Inference Example 3: The Kalman Filter

**Question:** You would like to build an automatic system to land a spacecraft on the moon. To steer the spacecraft properly, the system needs to estimate the current location of the spacecraft relative to the moon surface. Unfortunately, the sensors are noisy. How can the system best estimate the current location given all the noisy measurements of the past trajectory of the spacecraft?

**Comment 1:** This is an example of a very common problem in a diverse range of fields, such as control, signal and image processing, computer vision, finance etc. The general problem is to recover an underlying signal from noisy observations and perhaps to predict its future trajectory. The signal may be an image, an audio signal, trajectory of an aircraft, quality of a stock, etc. This class of problem is called *filtering*, *denoising* or *prediction*. The idea is to separate out the underlying signal from the noise. What distinguishes the signal from the noise is that the signal is often “smooth”: location of the spacecraft from one measurement to the next does not change much, values of adjacent pixels of an image are likely to be similar. On the other hand, the noise is highly random and varies significantly from one measurement to the next.

**Comment 2:** This problem is yet another example of inference problems. We have considered examples where the unknown and the observations are discrete (communication over binary symmetric channels), and now we are considering an example where both the unknown (the underlying signal) and the noisy observations are continuous.

**Comment 3:** Historically, this problem first arose in the 1960’s in the Apollo space program to land Americans on the moon. The solution of this problem is the celebrated *Kalman filter*, which we will now describe (for a very simple model).

### Modeling

The situation is shown in Figure 1.

The underlying signal is modeled by a sequence of random variables $X_0, X_1, X_2, \ldots$. The noisy observations are $Y_1, Y_2, \ldots$, given by

$$Y_i = X_i + Z_i, \quad i = 1, \ldots$$

The $Z_i$'s are i.i.d. $N(0, \sigma^2_Z)$ r.v.’s and independent of the $X_i$’s. The signal $X_i$’s are described by:

$$X_0 \sim N(\mu_0, \sigma^2_0)$$

$$X_{i+1} = \alpha X_i + W_i, \quad i = 0, 1, 2, \ldots, \quad (1)$$

where the $W_i$’s are i.i.d. $N(0, \sigma^2_W)$ r.v.’s, independent of $X_0$ and of the $Z_i$’s.

Note that the observation noises $Z_i$’s are independent from measurement to measurement. On the other hand, the signal values at different times can be strongly dependent (think of the case when $\alpha$ is close to 1 and the perturbation $W_i$ has a small variance). Thus the signal varies relatively smoothly compared to the observation noise. Nevertheless, the signal is still random: it starts from a random initial condition $X_0$ and each perturbation is random as well.
The filtering problem can be posed as follows:

At each time $n$, given the observations $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, estimate $X_n$.

Note that for estimating the signal value $X_n$ at time $n$, not only the observation $Y_n = y_n$ at time $n$ but the observations at previous times $1, 2, \ldots, n-1$ are also relevant: the previous measurements provide information on $X_{n-1}$ which in turn provide information on $X_n$. The question is how to combine that information with the current observation in a systematic fashion.

The extreme case of a smooth signal is one that doesn’t change at all, i.e.

$$X_i = X \sim N(\mu_0, \sigma_0^2), \quad i = 0, 1, 2, \ldots$$

This is a special case of the general model when $\alpha = 1$ and $\sigma_W^2 = 0$. In these notes, we will focus on developing a solution for this case. Once you understand this simple case, it should not be very difficult to work out the general case when the signal is varying.

**Single Observation**

Let’s first solve a simpler problem when we want to estimate $X$ from a single observation $Y$. The problem is: estimate the signal $X$ given the received signal $Y = y$, where

$$Y = X + Z$$

with the signal $X \sim N(\mu_X, \sigma_X^2)$ and the noise $Z \sim N(0, \sigma_Z^2)$ and $X$ and $Z$ are independent. Note that while the signal can have an arbitrary mean $\mu_X$, the noise is naturally modeled as having zero-mean.

**Posterior Distribution**

As in all the other inference problems, the knowledge about $X$ is captured by the conditional distribution of $X$ given the observations, i.e. the posterior distribution. Since $X$ is continuous in this problem, this distribution is represented by the conditional density of $X$ given $Y = y$. A natural definition for the conditional density,
in analogy to the discrete case, is:

\[ f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}, \]

where \( f_{X,Y} \) is the joint density of \( X \) and \( Y \) and \( f_Y \) is the (marginal) density of \( Y \). This leads to the analog of Bayes’ rule for continuous r.v.’s:

\[ f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}. \] (3)

Now, \( X \sim N(\mu_X, \sigma_X^2) \), so

\[ f_X(x) = \frac{1}{\sqrt{2\pi} \sigma_X} \exp \left[ -\frac{(x - \mu_X)^2}{2\sigma_X^2} \right]. \]

Given \( X = x, Y \sim N(x, \sigma_Z^2) \) so

\[ f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi} \sigma_Z} \exp \left[ -\frac{(y - x)^2}{2\sigma_Z^2} \right]. \]

Substituting into (3):

\[ f_{X|Y}(x|y) = \frac{1}{f_Y(y)} \cdot \frac{1}{2\pi \sigma_X \sigma_Z} \exp \left[ -\frac{(x - \mu_X)^2}{2\sigma_X^2} - \frac{(y - x)^2}{2\sigma_Z^2} \right]. \] (4)

Since we are computing the conditional density of \( X \) given \( Y = y \), let us look at how the expression (4) depends on \( x \), treating \( y \) as a constant. It can be rewritten in this form:

\[ f_{X|Y}(x|y) = c \exp \left[ dx^2 + ex \right], \]

where \( c, d, e \) are constants that do not depend on \( x \). You are asked to verify that any density of this form must be a Normal density, with mean \(-e/(2d)\) and variance \(-1/(2d)\). From this fact, we come to an interesting conclusion: not only is \( X \) Normal, but conditional on \( Y = y, X \) is also Normal! Once we know it is Normal, all we need to do is to compute its mean, say \( \mu_1 \), and its variance, say \( \sigma^2_1 \). They can be computed by figuring out what \( d \) and \( e \) are from (4):

\[ \mu_1 = a \mu_X + (1 - a)y \] (5)

\[ \sigma^2_1 = \frac{1}{\sigma_X^2 + \sigma_Z^2}, \] (6)

where

\[ a := \frac{\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}. \]

**Estimating X**

The situation is summarized in Figure 2. Unconditional on the observation, \( X \) is Normal with mean \( \mu_X \) and variance \( \sigma_X^2 \). Conditional on the observation \( Y = y \), \( X \) is still Normal but with a new mean \( \mu_1 \) and a new variance \( \sigma^2_1 \). How should we guess the value of \( X \)? A natural guess, and one we have used in earlier
problems, is the MAP estimate: this is the value of $x$ where the peak of the conditional density of $X$ given $Y = y$ is attained. In this example, this conditional density is Normal, and so the peak is at the mean of this conditional density, i.e. the conditional mean. So a natural estimate of $X$ given the observation $Y = y$ is:

$$a \mu_X + (1-a)y. \quad (7)$$

Bayesian learning is basically a systematic way of updating knowledge based on observations. We see it very clearly in the above formula. Based on the prior knowledge only, the best estimate of $X$ would be $\mu_X$. After the observation $Y = y$, the estimate is revised to be (7), a weighted combination of $\mu_X$ and $y$. The relative weights placed on the prior knowledge and the observation depend on the signal-to-noise ratio $\text{SNR} := \sigma_X^2 / \sigma_Z^2$. If $\text{SNR}$ is small, then more weight is placed on the prior knowledge. This is intuitive since then the observation is very noisy and not very informative. If $\text{SNR}$ is large, then more weight is placed on the observation.

How do we quantify the residual uncertainty of $X$ after the observation? A natural choice is the variance $\sigma_1^2$ of the conditional density. The larger it is, the more the residual uncertainty. Note that the variance $1/(\sigma_X^{-2} + \sigma_Z^{-2})$ of the conditional density is always small than $\sigma_X^2$, the variance of the prior density of $X$. So observations always reduce uncertainty (which is good to know!).

Recursive Algorithm

With the above groundwork, it is not too difficult to provide a solution to the original problem. Basically, after each observation, we update the prior density to get a posterior density, all of which are Normals. Let’s proceed step by step. Recall that the prior density of $X$ is $N(\mu_0, \sigma_0^2)$. At time 1, we observe $Y_1 = y_1$. The conditional density of $X$ given $Y_1 = y_1$ is $N(\mu_1, \sigma_1^2)$ where, by eqns. (5) and (6),

$$\mu_1 = a_0 \mu_0 + (1-a_0)y_1$$
$$\sigma_1^2 = \frac{1}{\sigma_0^{-2} + \sigma_Z^{-2}},$$

where

$$a_0 := \frac{\sigma_Z^2}{\sigma_0^2 + \sigma_Z^2}.$$
Now, let us proceed inductively. Suppose at time $i$, the conditional density of $X$ given the observations $Y_1 = y_1, \ldots, Y_i = y_i$ is $N(\mu_i, \sigma^2_i)$. Now suppose we are given an additional observation $Y_{i+1} = y_{i+1}$, where

$$Y_{i+1} = X + Z_{i+1}.$$ 

Now, conditional on $Y_1 = y_1, \ldots, Y_i = y_i$, $X$ is $N(\mu_i, \sigma^2_i)$ and $Z_i$ is independent of $X$. So with the added observation $Y_{i+1} = y_{i+1}$ at time $i + 1$, the problem is identical to the single observation problem except that the prior density of $X$ is replaced by $N(\mu_i, \sigma^2_i)$. Hence, conditional on now all the observations up to time $i + 1$, $X \sim N(\mu_{i+1}, \sigma^2_{i+1})$, with:

$$\mu_{i+1} = a_i \mu_i + (1 - a_i) y_i,$$

$$\sigma^2_{i+1} = \frac{1}{\sigma_i^{-2} + \sigma_Z^{-2}},$$

where

$$a_i := \frac{\sigma_Z^2}{\sigma_i^2 + \sigma_Z^2}.$$ 

Here, we again use (5) and (6).

Now that we have the conditional densities, we can compute everything. The estimate $X$ at time $i$ (output of the filter at time $i$) is simply $\mu_i$. The variance of that estimate conditional on all the observations seen so far is $\sigma^2_i$. The estimate at time $i + 1$ is a weighted combination of the previous estimate $\mu_i$ and the new observation $y_{i+1}$, the weights depending on $\sigma^2_i$ and $\sigma^2_Z$. Hence, the outputs of the filter can be computed recursively, without starting from scratch every time.

Note that the principle we are applying here is identical to the one we used in Homework 10 for the problem of learning about the identity of a randomly chosen coin: conditional on the previous observations, we are working in a new sample space with all probabilities calculated based on the conditioning. Using Bayes’ rule, we can compute the posterior distribution. For general Bayesian learning, it is a pain to keep track of the conditional distributions/densities, but the simplification we get in the present problem is that all the conditional densities are Normals, so we only have to keep track of two numbers, the mean and the variance.

**Exercise:** Using the same principles, find a recursive algorithm to solve the problem in the general case when the signal is time varying, as described by eqns. (1) and (2).