_{s\to\infty} q_s(x)}{\lim_{s\to\infty} d(s)}\frac{h(x)}{q(x)} \\ &= \frac{q(x)}{\int h(x)dx}\frac{h(x)}{q(x)} \\ &= \frac{h(x)}{\int h(x)dx} \\ \lim_{s\to\infty} \overline{p}_s(x) &= p(x) \end{aligned}$$

Notice the unbiased property no longer holds for a finite sample size. We can readily show

$$\mathcal{E}_{\text{is}}\left(\overline{p}_s(x)\right) \neq p(x) \qquad \text{for finite } s \tag{4.49}$$

For example, take $s = 1$. We have from (4.48) that $w_1 = 1$, and therefore

$$\begin{aligned} \mathcal{E}_{\text{is}}\left(\overline{p}_s(x)\right) &= \int p_{\text{is}}(x_1)w_1\delta(x - x_1)dx_1 \\ &= \int q(x_1)\delta(x - x_1)dx_1 \\ \mathcal{E}_{\text{is}}\left(\overline{p}_s(x)\right) &= q(x) \end{aligned}$$

and we see that the expectation of the sampling process with a single sample gives back the importance function $q(x)$ rather than the desired $p(x)$. Obviously we should choose many more samples than $s = 1$ for this case to reduce this bias. Consider the next example in which we use a large number of samples.

Example 4.34: Importance sampling of a multimodal function

We revisit Example 4.33 but use the following function $h(x)$

$$h(x) = e^{-(1/2)(x-m_1)^2/P_1} + e^{-(1/2)(x-m_2)^2/P_2}$$
$$m_1 = -4, m_2 = 4, P_1 = P_2 = 1$$

and we do not have the normalization constant available. We again generate samples using the following importance function

$$q(x) = \frac{1}{\sqrt{2\pi P}} e^{-(1/2)(x-m)^2/P} \qquad m = 0, P = 4$$

Solution

The exact and sampled density of the importance function $q(x)$ using 5000 samples is the same as Figure 4.17. The weighted density for $p(x)$ is shown in Figure 4.19. Comparing Figure 4.19 to Figure 4.18 shows the representation of the bimodal distribution with 5000 samples using $h(x)$ is of comparable quality to the one using $p(x)$ itself. The bias is not noticeable using 5000 samples. □

Weighted importance sampling. In applications of importance sampling to state estimation, the importance function is often available as a *weighted* sample in which the weights are not all equal. Therefore, as a final topic in importance sampling, we consider the case in which a weighted sample of the importance function is available

$$q_s(x) = \sum_{i=1}^{s} w_i^- \delta(x - x_i) \qquad w_i^- \geq 0$$

We have the two cases of interest covered previously.

(a) We can evaluate $p(x)$. For this case we define the sampled density for $p(x)$ as

$$\overline{p}_s(x) = \sum_{i=1}^{s} w_i \delta(x - x_i) \qquad w_i = w_i^- \frac{p(x_i)}{q(x_i)}$$

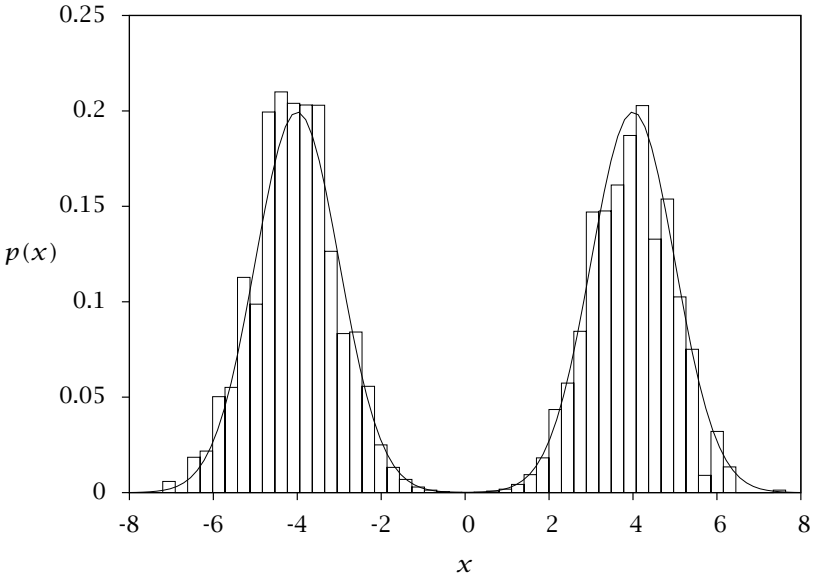

Figure 4.19: Exact density $p(x)$ and its histogram based on 5000 importance samples evaluating $h(x)$ in place of $p(x) = h(x)/\int h(x)dx$.

For this case, the sampled density is unbiased for all sample sizes and converges to $p(x)$ as the sample size increases.

(b) We cannot evaluate $p(x)$, but can evaluate only $h(x)$ with $p(x) = h(x)/\int h(x)dx$. For this case, we define the sampled density as

$$\overline{p}_s(x) = \sum_{i=1}^{s} \overline{w}_i \delta(x - x_i)$$

$$w_i = w_i^- \frac{h(x_i)}{q(x_i)} \qquad \overline{w}_i = \frac{w_i}{\sum_j w_j} \quad (4.50)$$

For this case, the sampled density is biased for all finite sample sizes, but converges to $p(x)$ as the sample size increases.

The proofs of these properties are covered in Exercises 4.21 and 4.22.

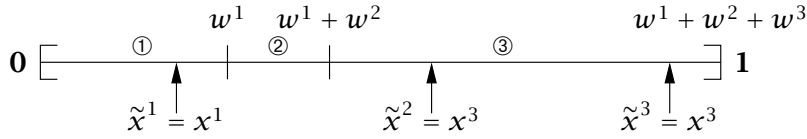

Figure 4.20: Interval $[0,1]$ partitioned by original sample weights, w_i. The arrows depict the outcome of drawing three uniformly distributed random numbers. For the case depicted here, the new samples are $\widetilde{x}_1 = x_1$, $\widetilde{x}_2 = x_3$, $\widetilde{x}_3 = x_3$ because the first arrow falls into the first interval and the other two arrows both fall into the third interval. Sample x_2 is discarded and sample x_3 is repeated twice in the resample. The new sample's weights are simply $\widetilde{w}^1 = \widetilde{w}^2 = \widetilde{w}^3 = 1/3$.

4.7.3 Resampling

Consider a set of samples at $x = x_i$, $i = 1, \ldots s$ and associated normalized weights w_i, $w_i \geq 0$, $\sum_{i=1}^{s} w_i = 1$. Define a probability density using these samples and weights by

$$p(x) = \sum_{i=1}^{s} w_i \delta(x - x_i)$$

Consider any function $f(x)$ defined on a set that contains the samples, x_i. Then the integral of f using the defined density is

$$\int f(x)p(x)dx = \sum_{i=1}^{s} w_i f(x_i) = \sum_{i=1}^{s} w_i f_i$$

in which $f_i = f(x_i)$. We now consider a resampling procedure that produces a new set of samples $\widetilde{x}_i$ with new weights $\widetilde{w}_i$. The resampling procedure is depicted in Figure 4.20 for the case $s = 3$. We partition the interval $[0,1]$ into s intervals using the original sample weights, w_i, as shown in Figure 4.20, in which the ith interval has width w_i. To choose s resamples, we generate s random numbers from a uniform distribution on $[0,1]$. Denote these random numbers as u_i, $i = 1, \ldots, s$. For each i, we find the interval in which the drawn random number falls. Denote the interval number as $m(i)$, defined by the relation

$$0 \leq w_1 + w_2 + \cdots + w_{m(i)-1} \leq u_i \leq w_1 + w_2 + \cdots + w_{m(i)} \leq 1$$

We then choose as resamples

$$\tilde{x}_i = x_{m(i)} \qquad i = 1, \ldots s$$

The resampling selects the new sample locations $\tilde{x}$ in regions of high density. We set all the $\tilde{w}$ weights equal to $1/s$. The result illustrated in Figure 4.20 is summarized in the following table

Original sample		Resample	
State	Weight	State	Weight
x_1	$w_1 = \frac{3}{10}$	$\tilde{x}_1 = x_1$	$\tilde{w}_1 = \frac{1}{3}$
x_2	$w_2 = \frac{1}{10}$	$\tilde{x}_2 = x_3$	$\tilde{w}_2 = \frac{1}{3}$
x_3	$w_3 = \frac{6}{10}$	$\tilde{x}_3 = x_3$	$\tilde{w}_3 = \frac{1}{3}$

The properties of the resamples are summarized by

$$p_{\text{re}}(\tilde{x}_i) = \begin{cases} w_j & \tilde{x}_i = x_j \\ 0 & \tilde{x}_i \neq x_j \end{cases}$$
$$\tilde{w}_i = 1/s \quad \text{all } i$$

We can associate with each resampling a sampled probability density

$$\tilde{p}(x) = \sum_{i=1}^{s} \tilde{w}_i \delta(x - \tilde{x}_i)$$

The resampled density is clearly *not the same* as the original sampled density. It is likely that we have moved many of the new samples to places where the original density has large weights. But by resampling in the fashion described here, we have not introduced bias into the estimates.

Consider taking many such resamples. We can calculate for each of these resamples a value of the integral of f as follows

$$\int f(x)\tilde{p}(x)dx = \sum_{i=1}^{s} \tilde{w}_i f(\tilde{x}_i)$$

To show this resampling procedure is valid, we show that the average over these values of the f integrals with $\tilde{p}(x)$ is equal to the original value of the integral using $p(x)$. We state this result for the resampling procedure described previously as the following theorem.

Theorem 4.35 (Resampling). *Consider a sampled density $p(x)$ with s samples at $x = x_i$ and associated weights w_i*

$$p(x) = \sum_{i=1}^{s} w_i \delta(x - x_i) \qquad w_i \geq 0, \qquad \sum_{i=1}^{s} w_i = 1$$

Consider the resampling procedure that gives a resampled density

$$\tilde{p}(x) = \sum_{i=1}^{s} \tilde{w}_i \delta(x - \tilde{x}_i)$$

in which the $\tilde{x}_i$ are chosen according to resample probability p_{re}

$$p_{\text{re}}(\tilde{x}_i) = \begin{cases} w_j & \tilde{x}_i = x_j \\ 0 & \tilde{x}_i \neq x_j \end{cases}$$

and with uniform weights $\tilde{w}_i = 1/s$. Consider a function $f(\cdot)$ defined on a set X containing the points x_i.

With this resampling procedure, the expectation over resampling of any integral of the resampled density is equal to that same integral of the original density

$$\mathcal{E}_{\text{re}}\left(\int f(x)\tilde{p}(x)dx\right) = \int f(x)p(x)dx \qquad \text{all } f$$

The proof of this theorem is discussed in Exercise 4.16. To get a feel for why this resampling procedure works, however, consider the case $s = 2$. There are four possible outcomes of $\tilde{x}_1, \tilde{x}_2$ in resampling. Because of the resampling procedure, the random variables $\tilde{x}_i$ and $\tilde{x}_j$, $j \neq i$ are independent, and their joint density is

$$p_{\text{re}}(\tilde{x}_1, \tilde{x}_2) = \begin{cases} w_1^2 & \tilde{x}_1 = x_1, \tilde{x}_2 = x_1 \\ w_1 w_2 & \tilde{x}_1 = x_1, \tilde{x}_2 = x_2 \\ w_2 w_1 & \tilde{x}_1 = x_2, \tilde{x}_2 = x_1 \\ w_2^2 & \tilde{x}_1 = x_2, \tilde{x}_2 = x_2 \end{cases}$$

The values of the integral of f for each of these four outcomes is

$$\sum_{i=1}^{2} \tilde{w}_i f(\tilde{x}_i) = \begin{cases} \frac{1}{2}(f_1 + f_1) & \tilde{x}_1 = x_1, \tilde{x}_2 = x_1 \\ \frac{1}{2}(f_1 + f_2) & \tilde{x}_1 = x_1, \tilde{x}_2 = x_2 \\ \frac{1}{2}(f_2 + f_1) & \tilde{x}_1 = x_2, \tilde{x}_2 = x_1 \\ \frac{1}{2}(f_2 + f_2) & \tilde{x}_1 = x_2, \tilde{x}_2 = x_2 \end{cases}$$

Notice there are only three different values for the integral of f. Next, calculating the expectation over the resampling process gives

$$\begin{aligned}\mathcal{E}_{\text{re}}\left(\sum_{i=1}^{2}\widetilde{w}_i f(\widetilde{x}_i)\right) &= w_1^2 f_1 + w_1 w_2 (f_1 + f_2) + w_2^2 f_2^2 \\ &= (w_1^2 + w_1 w_2) f_1 + (w_1 w_2 + w_2^2) f_2 \\ &= w_1 (w_1 + w_2) f_1 + w_2 (w_1 + w_2) f_2 \\ &= w_1 f_1 + w_2 f_2 \\ &= \int f(x) p(x) dx\end{aligned}$$

and the conclusion of the theorem is established for $s = 2$.

One can also change the total number of samples in resampling without changing the conclusions of Theorem 4.35. Exercise 4.17 explores this issue in detail. In many applications of sampling, we use the resampling process to discard samples with excessively small weights in order to reduce the storage requirements and computational burden associated with a large number of samples.

To make this discussion explicit, consider again the bimodal distribution of Example 4.33 shown in Figure 4.18 that was sampled using importance sampling. Many of the samples are located in the interval $[-1, 1]$ because the importance function q has large density in this interval. In fact, 1964 of the 5000 samples fall in this interval given the random sample corresponding to Figure 4.18. But notice the weights in this interval are quite small. If we resample p, we can retain the accuracy with many fewer samples as we show in the next example.

Example 4.36: Resampling a bimodal density

Consider the bimodal sampled density obtained in Example 4.33 using importance sampling. Resample this sampled density with 500 samples. Compare the accuracy to the original density with 5000 samples.

Solution

The histogram of the resampled density with 500 samples is shown in Figure 4.21. The weights in the resampled density are all equal to 1/500. Notice that the accuracy is comparable to Figure 4.18 with one tenth as many samples because most of the samples with small weights have been removed by the resampling process. In fact, none of the 500 resamples fall in the interval $[-1, 1]$. □

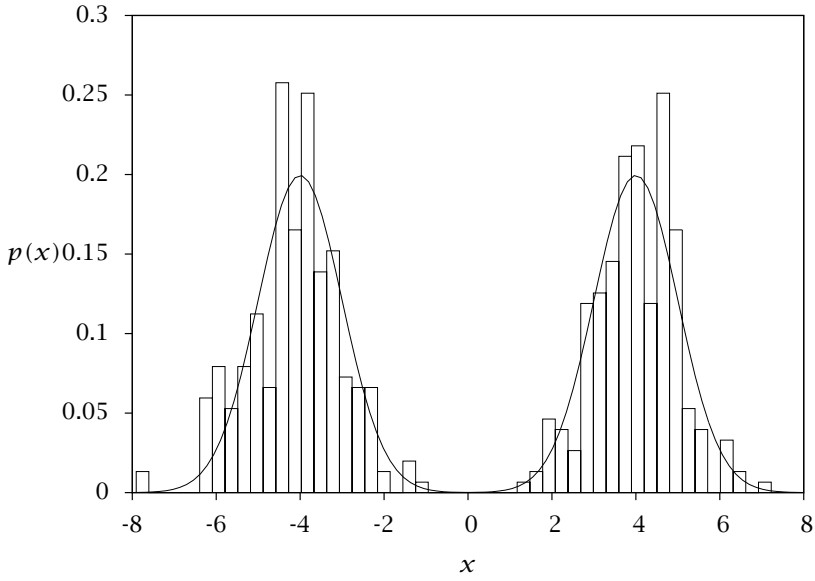

Figure 4.21: Resampled density of Example 4.33 using 500 samples. Compare to Figure 4.18 that uses 5000 samples.

4.7.4 The Simplest Particle Filter

Next we implement these sampling ideas for state estimation. This first version follows the approach given by Gordon, Salmond, and Smith (1993). In state estimation, the density $p(x(k)|\mathbf{y}(k))$ contains the information of most interest. The model is of the form

$$\begin{aligned} x(k+1) &= f(x(k), n(k)) \\ y(k) &= h(x(k), m(k)) \end{aligned}$$

in which f is a possibly nonlinear function of the state and process noise, n, and h is a possibly nonlinear function of the state and measurement noise, m. We assume that the densities of m, n and $x(0)$ are available. To start things off, first assume the conditional density $p(x(k)|\mathbf{y}(k))$ is available as a sampled density

$$p(x(k)|\mathbf{y}(k)) = \sum_{i=1}^{s} w_i(k)\delta(x(k) - x_i(k))$$

and we wish to find samples for $p(x(k+1)|\mathbf{y}(k))$. The state evolution can be considered a noninvertible transformation from $x(k), n(k)$ to $x(k+1)$, in which $n(k)$ is statistically independent of $x(k)$ and $\mathbf{y}(k)$. We generate s samples of $n(k)$, call these $n_i(k)$, and we have s samples of the conditional density $p(x(k), n(k)|\mathbf{y}(k))$ given by $\{x_i(k), n_i(k)\}, i = 1, \ldots, s$. As shown in Section 4.7.1, the sampled density of $p(x(k+1)|\mathbf{y}(k))$ is given by

$$p(x(k+1)|\mathbf{y}(k)) = \{x_i(k+1), w_i^-(k+1)\}$$
$$x_i(k+1) = f(x_i(k), n_i(k)) \qquad w_i^-(k+1) = w_i(k)$$

Next, given the sampled density for the conditional density $p(x(k)|\mathbf{y}(k-1))$

$$p(x(k)|\mathbf{y}(k-1)) = \sum_{i=1}^{s} w_i^-(k)\delta(x(k) - x_i(k))$$

we add the measurement $y(k)$ to obtain the sampled density $p(x(k)|\mathbf{y}(k))$. Notice that $\mathbf{y}(k) = (y(k), \mathbf{y}(k-1))$ and use the relationship (see Exercise 1.47)

$$p_{A|B,C}(a|b,c) = p_{C|A,B}(c|a,b)\frac{p_{A|B}(a|b)}{p_{C|B}(c|b)}$$

to obtain

$$p(x(k)|\mathbf{y}(k)) = \frac{p(y(k)|x(k), \mathbf{y}(k-1))p(x(k)|\mathbf{y}(k-1))}{p(y(k)|\mathbf{y}(k-1))}$$

Because the process is Markov, $p(y(k)|x(k), \mathbf{y}(k-1)) = p(y(k)|x(k))$, and we have

$$p(x(k)|\mathbf{y}(k)) = \frac{p(y(k)|x(k))p(x(k)|\mathbf{y}(k-1))}{p(y(k)|\mathbf{y}(k-1))}$$

The density of interest is in the form

$$p(x(k)|\mathbf{y}(k)) = g(x(k))p(x(k)|\mathbf{y}(k-1))$$
$$g(x(k)) = \frac{p(y(k)|x(k))}{p(y(k)|\mathbf{y}(k-1))}$$

and we have a sampled density for $p(x(k)|\mathbf{y}(k-1))$. If we could conveniently evaluate g, then we could obtain a sampled density using the product rule given in (4.36)

$$p(x(k)|\mathbf{y}(k)) = \{x_i(k), \tilde{w}_i(k)\}$$

in which

$$\tilde{w}_i(k) = w_i^-(k)\frac{p(y(k)|x_i(k))}{p(y(k)|\mathbf{y}(k-1))} \tag{4.51}$$

This method would provide an unbiased sampled density, but it is inconvenient to evaluate the term $p(y(k)|\mathbf{y}(k-1))$. So we consider an alternative in which the available sampled density is used as a weighted importance function for the conditional density of interest. If we define the importance function $q(x(k)) = p(x(k)|\mathbf{y}(k-1))$, then the conditional density is of the form

$$p(x(k)|\mathbf{y}(k)) = \frac{h(x(k))}{\int h(x(k))dx(k)}$$
$$h(x(k)) = p(y(k)|x(k))p(x(k)|\mathbf{y}(k-1))$$

We then use weighted importance sampling and (4.50) to obtain

$$p(x(k)|\mathbf{y}(k)) = \{x_i(k), \overline{w}_i(k)\} \qquad w_i(k) = w_i^-(k)p(y(k)|x_i(k))$$
$$\overline{w}_i(k) = \frac{w_i(k)}{\sum_j w_j(k)}$$

By using this form of importance sampling, the sampled density is biased for all finite sample sizes, but converges to $p(x(k)|\mathbf{y}(k))$ as the sample size increases.

Summary. Starting with s samples of $p(n(k))$ and s samples of $p(x(0))$, we assume that we can evaluate $p(y(k)|x(k))$ using the measurement equation. The iteration for the simple particle filter is summarized by the following recursion.

$$\begin{array}{rcl}
p(x(0)) & = & \{x_i(0), w_i(0) = 1/s\} \\
p(x(k)|\mathbf{y}(k)) & = & \{x_i(k), \overline{w}_i(k)\} \\
w_i(k) & = & \overline{w}_i(k-1)p(y(k)|x_i(k)) \\
\overline{w}_i(k) & = & \dfrac{w_i(k)}{\sum_j w_j(k)} \\
p(x(k+1)|\mathbf{y}(k)) & = & \{x_i(k+1), \overline{w}_i(k)\} \\
x_i(k+1) & = & f(x_i(k), n_i(k))
\end{array}$$

The sampled density of the simplest particle filter converges to the conditional density $p(x(k)|\mathbf{y}(k))$ in the limit of large sample size. The sampled density is biased for all finite sample sizes.

Analysis of the simplest particle filter. The simplest particle filter has well-known weaknesses that limit its use as a practical method for state estimation. The variances in both the particle locations and the filter weights can increase without bound as time increases and more measurements become available. Consider first the particle locations. For even the simple linear model with Gaussian noise, we have

$$\begin{aligned} x_i(k+1) &= Ax_i(k) + Bu(k) + Gw_i(k) \\ x_i(0) \sim N(\overline{x}(0), Q_0) &\qquad w_i(k) \sim N(0, Q) \end{aligned}$$

which gives the following statistical properties for the particle locations

$$\begin{aligned} x_i(k) &\sim N(\overline{x}(k), \overline{P}(k)) \quad i = 1, \ldots, s \\ \overline{x}(k) &= A\overline{x}(k-1) + Bu(k) \\ \overline{P}(k) &= A\overline{P}(k-1)A' + GQG' \end{aligned} \tag{4.52}$$

If A is not strictly stable, the variance of the samples locations, $P(k)$, increases without bound despite the availability of the measurement at every time. In this simplest particle filter, one is expecting the particle weights to carry all the information in the measurements. As we will see in the upcoming example, this idea does not work and after a few time iterations the resulting state estimates are useless.

To analyze the variance of the resulting particle weights, it is helpful to define the following statistical properties and establish the following identities. Consider two random variables A and B. Conditional expectations of A and functions of A and conditional variance of A are defined as

$$\begin{aligned} \mathcal{E}(A|B) &:= \int p_{A|B}(a|b)a\,da \\ \mathcal{E}(A^2|B) &:= \int p_{A|B}(a|b)a^2\,da \\ \mathcal{E}(g(A)|B) &:= \int p_{A|B}(a|b)g(a)\,da \\ \mathrm{var}(A|B) &:= \mathcal{E}(A^2|B) - \mathcal{E}^2(A|B) \end{aligned}$$

in which we assume as usual that B's marginal is nonzero so the conditional density is well defined. We derive a first useful identity

$$E(E(g(A)|B)) = E(g(A)) \tag{4.53}$$

as follows

$$\begin{aligned}\mathcal{E}(\mathcal{E}(g(A)|B)) &= \int p_B(b) \int p_{A|B}(a|b) g(a) da db \\ &= \int p_B(b) \int \frac{p_{A,B}(a,b)}{p_B(b)} g(a) da db \\ &= \iint p_{A,B}(a,b) g(a) da db \\ &= \int p_A(a) g(a) da \\ &= \mathcal{E}(g(A))\end{aligned}$$

We require a second identity

$$\text{var}(A) = \mathcal{E}(\text{var}(A|B)) + \text{var}(\mathcal{E}(A|B)) \tag{4.54}$$

which is known as the conditional variance formula or the law of total variance. We establish this identity as follows. Starting with the definition of variance

$$\text{var}(A) = \mathcal{E}(A^2) - \mathcal{E}^2(A)$$

we use (4.53) to obtain

$$\text{var}(A) = \mathcal{E}(\mathcal{E}(A^2|B)) - \mathcal{E}^2(\mathcal{E}(A|B))$$

Using the definition of variance on the first term on the right-hand side gives

$$\begin{aligned}\text{var}(A) &= \mathcal{E}\left(\text{var}(A|B) + \mathcal{E}^2(A|B)\right) - \mathcal{E}^2(\mathcal{E}(A|B)) \\ &= \mathcal{E}(\text{var}(A|B)) + \mathcal{E}(\mathcal{E}^2(A|B)) - \mathcal{E}^2(\mathcal{E}(A|B))\end{aligned}$$

and using the definition of variance again on the last two terms on the right-hand side gives

$$\text{var}(A) = \mathcal{E}(\text{var}(A|B)) + \text{var}(\mathcal{E}(A|B))$$

which establishes the result. Notice that since variance is nonnegative, this result also implies the inequality

$$\text{var}(\mathcal{E}(A|B)) \leq \text{var}(A)$$

which shows that the conditional expectation of random variable A has less variance than A itself.

We proceed to analyze the simplest particle filter. Actually we analyze the behavior of the weights for the idealized, unbiased case given by (4.51)

$$w_i(k) = w_i(k-1)\frac{p(y(k)|x_i(k))}{p(y(k)|\mathbf{y}(k-1))}$$

in which we consider the random variable $w_i(k)$ to be a function of the random variables $y(k), x_i(k)$. We next consider the conditional density of the random variables $y(k), x_i(k)$ relative to the previous samples $x_i(k-1)$, and the data $\mathbf{y}(k-1)$. We have

$$\begin{aligned} &p(y(k), x_i(k)|\mathbf{y}(k-1), x_i(k-1)) \\ &\quad = p(y(k)|\mathbf{y}(k-1), x_i(k-1))p(x_i(k)|\mathbf{y}(k-1), x_i(k-1)) \\ &\quad = p(y(k)|\mathbf{y}(k-1))p(x_i(k)|x_i(k-1)) \end{aligned}$$

The first equation results from the statistical independence of $y(k)$ and $x_i(k)$, and the second results from the sampling procedure used to generate $x_i(k)$ given $x_i(k-1)$. Note that in the next section, we use a different sampling procedure in which $x_i(k)$ depends on both the new data $y(k)$ as well as the $x_i(k-1)$. Now we take the expectation of the weights at time k conditional on the previous samples and previous measurement trajectory

$$\begin{aligned} &E(w_i(k)|x_i(k-1), \mathbf{y}(k-1)) \\ &\quad = \iint w_i(k)p(y(k), x_i(k)|x_i(k-1), \mathbf{y}(k-1))dx_i(k)dy(k) \\ &\quad = \iint w_i(k)p(y(k)|\mathbf{y}(k-1))p(x_i(k)|x_i(k-1))dx_i(k)dy(k) \end{aligned}$$

Substituting the weight recursion and simplifying yields

$$\begin{aligned} &E(w_i(k)|x_i(k-1), \mathbf{y}(k-1)) \\ &\qquad = \iint w_i(k-1)\frac{p(y(k)|x_i(k))}{p(y(k)|\mathbf{y}(k-1))} \\ &\qquad p(y(k)|\mathbf{y}(k-1))p(x_i(k)|x_i(k-1))dx_i(k)dy(k) \end{aligned}$$

$$\begin{aligned} &E(w_i(k)|x_i(k-1), \mathbf{y}(k-1)) \\ &\quad = \iint w_i(k-1)p(y(k)|x_i(k))p(x_i(k)|x_i(k-1))dx_i(k)dy(k) \end{aligned}$$

Taking $w_i(k-1)$ outside the integral and performing the integral over $x_i(k)$ and then $y(k)$ gives

$$E(w_i(k)|x_i(k-1),\mathbf{y}(k-1)) = w_i(k-1)\int p(y(k)|x_i(k-1))dy(k)$$
$$E(w_i(k)|x_i(k-1),\mathbf{y}(k-1)) = w_i(k-1)$$

Taking the variance of both sides and using the conditional variance formula (4.54) gives

$$\text{var}(E(w_i(k)|x_i(k-1),\mathbf{y}(k-1))) = \text{var}(w_i(k-1))$$
$$\text{var}(w_i(k)) - E(\text{var}(w_i(k)|x_i(k-1),\mathbf{y}(k-1))) = \text{var}(w_i(k-1))$$

Again, noting that variance is nonnegative gives the inequality

$$\text{var}(w_i(k)) \geq \text{var}(w_i(k-1))$$

and we see that the variance for the unbiased weights of the simplest particle filter increases with time.

Next we present two examples that show the serious practical limitations of the simplest particle filter and the simplest particle filter with resampling.

Example 4.37: What's wrong with the simplest particle filter?

Consider the following linear system with Gaussian noise.

$$A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \quad \theta = 6 \qquad C = \begin{bmatrix} 0.5 & 0.25 \end{bmatrix} \qquad G = I \qquad B = I$$
$$\overline{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad Q_0 = \frac{1}{4}\begin{bmatrix} 7 & 5 \\ 5 & 7 \end{bmatrix} \qquad Q = 0.01\, I \qquad R = 0.01$$
$$u(0,1,\cdots,4) = \begin{bmatrix} 7 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

(a) Plot the particle locations versus time from $k = 0$ to $k = 5$. Plot also the 95% contour of the true conditional density $p(x(k)|\mathbf{y}(k))$. Discuss the locations of the particles using the simplest particle filter.

(b) Write out the recursions for the conditional density of the particle locations $p(x_i(k)|\mathbf{y}(k))$ as well as the true conditional density $p(x(k)|\mathbf{y}(k))$. Discuss the differences.

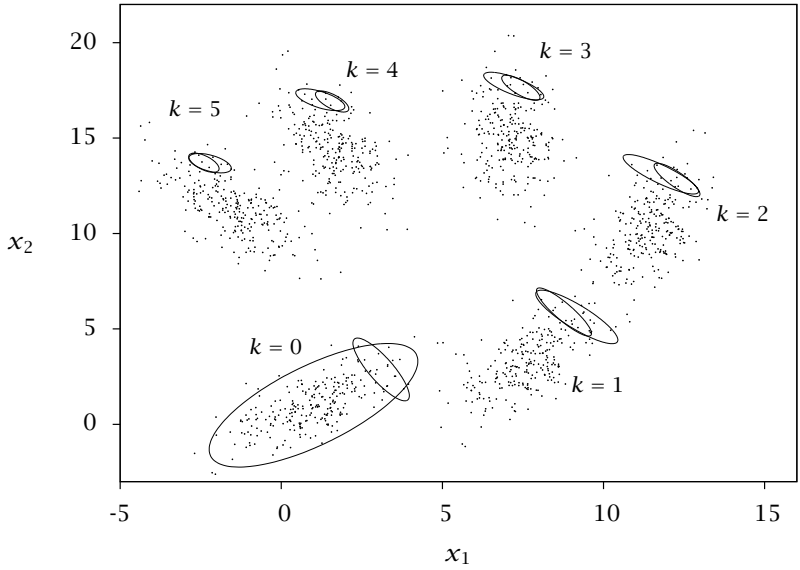

Figure 4.22: Particles' locations versus time for the simplest particle filter; 250 particles. Ellipses show the 95% contour of the true conditional densities before and after measurement.

Solution

(a) The samples and 95% conditional density contour are shown in Figure 4.22. The particles are located properly at $k = 0$ and about 95% of them are inside the state's initial density. But notice that the particles spread out quickly and few particles remain inside the 95% contour of the true conditional density after a few time steps.

(b) The true conditional density is the normal density given by the time-varying Kalman filter recursion. The conditional density of the particle location is given by (4.52) and the samples are identi-

cally distributed

$$
\begin{aligned}
p(x(k)|\mathbf{y}(k)) &\sim N(\hat{x}(k), P(k)) \\
\hat{x}(k+1) &= A\hat{x}(k) + Bu(k) + L(k)(y(k) - C(A\hat{x}(k) + Bu(k))) \\
P(k+1) &= AP(k)A' + GQG' - L(k+1)C(AP(k)A' + GQG') \\
p(x_i(k)|\mathbf{y}(k)) &\sim N(\overline{x}(k), \overline{P}(k)), \quad i = 1, \ldots, s \\
\overline{x}(k+1) &= A\overline{x}(k) + Bu(k) \\
\overline{P}(k+1) &= A\overline{P}(k)A' + GQG'
\end{aligned}
$$

The major differences are underlined. Notice that the mean of the particle samples is independent of $\mathbf{y}(k)$, which causes the samples to drift away from the conditional density's mean with time. Notice that the covariance does not have the reduction term present in the Kalman filter, which causes the variance of the particles to increase with time. Therefore, due to the missing underlined terms, the mean of the samples drifts and the variance increases with time. The particle weights cannot compensate for the inaccurate placement of the particles, and the state estimates from the simplest particle filter are not useful after a few time iterations.

□

Example 4.38: Can resampling fix the simplest particle filter?

Repeat the simulation of Example 4.37, but use resampling after each time step. Discuss the differences.

Solution

Applying the resampling strategy gives the results in Figure 4.23. Notice that resampling prevents the samples from drifting away from the mean of the conditional density. Resampling maintains a high concentration of particles in the 95% probability ellipse. If we repeat this simulation 500 times and compute the fraction of particles within the conditional density's 95% probability contour, we obtain the results shown in Figure 4.24. Notice the dramatic improvement. Without resampling, fewer than 10% of the particles are in the 95% confidence ellipse after only five time steps. With resampling, about 80% of the samples are inside the 95% confidence ellipse. There is one caution against resampling too frequently, however. If the measurement has a small covariance, then the weights computed from

$$w_i(k) = w_i(k-1)p(y(k)|x_i(i))$$

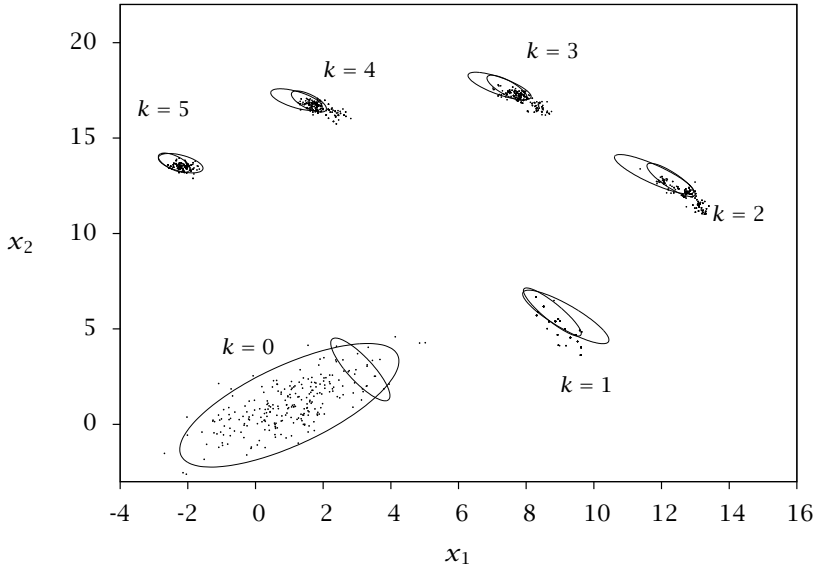

Figure 4.23: Particles' locations versus time for the simplest particle filter with resampling; 250 particles. Ellipses show the 95% contour of the true conditional densities before and after measurement.

will be dominated by only a few particles whose prediction of y is closest to the measurement. Resampling in this situation gives only those few particles repeated many times in the resample. For a sufficiently small covariance, this phenomenon can produce a single x_i value in the resample. This phenomenon is known as sample *impoverishment* (Doucet, Godsill, and Andrieu, 2000; Rawlings and Bakshi, 2006). □

4.7.5 A Particle Filter Based on Importance Sampling

Motivated by the drawbacks of the simplest particle filter of the previous section, researchers have developed alternatives based on a more flexible importance function (Arulampalam, Maskell, Gordon, and Clapp, 2002). We present this approach next. Rather than start with the statistical property of most interest, $p(x(k)|\mathbf{y}(k))$, consider instead the density of the entire *trajectory* of states conditioned on the measurements, $p(\mathbf{x}(k)|\mathbf{y}(k))$, as we did in moving horizon estimation. Our first

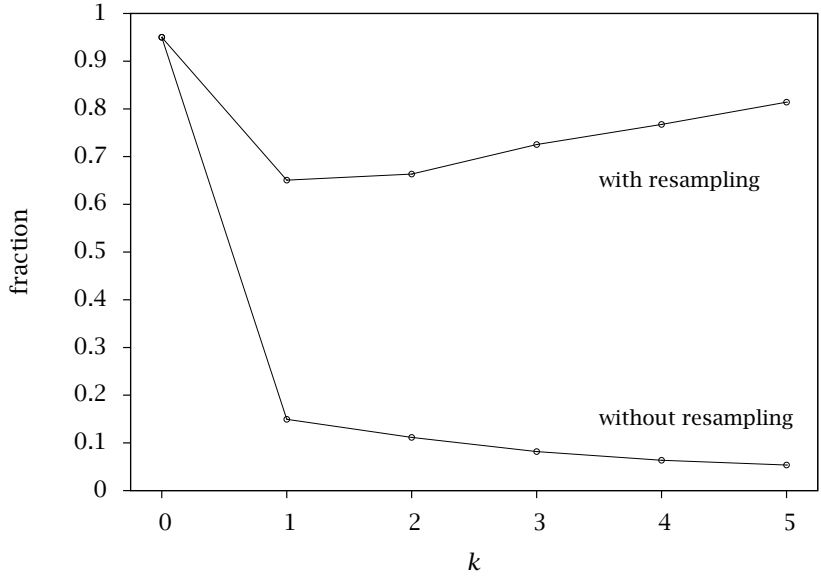

Figure 4.24: Fraction of particles inside the 95% contour of the true conditional density versus time; with and without resampling; average of 500 runs.

objective then is to obtain samples of $p(\mathbf{x}(k+1)|\mathbf{y}(k+1))$ from samples of $p(\mathbf{x}(k)|\mathbf{y}(k))$ and the model. We use importance sampling to accomplish this objective. Assume we have s weighted samples of the trajectory conditioned on measurements up to time k

$$p(\mathbf{x}(k)|\mathbf{y}(k)) = \{\mathbf{x}_i(k), \overline{w}_i(k)\} \quad i = 1, \ldots, s$$

in which the samples have been drawn from an importance function q, whose properties will be chosen as we proceed further. The weights $\overline{w}_i(k)$ are given by

$$\begin{aligned} w_i(k) &= \frac{h(\mathbf{x}_i(k))}{q(\mathbf{x}_i(k)|\mathbf{y}(k))} \\ p(\mathbf{x}_i(k)|\mathbf{y}(k)) &= \frac{h(\mathbf{x}_i(k))}{\int h(\mathbf{x}_i(k))d\mathbf{x}_i(k)} \\ \overline{w}_i(k) &= \frac{w_i(k)}{\sum_j w_j(k)} \end{aligned}$$

Notice $\mathbf{x}_i(k)$ is a set of ks n-vector samples, and, as in full information estimation, the storage requirements grow linearly with time. We remove this drawback subsequently, but for now we wish to obtain samples of $p(\mathbf{x}(k+1)|\mathbf{y}(k+1))$ in which $\mathbf{x}(k+1) = \{x(k+1), \mathbf{x}(k)\}$ and $\mathbf{y}(k+1) = \{y(k+1), \mathbf{y}(k)\}$. We start with

$$p(\mathbf{x}(k+1)|\mathbf{y}(k+1)) = \frac{p(y(k+1)|\mathbf{x}(k+1))p(\mathbf{x}(k+1)|\mathbf{y}(k))}{p(y(k+1)|\mathbf{y}(k))} \quad (4.55)$$

in which we have used the second identity in Exercise 1.47 and the Markov property, which implies

$$p(y(k+1)|\mathbf{x}(k+1), y(k)) = p(y(k+1)|\mathbf{x}(k+1))$$

Again, because the process is Markov $p(y(k+1)|\mathbf{x}(k+1)) = p(y(k+1)|x(k+1))$. We next use the identity $p_{A,B|C}(a,b|c) = p_{A|B,C}(a|b,c)p_{B|C}(b|c)$ (see Exercise 1.46) and obtain

$$p(\mathbf{x}(k+1)|\mathbf{y}(k)) = p(x(k+1)|\mathbf{x}(k), \mathbf{y}(k))p(\mathbf{x}(k)|\mathbf{y}(k))$$

Again using the Markov property in this equation, we know $p(x(k+1)|\mathbf{x}(k), \mathbf{y}(k)) = p(x(k+1)|x(k))$ and therefore

$$p(\mathbf{x}(k+1)|\mathbf{y}(k)) = p(x(k+1)|x(k))p(\mathbf{x}(k)|\mathbf{y}(k))$$

Substituting these relations into (4.55) gives

$$p(\mathbf{x}(k+1)|\mathbf{y}(k+1)) = \frac{p(y(k+1)|x(k+1))p(x(k+1)|x(k))}{p(y(k+1)|\mathbf{y}(k))} p(\mathbf{x}(k)|\mathbf{y}(k)) \quad (4.56)$$

We use importance sampling to sample this density. Notice the denominator does not depend on $\mathbf{x}(k+1)$ and is therefore not required when using importance sampling. We use instead

$$p(\mathbf{x}(k+1)|\mathbf{y}(k+1)) = \frac{h(\mathbf{x}(k+1))}{\int h(\mathbf{x}(k+1))d\mathbf{x}(k+1)}$$

$$h(\mathbf{x}(k+1)) = p(y(k+1)|x(k+1))p(x(k+1)|x(k))p(\mathbf{x}(k)|\mathbf{y}(k)) \quad (4.57)$$

Note also that using importance sampling here when we do not wish to evaluate the normalizing constant introduces bias for finite sample size

as stated in (4.49). We now state the two properties of q that provide a convenient importance function

$$q(x(k+1)|\mathbf{x}(k), \mathbf{y}(k+1)) = q(x(k+1)|x(k), y(k+1))$$

$$q(\mathbf{x}(k+1)|\mathbf{y}(k+1)) = q(x(k+1)|x(k), y(k+1))\, q(\mathbf{x}(k)|\mathbf{y}(k)) \quad (4.58)$$

The first property of q is satisfied also by the density p, so it is not unusual to pick an importance function to share this behavior. The second property is *not* satisfied by the density, however, and it is chosen strictly for convenience; it allows a recursive evaluation of q at time $k+1$ from the value at time k. See Exercise 4.18 for further discussion of this point.

Next we need to generate the samples of $q(\mathbf{x}(k+1)|\mathbf{y}(k+1))$. Given the second property in (4.58), we have

$$\begin{aligned} q(\mathbf{x}(k+1)|\mathbf{y}(k+1)) &= q(x(k+1), x(k), \mathbf{x}(k-1)|y(k+1), \mathbf{y}(k)) \\ &= q(x(k+1)|x(k), y(k+1))\, q(x(k), \mathbf{x}(k-1)|\mathbf{y}(k)) \end{aligned}$$

which is of the form studied in Example 4.32 with the substitution

$$a = x(k+1) \qquad b = x(k) \qquad c = \mathbf{x}(k-1) \qquad d = y(k+1) \qquad e = \mathbf{y}(k)$$

Using the results of that example, our sampling procedure is as follows. We have available samples of $q(\mathbf{x}(k), \mathbf{y}(k)) = q(x(k), \mathbf{x}(k-1)|\mathbf{y}(k))$. Denote these samples by $(x_i(k), \mathbf{x}_i(k-1)), i = 1, \ldots, s$. Then we draw one sample from $q(x(k+1)|x_i(k), y(k+1))$ for each $i = 1, \ldots, s$. Denote these samples as $x_i(k+1)$. Then the samples of $q(\mathbf{x}(k+1)|\mathbf{y}(k+1))$ are given by $(x_i(k+1), x_i(k), \mathbf{x}_i(k-1)) = (x_i(k+1), \mathbf{x}_i(k))$. So we have

$$\mathbf{x}_i(k+1) = (x_i(k+1), \mathbf{x}_i(k)) \qquad i = 1, \ldots, s$$

Next we evaluate the weights for these samples

$$w_i(k+1) = \frac{h(\mathbf{x}_i(k+1)|\mathbf{y}(k+1))}{q(\mathbf{x}_i(k+1)|\mathbf{y}(k+1))}$$

Using (4.57) to evaluate h and the second property of the importance

function to evaluate q gives

$$w_i(k+1) = \frac{p(y(k+1)|x_i(k+1))p(x_i(k+1)|x_i(k))}{q(x_i(k+1)|x_i(k), y(k+1))} \frac{h(\mathbf{x}_i(k)|\mathbf{y}(k))}{q(\mathbf{x}_i(k)|\mathbf{y}(k))}$$

$$w_i(k+1) = \frac{p(y(k+1)|x_i(k+1))p(x_i(k+1)|x_i(k))}{q(x_i(k+1)|x_i(k), y(k+1))} w_i(k) \quad (4.59)$$

$$\overline{w}_i(k+1) = \frac{w_i(k+1)}{\sum_j w_j(k+1)}$$

Notice we obtain a convenient recursion for the weights that depends only on the values of the samples $x_i(k+1)$ and $x_i(k)$ and not the rest of the trajectory contained in the samples $\mathbf{x}_i(k)$. The trajectory's sampled density is given by

$$p(x(k+1), \mathbf{x}(k)|\mathbf{y}(k+1)) = \sum_{i=1}^{s} \overline{w}_i(k+1)\delta(x(k+1) - x_i(k+1))\delta(\mathbf{x}(k) - \mathbf{x}_i(k))$$

Integrating both sides over the $\mathbf{x}(k)$ variables gives the final result

$$p(x(k+1)|\mathbf{y}(k+1)) = \sum_{i=1}^{s} \overline{w}_i(k+1)\delta(x(k+1) - x_i(k+1))$$

Since we generate $x_i(k+1)$ from sampling $q(x(k+1)|x_i(k), y(k+1))$, the trajectory samples, $\mathbf{x}_i(k)$, and measurement trajectory, $\mathbf{y}(k)$, are not required at all, and the particle filter storage requirements do not grow with time. Notice also that if we choose the importance function

$$q(x_i(k+1)|x_i(k), y(k+1)) = p(x_i(k+1)|x_i(k))$$

which ignores the current measurement when sampling, we obtain for the weights

$$w_i(k+1) = w_i(k)\ p(y(k+1)|x_i(k+1))$$

This choice of importance function reduces to the simplest particle filter of the previous section, with its concomitant drawbacks.

Summary. We select an importance function $q(x(k+1)|x(k), y(k+1))$. We start with s samples of $p(x(0))$. We assume that we can evaluate $p(y(k)|x(k))$ using the measurement equation and $p(x(k+$

$1)|x(k))$ using the model equation. The importance function particle filter is summarized by the following recursion

$$
\begin{aligned}
p(x(0)|y(0)) &= \{x_i(0), \overline{w}_i(0)\} \\
w_i(0) &= p(y(0)|x_i(0)) \qquad \overline{w}_i(0) = \frac{w_i(0)}{\sum_j w_j(0)} \\
p(x(k)|\mathbf{y}(k)) &= \{x_i(k), \overline{w}_i(k)\} \\
w_i(k+1) &= \overline{w}_i(k)\, \frac{p(y(k+1)|x_i(k+1))p(x_i(k+1)|x_i(k))}{q(x_i(k+1)|x_i(k), y(k+1))} \\
\overline{w}_i(k+1) &= \frac{w_i(k+1)}{\sum_j w_j(k+1)}
\end{aligned}
$$

and $x_i(k+1)$ is a sample of $q(x(k+1)|x_i(k), y(k+1)), \quad i = 1, \ldots, s$. The sampled density of the importance-sampled particle filter converges to the conditional density $p(x(k)|\mathbf{y}(k))$ in the limit of infinite samples. Because of the way importance sampling was used, the sampled density is biased for all finite sample sizes.

Exercise 4.23 provides the recursion for the weights in the unbiased particle filter; these weights require the evaluation of $p(y(k)|\mathbf{y}(k-1))$. Exercise 4.24 shows that the variance of the unbiased weights increases with time.

4.7.6 Optimal Importance Function

In this section we develop the so-called "optimal" importance function $q(x(k)|x_i(k-1), y(k))$. We start with the weight recursion for the importance function particle filter given in (4.59), repeated here with k replacing $k+1$

$$
w_i(k) = w_i(k-1)\, \frac{p(y(k)|x_i(k))p(x_i(k)|x_i(k-1))}{q(x_i(k)|x_i(k-1), y(k))}
$$

We consider the $w_i(k)$ conditioned on the random variables $x_i(k-1), \mathbf{y}(k)$. The weight $w_i(k)$ is then a function of the random variable $x_i(k)$, which is sampled from the importance function $q(x(k)|x_i(k-$

$1), y(k))$. Taking the expectation gives

$$\begin{aligned}
&\mathcal{E}\left(w_i(k)|x_i(k-1), \mathbf{y}(k)\right) \\
&\qquad = \int w_i(k)\; q(x_i(k)|x_i(k-1), y(k))\; dx_i(k) \\
&\qquad = \int \frac{p(y(k)|x_i(k))p(x_i(k)|x_i(k-1))}{q(x_i(k)|x_i(k-1), y(k))} \\
&\qquad\qquad w_i(k-1)\; q(x_i(k)|x_i(k-1), y(k))\; dx_i(k) \\
&\qquad = \int p(y(k)|x_i(k))\; p(x_i(k)|x_i(k-1))\; w_i(k-1)\; dx_i(k) \\
&\qquad = w_i(k-1)\; p(y(k)|x_i(k-1))
\end{aligned}$$

Next we compute the conditional variance of the weights

$$\begin{aligned}
&\text{var}(w_i(k)|x_i(k-1), \mathbf{y}(k)) \\
&\qquad = \mathcal{E}(w_i^2(k)|x_i(k-1), \mathbf{y}(k)) - \mathcal{E}^2(w_i|x_i(k-1), \mathbf{y}(k))
\end{aligned}$$

Using the recursion in the first term and the expectation just derived in the second term gives

$$\begin{aligned}
&\text{var}(w_i(k)|x_i(k-1), \mathbf{y}(k)) = \\
&\qquad \int w_i^2(k)\; q(x_i(k)|x_i(k-1), y(k))\; dx_i(k) \\
&\qquad\quad - \left(w_i(k-1)\; p(y(k)|x_i(k-1))\right)^2
\end{aligned}$$

$$\begin{aligned}
&\text{var}(w_i(k)|x_i(k-1), \mathbf{y}(k)) \\
&= \int w_i^2(k-1) \frac{\left(p(y(k)|x_i(k))\; p(x_i(k)|x_i(k-1))\right)^2}{q^2(x_i(k)|x_i(k-1), y(k))} \\
&\quad q(x_i(k)|x_i(k-1), y(k))\; dx_i(k) - \left(w_i(k-1)\; p(y(k)|x_i(k-1))\right)^2 \\
&= w_i^2(k-1)\Big[\int \frac{p^2(y(k)|x_i(k))\; p^2(x_i(k)|x_i(k-1))}{q(x_i(k)|x_i(k-1), y(k))}\; dx_i(k) \\
&\qquad - p^2(y(k)|x_i(k-1))\Big]
\end{aligned}$$

We can now optimize the choice of $q(x_i(k)|x_i(k-1), y(k))$ to minimize this conditional variance. Consider the choice

$$\boxed{q(x_i(k)|x_i(k-1), y(k)) = p(x_i(k)|x_i(k-1), y(k))} \tag{4.60}$$

which makes the samples at k depend on current measurement $y(k)$ as well as the past samples. We know from Bayes's rule and the Markov property

$$\begin{aligned} q(x_i(k)|x_i(k-1), y(k)) &= p(x_i(k)|x_i(k-1), y(k)) \\ &= \frac{p(y(k)|x_i(k), x_i(k-1))p(x_i(k)|x_i(k-1))}{p(y(k)|x_i(k-1))} \\ q(x_i(k)|x_i(k-1), y(k)) &= \frac{p(y(k)|x_i(k))p(x_i(k)|x_i(k-1))}{p(y(k)|x_i(k-1))} \end{aligned}$$

Using this result we have for the integral term

$$\begin{aligned} &\int \frac{p^2(y(k)|x_i(k))\ p^2(x_i(k)|x_i(k-1))}{q(x_i(k)|x_i(k-1), y(k))}\, dx_i(k) \\ &= p(y(k)|x_i(k-1)) \int p(y(k)|x_i(k))\ p(x_i(k)|x_i(k-1))\, dx_i(k) \\ &= p^2(y(k)|x_i(k-1)) \end{aligned}$$

Substituting this result into the previous equation for conditional variance gives

$$\text{var}(w_i(k)|x_i(k-1), \mathbf{y}(k)) = 0$$

Since variance is nonnegative, the choice of importance function given in (4.60) is optimal for reducing the conditional variance of the weights. This choice has the important benefit of making the samples $x_i(k)$ more responsive to the measurement $y(k)$, which we show in the next example is a big improvement over the simplest particle filter.

Example 4.39: Optimal importance function applied to a linear estimation problem

Given the linear system of Example 4.37 and 250 particles, show the particles' locations for times $k = 0, 1, \ldots, 5$ along with the 95% elliptical contour of the true conditional density $p(x(k)|\mathbf{y}(k))$. Perform this calculation with and without resampling after every time step.

Solution

The optimal importance function is given in (4.60)

$$q(x_i(k)|x_i(k-1), y(k)) = p(x_i(k)|x_i(k-1), y(k))$$

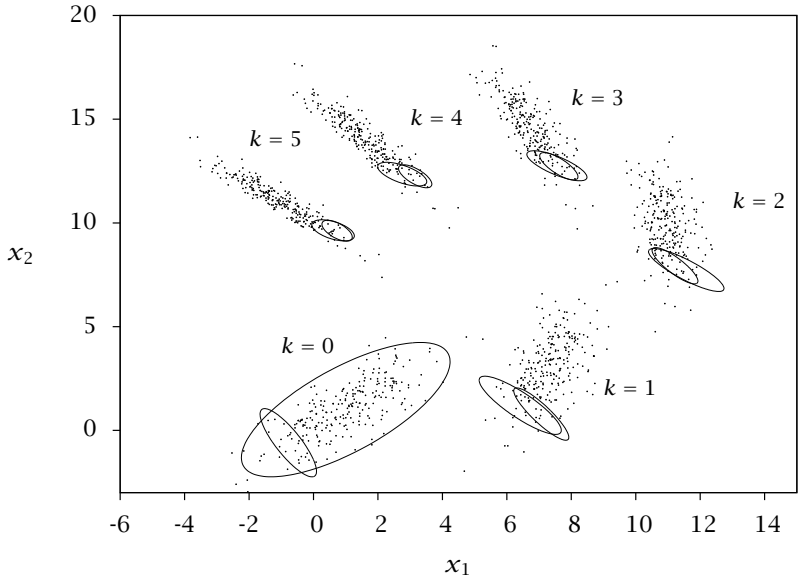

Figure 4.25: Particles' locations versus time using the optimal importance function; 250 particles. Ellipses show the 95% contour of the true conditional densities before and after measurement.

The conditional density on the right-hand side is given by

$$p(x_i(k)|x_i(k-1), y(k)) \sim N(\overline{x}(k), \overline{P})$$

$$\overline{x}(k) = \overline{P}\left(Q^{-1}(Ax_i(k-1) + Bu(k-1)) + C'R^{-1}y(k)\right)$$
$$\overline{P} = \left(Q^{-1} + C'R^{-1}C\right)^{-1}$$

Exercise 4.25 discusses establishing this result. So the $x_i(k)$ are generated by sampling this normal, and the results are shown in Figure 4.25. We see that the optimal importance function adds a $y(k)$ term to the evolution of the particle mean. This term makes the particles more responsive to the data and the mean particle location better tracks the conditional density's mean. Compare Figure 4.22 for the simplest particle filter with Figure 4.25 to see the improvement. Also the variance no longer increases with time as in the simplest particle filter so the particles do not continue to spread apart.

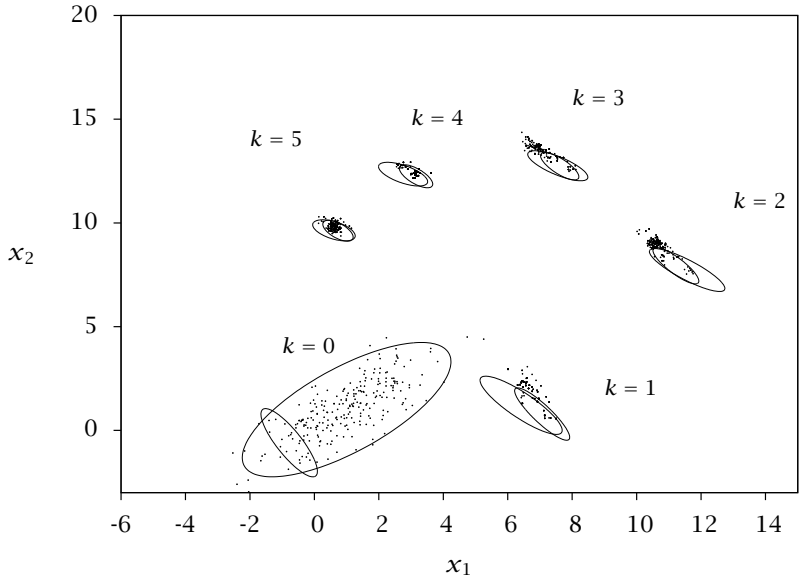

Figure 4.26: Particles' locations versus time using the optimal importance function with resampling; 250 particles. Ellipses show the 95% contour of the true conditional densities before and after measurement.

If we apply resampling at every time step, we obtain the results in Figure 4.26. As we saw in the case of the simplest particle filter, resampling greatly increases the number of samples inside the 95% probability ellipse of the conditional density.

If we rerun the simulation 500 times and plot versus time the fraction of particles that are inside the 95% contour of the true conditional density, we obtain the result shown in Figure 4.27. The optimal importance function is able to maintain about 20% of the particles in the 95% probability ellipse. With the optimal importance function and resampling, more than 90% of the particles are inside the 95% probability ellipse. The earlier warning about sample impoverishment applies here as well. □

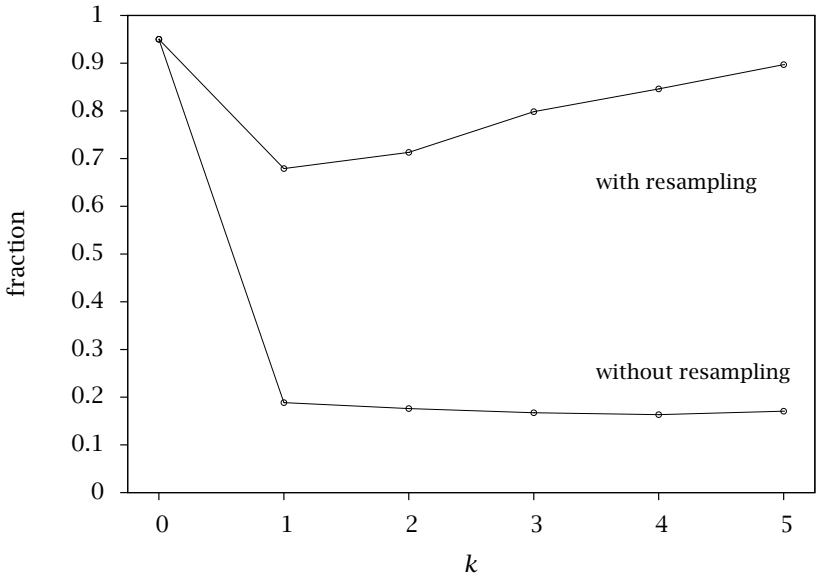

Figure 4.27: Fraction of particles inside the 95% contour of the true conditional density versus time; with and without resampling; average of 500 runs.

4.8 Combined MHE/Particle Filtering

We next propose a new state estimation method that combines some of the best elements of MHE and PF. This type of combination has several design parameters and can take different forms, and we use the general term combined MHE/PF to designate this entire class of state estimators. To motivate the design of MHE/PF, consider the strengths and weaknesses of pure MHE and pure PF. The main *strengths of MHE* are

1. MHE propagates the state using the full nonlinear model.
2. MHE uses optimization to find the most likely estimate. Physical constraints can be included in the optimization.
3. MHE employs a horizon of measurements.

Using the full nonlinear model prevents inaccurate model linearizations from interfering with the fitting of the model to the data. The

use of optimization produces the best state or state trajectory to describe the current snapshot of data. Optimization methods generally evaluate a small set of points in the state space to find the best estimate compared to exhaustive enumeration, gridding, and sampling strategies. That becomes a significant strength as the dimension of the state space model increases past $n \approx 2$–3. The use of a moving window of data provides some robustness to unmodeled disturbances entering the system. The goal in most recursive estimation is to consider measurements one at a time. That is often a valid goal, mainly because it allows faster computation of the current estimate given the current measurement. But unmodeled disturbances are often problematic when measurements are considered one at a time. No single measurement is sufficient to conclude that an unmodeled disturbance has shifted the state significantly from its current estimated value. Only when several sequential measurements are considered at once is the evidence sufficient to overturn the current state estimate and move the state a significant distance to better match all of the measurements. MHE has this capability built in.

The main *weaknesses of MHE* are

1. MHE may take significant computation time.
2. MHE uses local instead of global optimization.

Of course attempting global optimization is possible, but that exacerbates weakness 1 significantly and no guarantees of finding a global optimum are available for anything but the simplest nonlinear models. Note that for the special case of linear models, MHE finds the global optimum and weakness 2 is removed.

Particle filtering displays quite different characteristics than those of MHE. The main *strengths of PF* are

1. PF uses the full nonlinear model to propagate the samples.
2. The PF sampled density can represent a general conditional density.
3. PF is simple to program and executes quickly for small sample sizes.

As we have illustrated with simple examples, pure PF also demonstrates significant weaknesses, and these are not remedied by any suggestions in the research literature of which we are aware. The *weaknesses of PF* include

1. PF exhibits significant decrease in performance with increasing state dimension.

2. PF displays poor robustness to unmodeled disturbances.

The lack of robustness is a direct outcome of the sampling strategies. Sampling any of the proposed PF importance functions does not locate the samples close to the true state after a significant and unmodeled disturbance. Once the samples are in the wrong place with respect to the peak in the conditional density, they do not recover. If the samples are in the wrong part of the state space, the weights cannot carry the load and represent the conditional density. Resampling does not successfully reposition the particles if they are already out of place. An appeal to sampled density convergence to the true conditional density with increasing sample number is unrealistic. The number of samples required is simply too large for even reasonably small state dimensions considered in applications; $n > 50$ is not unusual in applications.

In constructing a class of combined methods we propose to

1. Use MHE to locate/relocate the samples.

2. Use PF to obtain fast recursive estimation between MHE optimizations.

We overcome the potentially expensive MHE optimization by using PF to process the measurements and provide rapid online estimates while an MHE computation is underway. We position the samples in regions of high conditional density after every run of the MHE optimization, which allows recovery from unmodeled disturbances as soon as an MHE computation completes. A challenge that is not addressed is the appearance of multiple peaks in the conditional density when using nonlinear models. Handling the multimodal conditional density remains a challenge for any online, and indeed offline, state estimation procedure.

Next we propose a specific state estimator in this general MHE/PF class and examine its performance with some simple computational examples. Because this class of estimators is new, we fully expect significant modifications and improvements to come along. At this early juncture we expect only to be able to illustrate some of the new capabilities of the approach.

Let $\hat{Z}_k(x)$ denote the MHE arrival cost function given in Definition 4.16. We let $\hat{V}_k^0$ denote the optimal cost and $\hat{x}(k)$ the optimal estimate of the last stage at time k. We consider the quadratic approximation of

$\hat{Z}_k(\cdot)$ at the optimum $\hat{x}(k)$

$$V(x) = V_k^0(\hat{x}(k)) + (1/2)(x - \hat{x}(k))' H (x - \hat{x}(k))$$

in which H is the Hessian of $\hat{Z}_k(x)$ evaluated at the optimum $\hat{x}(k)$. We use this function as an importance function for sampling the conditional density. Notice that this procedure is not the same as assuming the conditional density itself is a normal distribution. We are using $N(\hat{x}(k), H^{-1})$ strictly as an importance function for sampling the unknown conditional density. The samples $x_i(k)$ are drawn from $N(\hat{x}(k), H^{-1})$. The weights are given by

$$w_i(k) = V(x_i(k)) \qquad \overline{w}_i(k) = \frac{w_i(k)}{\sum_j w_j(k)} \tag{4.61}$$

and the sampled density is given by

$$p_s(x) = \{x_i(k), \overline{w}_i(k)\}$$

If the conditional density is well represented by the normal approximation, then the normalized weights are all nearly equal to $1/s$. The MHE cost function modifies these ideal weights as shown in (4.61).

Example 4.40: Comparison of MHE, PF, and combined MHE/PF

Consider a well-mixed semibatch chemical reactor in which the following reaction takes place

$$2\text{A} \xrightarrow{k} \text{B} \qquad r = kc_A^2$$

The material balances for the two components are

$$\frac{dc_A}{dt} = -2kc_A^2 + \frac{Q_f}{V} c_{Af}$$
$$\frac{dc_B}{dt} = kc_A^2 + \frac{Q_f}{V} c_{Bf}$$

with constant parameter values

$$\frac{Q_f}{V} = 0.4 \qquad k = 0.16 \qquad c_{Af} = 1 \qquad c_{Bf} = 0$$

The scalar measurement is the total pressure, which is the sum of the two states. The sample time is $\Delta = 0.1$. The initial state is $x(0) = [3\ 1]'$ and the initial prior mean is $\hat{x}(0) = [0.1\ 4.5]'$. Moreover, the input

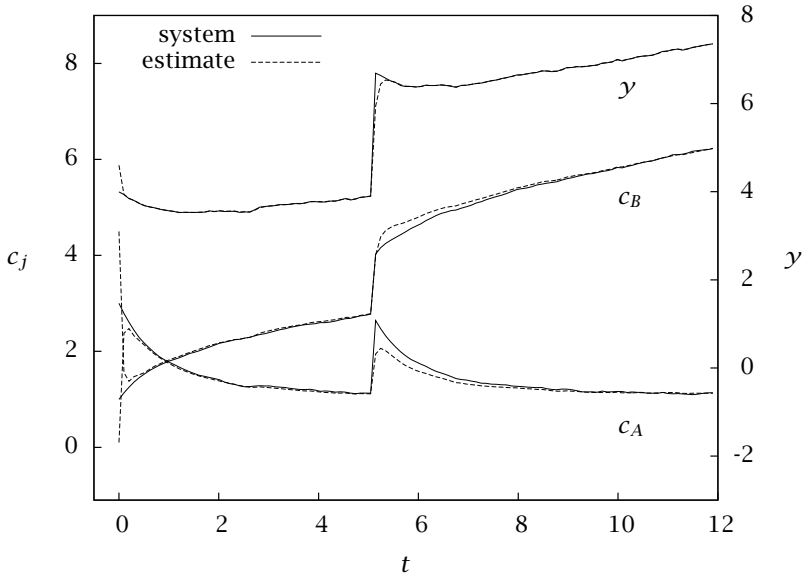

Figure 4.28: Pure MHE.

suffers an unmodeled step disturbance at $t = 5$ for two samples. So this example tests robustness of the estimator to initial state error and unmodeled disturbances.

First we apply MHE to the example and the results are displayed in Figure 4.28. The horizon is chosen as $N = 15$. The initial covariance is chosen to be $P_0 = 10I_2$ to reflect the poor confidence in the initial state. Notice that MHE is able to recover from the poor initial state prior in only 4 or 5 samples.

Next we apply pure particle filtering using 50 particles. We use the optimal importance function because the measurement equation is linear. The particles are initialized using the same initial density as used in the MHE estimator.

$$p_{x(0)}(x) = n(x, \hat{x}(0), P(0))$$

The results are shown in Figure 4.29. The figure shows the state and output mean versus time. We notice two effects. The particle filter is unable to recover from the poor initial samples. The measurement is predicted well but neither state is estimated accurately. The A concentration estimate is also negative, which is physically impossible. The

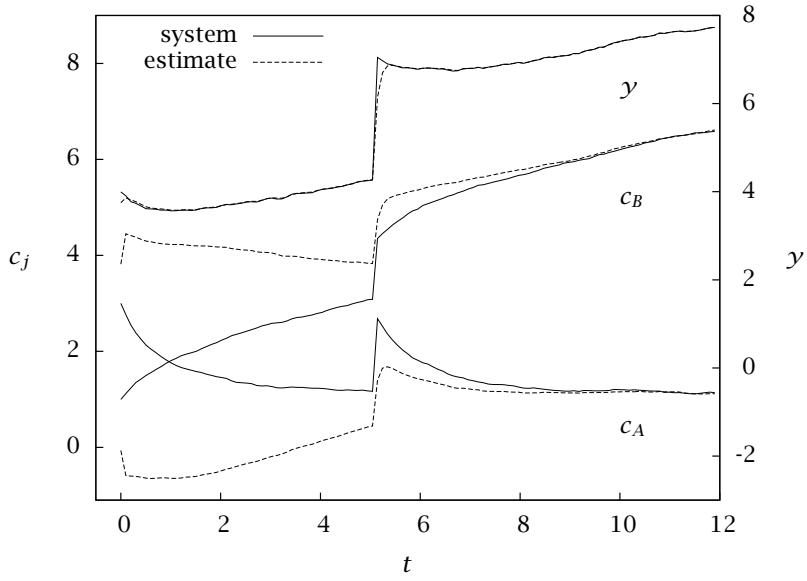

Figure 4.29: Pure PF with optimal importance function.

disturbance at $t = 5$ is fortuitous and helps the PF get back on track.

Next we assume that the MHE optimization cannot finish in one sample, but requires M samples. If we attempt a pure MHE solution in this situation, the estimator falls hopelessly behind; an estimate using data $\mathbf{y}(k-M,k), k \geq M$ is not available until time Mk. Instead we use MHE/PF as follows.

1. At time k run MHE on data $\mathbf{y}(k-M,k)$. This computation is assumed to finish at time $k+M$. For simplicity, assume N large and a noninformative prior.

2. Draw samples from $N(\hat{x}(k), P(k))$. Run the particle filtering update from time k to time $k+M$. For illustrative purposes, we assume this PF step finishes in one sample.

3. Update k to $k+M$ and repeat.

For illustrative purposes, we choose $M = 10$ and apply the combination of MHE and PF with the simple importance function, also using 50 particles as before. The results are shown in Figure 4.30. Notice that again the poor initial samples lead to significant estimate error.

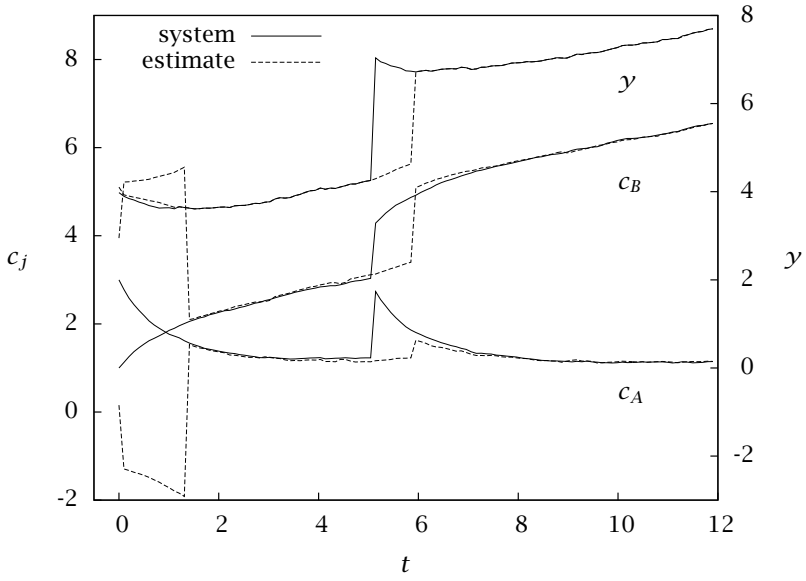

Figure 4.30: Combination MHE/PF with simple importance function.

But the inaccurate sample is repaired after $M = 10$ samples. The MHE calculation completes by about $t = 2$, and the samples are reinitialized from the MHE cost function at $t = 1$, and run forward from $t = 1$. These reinitialized samples converge to the true state shortly after $t = 1$.[7]

The disturbance at $t = 5$ also causes the PF samples with the simple importance function to be in the wrong locations. They do not recover and inaccurate estimates are produced by the PF. Another MHE calculation starts at $t = 5$ and finishes at $t = 6$, and the samples are reinitialized with the MHE cost function at $t = 5$ and run forward. After this resampling, the PF estimates again quickly converge to the true estimates after $t = 6$.

Next we use the combination of MHE and PF with the optimal importance function. These results are shown in Figure 4.31. We see as

[7]Even with only 50 particles, we find that particle filtering is not so much faster than MHE, that its computation time can be neglected as we have done here. The two computations take about the *same* time with 50 particles. The computational expense in PF arises from calling an ODE solver 50 times at each sample time. No attempt was made to tailor the ODE solver for efficiency by exploiting the fact that the sample time is small. Note, however, that tailoring the ODE solver would speed up MHE as well as PF.

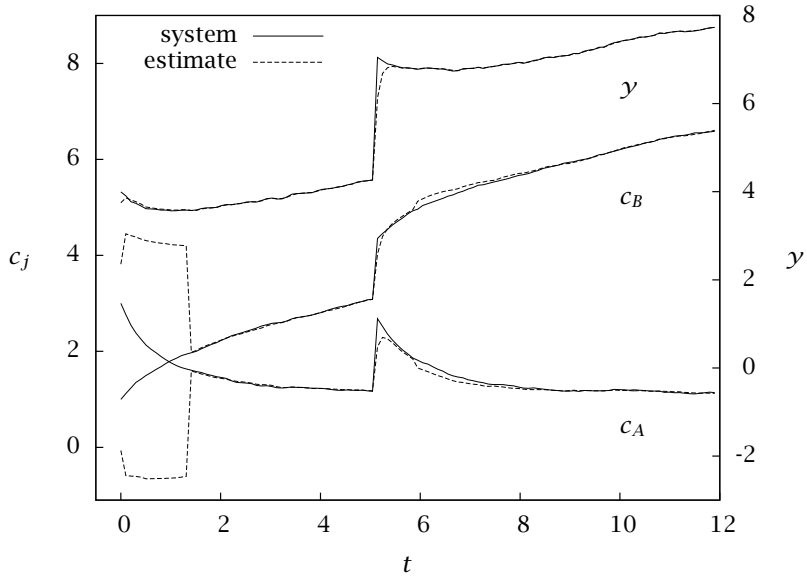

Figure 4.31: Combination MHE/PF with optimal importance function.

in the early part of Figure 4.29 that the samples cannot recover from the poor initial state prior and resampling from the MHE cost function takes place at $t = 1$ after the first MHE calculation finishes at $t = 2$. But as in the case of pure PF with the optimal importance function, the disturbance does not move the state so far from the samples that they are unable to recover and continue to provide accurate estimates. The MHE resampling that takes place at $t = 5$ after MHE finishes at $t = 6$ does not modify significantly the PF samples that are already well placed. □

Of course, the simulations shown in Figures 4.28–4.31 display the outcome of only single random realizations. A full characterization of the behavior of the four estimators is determined by running many such random simulations and computing the statistics of interest. We have not compiled these statistics because the single simulations are rather time consuming. After running several random simulations for each estimator, these single simulations were selected manually as representative behavior of the different estimators.

Substitute the results for $\hat{x}(k)$ and $P(k)$ above and show

$$
\begin{aligned}
V_{k+1}^{-}(z) &= (1/2)(z-\hat{x}^{-}(k+1))'(P^{-}(k+1))^{-1}(z-\hat{x}(k+1)) \\
P^{-}(k+1) &= Q + AP^{-}(k)A' - AP^{-}(k)C'(CP^{-}(k)C'+R)^{-1}CP^{-}(k)A \\
\hat{x}^{-}(k+1) &= A\hat{x}^{-}(k) + \widetilde{L}(k)(y(k)-C\hat{x}^{-}(k)) \\
\widetilde{L}(k) &= AP^{-}(k)C'(CP^{-}(k)C'+R)^{-1}
\end{aligned}
$$

(c) Compare and contrast this form of the estimation problem to the one given in Exercise 1.29 that describes the Kalman filter.

Exercise 4.13: Duality, cost to go, and covariance

Using the duality variables of Table 4.2, translate Theorem 4.10 into the version that is relevant to the state estimation problem.

Exercise 4.14: Estimator convergence for (A, G) not stabilizable

What happens to the stability of the optimal estimator if we violate the condition

$$(A, G) \text{ stabilizable}$$

(a) Is the steady-state Kalman filter a stable estimator? Is the full information estimator a stable estimator? Are these two answers contradictory? Work out the results for the case $A = 1, G = 0, C = 1, P^{-}(0) = 1, Q = 1, R = 1$.
Hint: you may want to consult de Souza, Gevers, and Goodwin (1986).

(b) Can this phenomenon happen in the LQ regulator? Provide the interpretation of the time-varying regulator that corresponds to the time-varying filter given above. Does this make sense as a regulation problem?

Exercise 4.15: Exponential stability of the Kalman predictor

Establish that the Kalman predictor defined in Section 4.2.1 is a globally exponentially stable estimator. What is the corresponding linear quadratic regulator?

Exercise 4.16: The resampling theorem

Generalize the proof of Theorem 4.35 to cover any number of samples.
Hint: you may find the multinomial expansion formula useful

$$(x_1 + x_2 + \cdots + x_s)^k = \sum_{r_1=0}^{k} \sum_{r_2=0}^{k} \cdots \sum_{r_s=0}^{k} a(r_1, r_2, \ldots, r_s) x_1^{r_1} x_2^{r_2} \cdots x_s^{r_s}$$

in which the coefficients in the expansion formula are given by Feller (1968, p.37)

$$a(r_1, r_2, \ldots, r_s) = \begin{cases} \dfrac{k!}{r_1! r_2! \cdots r_s!} & r_1 + r_2 + \cdots + r_s = k \\ 0 & r_1 + r_2 + \cdots + r_s \neq k \end{cases} \tag{4.63}$$

Exercise 4.17: Pruning while resampling

Sometimes it is convenient in a simulation to reduce the number of samples when resampling a density. In many discrete processes, for example, the number of possible states that may be reached in the simulation increases with time. To keep the number of samples constant, we may wish to remove samples at each time through the resampling process. Consider a modification of Theorem 4.35 in which the number of resamples is $\tilde{s}$, which does not have to be equal to s.

Theorem 4.41 (Resampling and pruning)**.** *Consider a sampled density $p(x)$ with s samples at $x = x_i$ and associated weights w_i*

$$p(x) = \sum_{i=1}^{s} w_i \delta(x - x_i) \qquad w_i \geq 0 \qquad \sum_{i=1}^{s} w_i = 1$$

Consider the resampling procedure that gives a resampled density with $\tilde{s} > 0$ samples

$$\tilde{p}(x) = \sum_{i=1}^{\tilde{s}} \tilde{w}_i \delta(x - \tilde{x}_i)$$

in which the $\tilde{x}_i$ are chosen according to resample probability p_r

$$p_r(\tilde{x}_i) = \begin{cases} w_j, & \tilde{x}_i = x_j \\ 0, & \tilde{x}_i \neq x_j \end{cases}$$

and with uniform weights $\tilde{w}_i = 1/\tilde{s}$. Consider a function $f(\cdot)$ defined on a set X containing the points x_i.

Under this resampling procedure, the expectation over resampling of any integral of the resampled density is equal to that same integral of the original density

$$\mathcal{E}_r \left(\int f(x) \tilde{p}(x) dx \right) = \int f(x) p(x) dx \qquad \text{all } f$$

(a) Is the proposed theorem correct? If so, prove it. If not, provide a counterexample.

(b) What do you suppose happens in a simulation if we perform aggressive pruning by always choosing $\tilde{s} = 1$?

Exercise 4.18: Properties of the importance function

It is stated in the chapter that $p(\mathbf{x}(k+1)|\mathbf{y}(k+1))$ does not satisfy the second importance function property listed in (4.58)

$$q(\mathbf{x}(k+1)|\mathbf{y}(k+1)) = q(x(k+1)|x(k), y(k+1))\, q(\mathbf{x}(k)|\mathbf{y}(k)) \tag{4.64}$$

Derive a similar property that $p(\mathbf{x}(k+1)|\mathbf{y}(k+1))$ *does* satisfy. What has been altered in (4.64)? Why do you think this change has been made?

Exercise 4.19: A single sample of joint density

Consider again Example 4.31 in which we have s_x and s_y samples of the marginals of independent random variables ξ and η, respectively

$$\xi \sim \{x_i, w_{xi}\} \qquad w_{xi} = 1/s_x, \quad i = 1, \ldots, s_x$$
$$\eta \sim \{y_j, w_{yj}\} \qquad w_{yj} = 1/s_y, \quad j = 1, \ldots, s_y$$

and wish to sample the joint density $p_{\xi,\eta}(x,y) = p_\xi(x)p_\eta(y)$. Show that selecting any single sample is a valid sample of the joint density

$$\{(x_1, y_1), w\}, \quad w = 1$$

Exercise 4.20: Kolmogorov-Smirnov limit theorem for sampling error

Consider again s mutually independent samples taken from cumulative distribution $P(x)$ to produce the sampled cumulative distribution $P_s(x;s)$ as discussed in Section 4.7.2. Define sampling error as in the chapter

$$D_s = \sup_x |P_s(x;s) - P(x)|$$

(a) Reproduce the results of Example 4.30. Plot the actual and limiting distributions for D_s for $s = 10, 100, 1000$ when sampling a normal distribution with unit variance. Your result should resemble Figure 4.14

(b) Now compute the actual and limiting probability *densities* of the sampling error $p(D_s)$ rather than the distribution $\Pr(D_s)$. Give a formula for $l(z) = dL(z)/dz$. Plot $p(D_s)$ for $s = 10, 100, 1000$ samples for sampling the normal distribution with unit variance.

Exercise 4.21: Sampled density from a weighted importance function

Given a weighted sample of an importance function $q(x)$

$$q_s(x) = \sum_{i=1}^{s} w_i^- \delta(x - x_i) \qquad \sum_i w_i^- = 1$$

(a) Show that the sampled density

$$\overline{p}_s(x) = \sum_{i=1}^{s} w_i \delta(x - x_i) \qquad w_i = w_i^- \frac{p(x_i)}{q(x_i)}$$

converges to $p(x)$ as sample size increases.

(b) Show that the sampled density is unbiased for all samples sizes.

Exercise 4.22: Sampled density from a weighted importance function when unable to evaluate the density

Given a weighted sample of an importance function $q(x)$

$$q_s(x) = \sum_{i=1}^{s} w_i^- \delta(x - x_i) \qquad \sum_i w_i^- = 1$$

and a density of the following form

$$p(x) = \frac{h(x)}{\int h(x)dx}$$

in which $p(x)$ cannot be conveniently evaluated but $h(x)$ can be evaluated.

(a) Show that the sampled density

$$\overline{p}_s(x) = \sum_{i=1}^{s} \overline{w}_i \delta(x - x_i) \qquad w_i = w_i^- \frac{h(x_i)}{q(x_i)} \qquad \overline{w}_i = \frac{w_i}{\sum_j w_j}$$

converges to $p(x)$ as sample size increases.

(b) Show that the sampled density is biased for all finite sample sizes.

Exercise 4.23: Unbiased particle filter with importance sampling

Show that an unbiased particle filter using importance sampling is given by

$$\overline{p}_s(x(k)|\mathbf{y}(k)) = \{x_i(k), \widetilde{w}_i(k)\}$$

$$\widetilde{w}_i(k+1) = \widetilde{w}_i(k) \frac{p(y(k+1)|x_i(k+1))\, p(x_i(k+1)|x_i(k))}{p(y(k+1)|\mathbf{y}(k))\, q(x_i(k+1)|x_i(k), y(k+1))}$$

in which $x_i(k)$ are samples of the importance function $q(x(k)|x_i(k-1), y(k))$. Note that normalization of $\widetilde{w}_i$ is not required in this form of a particle filter, but evaluation of $p(y(k+1)|\mathbf{y}(k))$ is required.

Exercise 4.24: Variance of the unbiased particle filter with importance sampling

Show that the variance of the weights of the unbiased particle filter given in Exercise 4.23 increases with time.

Exercise 4.25: Optimal importance function for a linear system

The optimal importance function is given in (4.60), repeated here

$$q(x_i(k)|x_i(k-1), y(k)) = p(x_i(k)|x_i(k-1), y(k))$$

For the linear time-invariant model, this conditional density is the following normal density (Doucet et al., 2000)

$$p(x_i(k)|x_i(k-1), y(k)) \sim N(\overline{x}(k), \overline{P})$$

$$\overline{x}(k) = \overline{P}Q^{-1}(Ax_i(k-1) + Bu(k-1)) + \overline{P}C'R^{-1}y(k)$$

$$\overline{P} = \left(Q^{-1} + C'R^{-1}C\right)^{-1}$$

Establish this result by first considering the linear transformation between $(x_i(k), y(k))$ and $x_i(k-1), w(k), v(k)$, and then using the formulas for taking conditional densities of normals.

Exercise 4.26: Equivalance of detectability and IOSS for continuous-time, linear, time-invariant system

Consider the continuous-time, linear, time-invariant system with input

$$\dot{x} = Ax + Bu \qquad y = Cx$$

Show that the system is detectable if and only if the system is IOSS.

Bibliography

M. S. Arulampalam, S. Maskell, N. J. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.*, 50(2):174–188, February 2002.

K. J. Åström. *Introduction to Stochastic Control Theory*. Academic Press, San Diego, California, 1970.

D. P. Bertsekas. *Dynamic Programming*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

A. E. Bryson and Y. Ho. *Applied Optimal Control*. Hemisphere Publishing, New York, 1975.

F. M. Callier and C. A. Desoer. *Linear System Theory*. Springer-Verlag, New York, 1991.

M. Chaves and E. D. Sontag. State-estimators for chemical reaction networks of Feinberg-Horn-Jackson zero deficiency type. *Eur. J. Control*, 8(4):343–359, 2002.

C. E. de Souza, M. R. Gevers, and G. C. Goodwin. Riccati equation in optimal filtering of nonstabilizable systems having singular state transition matrices. *IEEE Trans. Auto. Cont.*, 31(9):831–838, September 1986.

A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. and Comput.*, 10:197–208, 2000.

W. Feller. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Ann. Math. Stat.*, 19(2):177–189, 1948.

W. Feller. *An Introduction to Probability Theory and Its Applications: Volume I*. John Wiley & Sons, New York, third edition, 1968.

A. Gelb, editor. *Applied Optimal Estimation*. The M.I.T. Press, Cambridge, Massachusetts, 1974.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F-Radar and Signal Processing*, 140(2):107–113, April 1993.

R. Gudi, S. Shah, and M. Gray. Multirate state and parameter estimation in an antibiotic fermentation with delayed measurements. *Biotech. Bioeng.*, 44: 1271–1278, 1994.

J. E. Handschin and D. Q. Mayne. Monte Carlo techniques to estimate the conditional expectation in multistage nonlinear filtering. *Int. J. Control*, 9 (5):547–559, 1969.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

Z.-P. Jiang and Y. Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37:857–869, 2001.

S. J. Julier and J. K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *International Symposium Aerospace/Defense Sensing, Simulation and Controls*, pages 182–193, 1997.

S. J. Julier and J. K. Uhlmann. Author's reply. *IEEE Trans. Auto. Cont.*, 47(8): 1408–1409, August 2002.

S. J. Julier and J. K. Uhlmann. Unscented filtering and nonlinear estimation. *Proc. IEEE*, 92(3):401–422, March 2004a.

S. J. Julier and J. K. Uhlmann. Corrections to unscented filtering and nonlinear estimation. *Proc. IEEE*, 92(12):1958, December 2004b.

S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Trans. Auto. Cont.*, 45(3):477–482, March 2000.

T. Kailath. A view of three decades of linear filtering theory. *IEEE Trans. Inform. Theory*, IT-20(2):146–181, March 1974.

R. Kandepu, L. Imsland, and B. A. Foss. Constrained state estimation using the unscented kalman filter. In *Procedings of the 16th Mediterranean Conference on Control and Automation*, pages 1453–1458, Ajaccio, France, June 2008.

A. Kolmogoroff. Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari*, 4:1–11, 1933.

A. N. Kolmogorov. Interpolation and extrapolation of stationary random sequences. *Bull. Moscow Univ., USSR, Ser. Math. 5*, 1941.

H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems*. John Wiley and Sons, New York, 1972.

T. Lefebvre, H. Bruyninckx, and J. De Schutter. Comment on "A new method for the nonlinear transformation of means and covariances in filters and estimators". *IEEE Trans. Auto. Cont.*, 47(8):1406–1408, August 2002.

E. S. Meadows, K. R. Muske, and J. B. Rawlings. Constrained state estimation and discontinuous feedback in model predictive control. In *Proceedings of the 1993 European Control Conference*, pages 2308–2312, 1993.

H. Michalska and D. Q. Mayne. Moving horizon observers and observer-based control. *IEEE Trans. Auto. Cont.*, 40(6):995–1006, 1995.

S. A. Middlebrooks and J. B. Rawlings. State estimation approach for determining composition and growth rate of $Si_{1-x}Ge_x$ chemical vapor deposition utilizing real-time ellipsometric measurements. *Applied Opt.*, 45:7043–7055, 2006.

K. R. Muske, J. B. Rawlings, and J. H. Lee. Receding horizon recursive state estimation. In *Proceedings of the 1993 American Control Conference*, pages 900–904, June 1993.

M. Nørgaard, N. K. Poulsen, and O. Ravn. New developments in state estimation for nonlinear systems. *Automatica*, 36:1627–1638, 2000.

V. Prasad, M. Schley, L. P. Russo, and B. W. Bequette. Product property and production rate control of styrene polymerization. *J. Proc. Cont.*, 12(3):353–372, 2002.

C. C. Qu and J. Hahn. Computation of arrival cost for moving horizon estimation via unscent Kalman filtering. *J. Proc. Cont.*, 19(2):358–363, 2009.

C. V. Rao. *Moving Horizon Strategies for the Constrained Monitoring and Control of Nonlinear Discrete-Time Systems*. PhD thesis, University of Wisconsin–Madison, 2000.

C. V. Rao, J. B. Rawlings, and J. H. Lee. Constrained linear state estimation – a moving horizon approach. *Automatica*, 37(10):1619–1628, 2001.

C. V. Rao, J. B. Rawlings, and D. Q. Mayne. Constrained state estimation for nonlinear discrete-time systems: stability and moving horizon approximations. *IEEE Trans. Auto. Cont.*, 48(2):246–258, February 2003.

H. E. Rauch, F. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA J.*, 3(8):1445–1450, 1965.

J. B. Rawlings and B. R. Bakshi. Particle filtering and moving horizon estimation. *Comput. Chem. Eng.*, 30:1529–1541, 2006.

K. Reif and R. Unbehauen. The extended Kalman filter as an exponential observer for nonlinear systems. *IEEE Trans. Signal Process.*, 47(8):2324–2328, August 1999.

K. Reif, S. Günther, E. Yaz, and R. Unbehauen. Stochastic stability of the discrete-time extended Kalman filter. *IEEE Trans. Auto. Cont.*, 44(4):714–728, April 1999.

K. Reif, S. Günther, E. Yaz, and R. Unbehauen. Stochastic stability of the continuous-time extended Kalman filter. *IEE Proceedings-Control Theory and Applications*, 147(1):45–52, January 2000.

D. G. Robertson and J. H. Lee. On the use of constraints in least squares estimation and control. *Automatica*, 38(7):1113–1124, 2002.

A. Romanenko and J. A. A. M. Castro. The unscented filter as an alternative to the EKF for nonlinear state estimation: a simulation case study. *Comput. Chem. Eng.*, 28(3):347–355, March 15 2004.

A. Romanenko, L. O. Santos, and P. A. F. N. A. Afonso. Unscented Kalman filtering of a simulated pH system. *Ind. Eng. Chem. Res.*, 43:7531–7538, 2004.

N. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Univesité de Moscou*, 2, 1939. fasc. 2.

A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: A sampling-resampling perspective. *Amer. Statist.*, 46(2):84–88, 1992.

E. D. Sontag. *Mathematical Control Theory*. Springer-Verlag, New York, second edition, 1998a.

E. D. Sontag. Comments on integral variants of ISS. *Sys. Cont. Let.*, 34:93–100, 1998b.

E. D. Sontag and Y. Wang. Output-to-state stability and detectability of nonlinear systems. *Sys. Cont. Let.*, 29:279–290, 1997.

R. F. Stengel. *Optimal Control and Estimation*. Dover Publications, Inc., 1994.

B. O. S. Teixeira, L. A. B. Tôrres, L. A. Aguirre, and D. S. Bernstein. Unscented filtering for interval-constrained nonlinear systems. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 5116–5121, Cancun, Mexico, December 9-11 2008.

P. Vachhani, S. Narasimhan, and R. Rengaswamy. Robust and reliable estimation via unscented recursive nonlinear dynamic data reconciliation. *J. Proc. Cont.*, 16(10):1075–1086, December 2006.

R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan. The unscented particle filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University Engineering Department, August 2000.

N. Wiener. *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. Wiley, New York, 1949. Originally issued as a classified MIT Rad. Lab. Report in February 1942.

D. I. Wilson, M. Agarwal, and D. W. T. Rippin. Experiences implementing the extended Kalman filter on an industrial batch reactor. *Comput. Chem. Eng.*, 22(11):1653–1672, 1998.