5 Bayesian estimation and inference

We started the last chapter by approaching the problem of state estimation from a maximum likelihood perspective. That is, for some hidden states $\mathbf{x}$ and observations $\mathbf{y}$, we formulated a generative model that described a probability distribution $p(\mathbf{y}|\mathbf{x})$ and then we found an estimate $\hat{\mathbf{x}}$ that maximized the probability of our observations. Formally, this is stated as

$$\hat{x} = \arg\max_{\mathbf{x}} \left(p(y \mid \mathbf{x}) \right). \tag{5.1}$$

The problem with our approach was that we could not incorporate a prior belief[1] into this formulation, and this was a serious flaw because everything that we perceive is likely an integration of a prior belief with observations. To remedy the situation, we considered the Kalman framework. In this framework, we found the mixing gain $\mathbf{k}^{(n)}$ that allowed us to integrate our prior belief $\hat{\mathbf{x}}^{(n|n-1)}$ with our observation $\mathbf{y}^{(n)}$ to form a posterior belief $\hat{\mathbf{x}}^{(n|n)}$. The gain that we found was one that minimized the trace of the variance of the posterior uncertainty $P^{(n|n)}$. This uncertainty represented the variance of the Gaussian distribution $p\left(\mathbf{x}^{(n)}\middle|\mathbf{y}^{(n)}\right)$. Finally, we found a relationship between the Kalman gain and maximum likelihood: we found that if we are naïve and have no prior beliefs about the hidden states, then the Kalman gain is in fact the mixing gain that we derived in the maximum likelihood approach.

Our approach thus far is a bit curious because what we are really after is the posterior probability distribution associated with our hidden states $p\left(\mathbf{x}^{(n)}\middle|\mathbf{y}^{(n)}\right)$. That is, we have some prior belief about the state $p\left(\mathbf{x}^{(n)}\right)$, we make an observation $\mathbf{y}^{(n)}$, and now we want to update our belief based on what we observed. We have not shown what this posterior probability distribution looks like. For example, if we apply the Kalman gain to the generative model that had Gaussian noise and form a posterior belief $\hat{\mathbf{x}}^{(n|n)}$, is this the expected value of $\mathbf{x}$ in the distribution $p\left(\mathbf{x}^{(n)}\middle|\mathbf{y}^{(n)}\right)$? Is the uncertainty of our posterior belief $P^{(n|n)}$ the variance in the distribution $p\left(\mathbf{x}^{(n)}\middle|\mathbf{y}^{(n)}\right)$?

[1] Here the term "belief" is not used with its ordinary meaning, but according to an accepted lexicon in statistical learning theory. In this more restricted sense, a belief is an expectation that the learning system has developed either from past experience or has encoded in its initial structural properties. There is not the assumption of conscious awareness associated with the more common use of the word.

Here, we will formulate the posterior and we will see that indeed $\hat{\mathbf{x}}^{(n|n)}$ and $P^{(n|n)}$ are the mean and variance of it. The approach is called *Bayesian state estimation.*

5.1 Bayesian state estimation

Bayesian estimation has its roots on a very simple and fundamental theorem, first discovered by Thomas Bayes shortly before his death, in the mid 18th century. In modern terms things go as follows. Consider two random variables, *x* and *y*. Suppose that *x* can take one of N_X values and *y* can take one of N_Y values. The joint probability of observing $x = x_i$ and $y = y_j$ is $\Pr(x = x_i, y = y_j)$. If *x* and *y* are statistically independent, this joint probability is just the product $\Pr(x = x_i)\Pr(y = y_j)$. If the *x* and *y* are not independent, then one has to multiply the probability that $x = x_i$ given that $y = y_j$ by the probability that $y = y_j$. Of course, this is the most general thing, since the conditional probability $\Pr(x = x_i \mid y = y_j)$ coincides with $\Pr(x = x_i)$ if the two variables are independent. Bayes' theorem is a direct consequence of the intuitive fact that the joint probability is commutative:

$$\Pr(x = x_i, y = y_j) = \Pr(y = y_j, x = x_i).$$

Then, expanding each side with the corresponding expression on conditional probability one obtains

$$\Pr(x = x_i \mid y = y_j)\Pr(y = y_j) = \Pr(y = y_j \mid x = x_i)\Pr(x = x_i).$$

Bayes' theorem is then obtained by rearranging the terms as

$$\Pr(x = x_i \mid y = y_j) = \Pr(y = y_j \mid x = x_i)\frac{\Pr(x = x_i)}{\Pr(y = y_i)}$$

Importantly, this simple algebra applies not only to probability values, but also to probability distributions. So, if *x* and *y* are continuous random variables, then Bayes' theorem allows to derive the relation between the respective distributions as

$$p(x \mid y) = p(y \mid x)\frac{p(x)}{p(y)}$$

If the variable x represents a "model" variable – for example the state of a dynamical system – and y represents an observed variable – for example the output of a sensor - then,

a) $p(y \mid x)$ is the likelihood of an observation given that the underlying model is true;

b) $p(x \mid y)$ is the probability distribution of the model given the observations; this gives the probability that the model is correct "after the fact" that we have collected an observation. Therefore is called the *posterior* distribution.

c) $p(x)$ is the probability distribution of the model independent of any observation, or the *prior* of x.

d) the prior probability to make an observation, $p(y)$, or *marginal* probability, is generally derived as a normalization factor, to insure that all distributions integrate to 1.

In the following discussion, we take advantage not of Bayes' theorem in its standard form, but of the underlying rule expressing the joint probability from the product of the posterior and the marginal distributions, that is from

$$p(x, y) = p(x \mid y)\, p(y).$$

To formulate the posterior distribution, we start with the prior and the likelihood. Say that our prior estimate of the hidden state is normally distributed with mean $\hat{\mathbf{x}}^{(n|n-1)}$ and variance $P^{(n|n-1)}$:

$$p(\mathbf{x}) = N\left(\hat{\mathbf{x}}^{(n|n-1)}, P^{(n|n-1)}\right). \tag{5.2}$$

Further assume that our measurements are related to the hidden states via the following relationship:

$$\begin{aligned} \mathbf{y} &= C\mathbf{x} + \boldsymbol{\varepsilon}_y \\ \boldsymbol{\varepsilon}_y &\sim N(\mathbf{0}, R) \end{aligned} \tag{5.3}$$

Therefore, the expected value and variance of observation $\mathbf{y}$ are:

$$\begin{aligned} E(\mathbf{y}) &= C\hat{\mathbf{x}}^{(n|n-1)} \\ \operatorname{var}(\mathbf{y}) &= C \operatorname{var}(\mathbf{x}) C^T + 2\operatorname{cov}(C\mathbf{x}, \boldsymbol{\varepsilon}) + \operatorname{var}(\boldsymbol{\varepsilon}) \\ &= CP^{(n|n-1)}C^T + R \end{aligned}$$

We have the distribution $p(\mathbf{y})$:

$$p(\mathbf{y}) = N\left(C\hat{\mathbf{x}}^{(n|n-1)}, CP^{(n|n-1)}C^T + R\right) \tag{5.4}$$

Our next step is to compute the joint probability distribution $p(\mathbf{x}, \mathbf{y})$. To form this distribution, we need to know the covariance between $\mathbf{x}$ and $\mathbf{y}$, which is computed below:

$$\begin{aligned}\operatorname{cov}(\mathbf{y},\mathbf{x}) &= \operatorname{cov}(C\mathbf{x}+\boldsymbol{\varepsilon},\mathbf{x}) = E\left[\left(C\mathbf{x}+\boldsymbol{\varepsilon}-CE(\mathbf{x})\right)\left(\mathbf{x}-E(\mathbf{x})\right)^T\right] \\ &= E\left[C\mathbf{x}\mathbf{x}^T - C\mathbf{x}E(\mathbf{x})^T - CE(\mathbf{x})\mathbf{x}^T + CE(\mathbf{x})E(\mathbf{x})^T + \boldsymbol{\varepsilon}\mathbf{x}^T - \boldsymbol{\varepsilon}E(\mathbf{x})^T\right] \\ &= CE\left[\mathbf{x}\mathbf{x}^T\right] - CE(\mathbf{x})E(\mathbf{x})^T \\ &= C\operatorname{var}(\mathbf{x}) = CP^{(n|n-1)}\end{aligned}$$

The joint probability distribution becomes:

$$p(\mathbf{x},\mathbf{y}) = p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = N\left(\begin{bmatrix}\mathbf{x}^{(n|n-1)}\\ C\mathbf{x}^{(n|n-1)}\end{bmatrix}, \begin{bmatrix}P^{(n|n-1)} & P^{(n|n-1)}C^T\\ CP^{(n|n-1)} & CP^{(n|n-1)}C^T + R\end{bmatrix}\right) \tag{5.5}$$

The Gaussian distribution in Eq. (5.5) is equal to the product of the posterior probability $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{y})$, that is:

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})\,p(\mathbf{y}) \tag{5.6}$$

The item that we are looking for is the posterior probability $p(\mathbf{x}|\mathbf{y})$. We have the joint probability $p(\mathbf{x},\mathbf{y})$ (in Eq. 5.5), and we also have $p(\mathbf{y})$ in Eq. (5.4). If we could factor Eq. (5.5) so that it becomes a multiplication of two normal distributions, one of which is $p(\mathbf{y})$, then we will have the posterior probability that we are looking for.

The general problem that we are trying to solve is to factor a normal distribution $p(\mathbf{x},\mathbf{y})$. In this distribution, $\mathbf{x}$ is a $p\times 1$ vector and $\mathbf{y}$ is a $q\times 1$ vector, and the distribution has the following general form:

$$p(\mathbf{x},\mathbf{y}) = N\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right) \tag{5.7}$$

in which our *marginal distributions* are

$$\begin{aligned}\mathbf{x} &\sim N(\boldsymbol{\mu}_x, \Sigma_{11}) \\ \mathbf{y} &\sim N(\boldsymbol{\mu}_y, \Sigma_{22})\end{aligned} \tag{5.8}$$

and $\operatorname{cov}(\mathbf{x},\mathbf{y}) = \Sigma_{12} = \Sigma_{21}$. We hope to find a *conditional distribution* $p(\mathbf{x}|\mathbf{y})$.

Our first step is to block-diagonalize the variance-covariance matrix in Eq. (5.5). To see how to do this, assume we have a matrix *M* composed of following blocks:

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix} \tag{5.9}$$

If we right and left multiply *M* by the following two matrices, each of which has an identity determinant, we will end up with a diagonalized version of *M*:

$$\begin{aligned} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} M \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} &= \begin{bmatrix} E - FH^{-1}G & F - FH^{-1}H \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \\ &= \begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix} \end{aligned} \tag{5.10}$$

The term M/H is called the Schur complement of *M* and is defined as:

$$M/H \equiv E - FH^{-1}G \tag{5.11}$$

If we now take the determinant of the matrix in Eq. (5.10), we have:

$$\begin{aligned} \det(M) &= \det\left(E - FH^{-1}G\right)\det(H) \\ &= \det(M/H)\det(H) \end{aligned} \tag{5.12}$$

The above equality relies on the fact that the determinant of a block-triangular matrix is the product of the determinants of the diagonal blocks.

Our second step is to compute the inverse of matrix *M*. We will do this by taking advantage of the diagonalization that we did in Eq. (5.10). Suppose that we call X the matrix that we left multiplied *M* in Eq. (5.10) , and the right multiple as *Z*. Eq. (5.10) is simply:

$$\begin{aligned} XMZ &= W \\ Z^{-1}M^{-1}X^{-1} &= W^{-1} \\ M^{-1} &= ZW^{-1}X \end{aligned} \tag{5.13}$$

That is, the inverse of matrix *M* is:

$$M^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \tag{5.14}$$

In the expression that describes a normal distribution, we have a determinant of the covariance matrix, and we have an inverse of the covariance matrix. We will use the results in Eq. (5.12) to factor the determinant term, and the result in Eq. (5.14) to factor the inverse term.

The distribution that we wish to factor has the following form:

$$p(\mathbf{x},\mathbf{y}) = (2\pi)^{-(p+q)/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix} \right)^T \Sigma^{-1} \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix} \right) \right\} \quad (5.15)$$

where Σ is the variance-covariance matrix of the above distribution, i.e.,

$$\Sigma \equiv \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Using Eq. (5.12), the determinant and the constants in the joint probability distribution can be factored as:

$$(2\pi)^{-(p+q)/2} \left| \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right|^{-1/2} = (2\pi)^{-p/2} |\Sigma / \Sigma_{22}|^{-1/2} (2\pi)^{-q/2} |\Sigma_{22}|^{-1/2} \quad (5.16)$$

In Eq. (5.16), the term Σ / Σ_{22} is the Schur complement of the matrix Σ. Using Eq. (5.14), the exponential term in the joint probability distribution can be factored as:

$$\begin{aligned} &\exp\left\{ -\frac{1}{2} \begin{bmatrix} \mathbf{x}-\boldsymbol{\mu}_x \\ \mathbf{y}-\boldsymbol{\mu}_y \end{bmatrix}^T \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & {\Sigma_{22}}^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{x}-\boldsymbol{\mu}_x \\ \mathbf{y}-\boldsymbol{\mu}_y \end{bmatrix} \right\} = \\ &\exp\left\{ -\frac{1}{2} \left(\mathbf{x} - \left(\boldsymbol{\mu}_x + \Sigma_{12}\Sigma_{22}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_y \right) \right) \right)^T (\Sigma/\Sigma_{22})^{-1} \left(\mathbf{x} - \left(\boldsymbol{\mu}_x + \Sigma_{12}\Sigma_{22}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_y \right) \right) \right) \right\} \\ &\exp\left\{ -\frac{1}{2} \left(\mathbf{y} - \boldsymbol{\mu}_y \right)^T {\Sigma_{22}}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_y \right) \right\} \end{aligned} \quad (5.17)$$

Therefore, we factored the joint probability distribution into two terms:

$$\begin{aligned} p(\mathbf{x},\mathbf{y}) &= N\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \\ &= N\left(\boldsymbol{\mu}_x + \Sigma_{12}\Sigma_{22}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_y \right), \Sigma/\Sigma_{22} \right) N\left(\boldsymbol{\mu}_y, \Sigma_{22} \right) \\ &= N\left(\boldsymbol{\mu}_x + \Sigma_{12}\Sigma_{22}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_y \right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right) N\left(\boldsymbol{\mu}_y, \Sigma_{22} \right) \\ &= p\left(\mathbf{x} | \mathbf{y} \right) p\left(\mathbf{y} \right) \end{aligned} \quad (5.18)$$

The posterior probability that we were looking for is the first of the two normal distributions in Eq. (5.18):

$$p\left(\mathbf{x} | \mathbf{y} \right) = N\left(\boldsymbol{\mu}_x + \Sigma_{12}\Sigma_{22}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_y \right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right) \quad (5.19)$$

In the distribution of Eq. (5.19), we have the following terms:

$$\begin{aligned}
\boldsymbol{\mu}_x &= \hat{\mathbf{x}}^{(n|n-1)} \\
\boldsymbol{\mu}_y &= C\hat{\mathbf{x}}^{(n|n-1)} \\
\Sigma_{11} &= P^{(n|n-1)} \\
\Sigma_{12} &= \Sigma_{21} = P^{(n|n-1)}C^T \\
\Sigma_{22} &= CP^{(n|n-1)}C^T + R
\end{aligned}$$

Re-writing the distribution in Eq. (5.19) with the above terms, we have our usual Kalman filter estimate of the posterior:

$$\begin{aligned}
E\left[\mathbf{x}^{(n)}\middle|\mathbf{y}^{(n)}\right] &= \hat{\mathbf{x}}^{(n|n-1)} + P^{(n|n-1)}C^T\left(CP^{(n|n-1)}C^T + R\right)^{-1}\left(\mathbf{y}^{(n)} - C\hat{\mathbf{x}}^{(n|n-1)}\right) \\
&= \hat{\mathbf{x}}^{(n|n-1)} + K^{(n)}\left(\mathbf{y}^{(n)} - C\hat{\mathbf{x}}^{(n|n-1)}\right) \\
\text{var}\left[\mathbf{x}^{(n)}\middle|\mathbf{y}^{(n)}\right] &= P^{(n|n-1)} - P^{(n|n-1)}C^T\left(CP^{(n|n-1)}C^T + R\right)^{-1}CP^{(n|n-1)} \\
&= \left(I - K^{(n)}C\right)P^{(n|n-1)}
\end{aligned} \tag{5.20}$$

So indeed we see that the Kalman gain in Eq. (4.55) is the term $\Sigma_{12}\Sigma_{22}^{-1}$ in Eq. (5.19), the posterior belief $\hat{\mathbf{x}}^{(n|n)}$ is the expected value of $p\left(\mathbf{x}\middle|\mathbf{y}^{(n)}\right)$, and the uncertainty of our posterior belief $P^{(n|n)}$ is the variance of $p\left(\mathbf{x}\middle|\mathbf{y}^{(n)}\right)$. By using Kalman's approach, we are computing the mean and variance of the posterior probability of the hidden state that we are trying to estimate.

5.2 Causal inference

Recall that in the hiking problem that we described in the last chapter we had two GPS devices that measured our position. We combined the reading from the two devices to form an estimate of our location. For example, in Fig. 4.8 the estimate of our location ended up being somewhere in between the two readings. This approach makes sense if our two readings are close to each other (i.e., if the two GPS devices are providing us with estimates that pretty much agree with each other). However, we can hardly be expected to combine the two readings if one of them is telling us that we are on the north bank of the river and the other is telling us that we are on the south bank. We know that we are not somewhere in the river! In this case the idea of combining the two readings makes little sense.

Consider another example. Say you are outside and see lightening and soon afterwards hear thunder. It seems reasonable that your two sensors (vision and audition) were driven by a single cause: a lightening occurring at a specific location. However, if the two sensory events are separated by a long time interval or the sound appears to come from a different direction than the light, then you would be less likely to believe that there was a single cause for the two observations. In principle, when the various sensors in our body report an event, the probability that there was a single source responsible for them should depend on the temporal and spatial consistency of the readings from the various sensors. This probability of a single source, i.e., the probability of a single cause, should then play a significant role in whether our brain will combine the readings from the sensors or leave them apart.

Wallace and colleagues (2004) examined this question by placing people in a room where LEDs and small speakers were placed around a semi-circle (Fig. 5.1A). A volunteer sitting in the center of the semi-circle held a pointer in hand. The experiment began by the volunteer fixating a location (fixation LED, Fig. 5.1A). An auditory stimulus was presented from one of the speakers, and then one of the LEDs was turned on 200, 500, or 800ms later. The volunteer estimated the location of the sound by pointing (pointer, Fig. 5.1A). Then he pressed a switch with their foot if they thought that the light and the sound came from the same location. The results of the experiment are plotted in Fig. 5.1B and C. The combination of the audio and visual stimuli in a single percept – perception of unity - was strongest when the two events occurred in close temporal and spatial proximity. Importantly, when the volunteers perceived a common source, their localization of the sound was highly affected by the location of the light. That is, if x_s represents the location of the sound and x_v represents the location of the LED, the estimate of the location of the sound $\hat{x}_s$ (i.e., where the subject pointed) was biased by x_v when the volunteer thought that there was a common source (Fig. 5.1C). This bias fell to near zero when the volunteer perceived light and sound to originate from different sources.

The experiment in Fig. 5.1 suggests that when our various sensory organs produce reports that are temporally and spatially in agreement, we tend to believe that there was a single source that was responsible for both observations (Fig. 5.2A). In this case, we combine the readings from the sensors to estimate the state of the source. On the other hand, if our sensory measurements are temporally or spatially inconsistent, then we view the events as having disparate sources (Fig. 5.2A), and we do not combine the sources. Therefore, the nature of our belief as to whether there

was a common source or not is not black or white. Rather, there is some probability that there was a common source. In that case, this probability should have a lot to do with how we combine the information from the various sensors.

The idea is that in principle, there are many generative models that could explain our observations. For example, we could have a model that says that the two observations come from the same source. We could also have another model that says that the two observations are independent. The one that we pick, or rather the probabilities that we assign to the various potential generative models, will determine how we will form our belief.

Konrad Kording and colleagues (Kording et al., 2007) suggested a simple way to frame this problem. Suppose that the binary random variable z specifies whether there is a single source ($z=1$), or whether there are two distinct sources that drive our sensors ($z=0$). If $\Pr(z=1|\mathbf{y})=1$, then our visual and sound measurements are reflecting a common source:

$$\begin{aligned}\mathbf{y} &= \begin{bmatrix} y_s \\ y_v \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_s \\ x_v \end{bmatrix} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(0,R) \\ &= C_1\mathbf{x} + \boldsymbol{\varepsilon}\end{aligned} \tag{5.21}$$

On the other hand, if $\Pr(z=1|\mathbf{y})=0$, then our measurements are reflecting different sources:

$$\begin{aligned}\mathbf{y} &= \begin{bmatrix} y_s \\ y_v \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_s \\ x_v \end{bmatrix} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(0,R) \\ &= C_2\mathbf{x} + \boldsymbol{\varepsilon}\end{aligned} \tag{5.22}$$

Starting with a prior belief about the location of the two stimuli $\mathbf{x} \sim N(\boldsymbol{\mu}, P)$, and a prior belief about the probability of a common source $\Pr(z=1)$, we can compute the posterior probability of a common source after making a measurement $\mathbf{y}$:

$$p(z=1|\mathbf{y}) = \frac{p(\mathbf{y}|z=1)\Pr(z=1)}{p(\mathbf{y}|z=0)\Pr(z=0)+p(\mathbf{y}|z=1)\Pr(z=1)} \tag{5.23}$$

The probability distribution of our measurements given that there is a common source can be computed from Eq. (5.21) as follows:

$$p(\mathbf{y}|z=1) = N\left(C_1\boldsymbol{\mu}, C_1PC_1^T + R\right) \tag{5.24}$$

Similarly, from Eq. (5.22) we have:

$$p(\mathbf{y}|z=0) = N\left(C_2\boldsymbol{\mu}, C_2PC_2^T + R\right) \tag{5.25}$$

In Fig. 5.2B we have plotted $p(z=1|\mathbf{y})$ for various values of our two sensory measurements (in computing Eq. 5.23, we assumed that $\Pr(z=1)=0.5$, that is, a common source was just as likely as two independent sources). When the two measurements y_s and y_v are close to each other, the probability of a common source is nearly one. When they are far from each other, the probability is close to zero. That is, as the spatial disparity between the two measurements increases, it is less likely that we are observing the consequences of a common source (Fig. 5.2C).

Let us now return to the experiment in Fig. 5.1A in which we hear a sound and see a light and need to know where the sound came from. Eq. (5.23), as plotted in Fig. 5.2C, gives us a probability of a common source as a functional of spatial disparity between the two sensory measurements. The question is how to use the probability of a common source to compute the location of the auditory stimulus. Let us suppose that $\Pr(z=1|\mathbf{y})=1$. In that case we would use the single source model (Eq. 5.21) to find the posterior probability $p(\mathbf{x}|\mathbf{y})$. Our best estimate for $\mathbf{x}$ is the expected value of this posterior distribution, which is the usual Kalman estimate derived from the model in Eq. (5.21), i.e., one in which we combine the two measurements to estimate the location of the sound. On the other hand, if $\Pr(z=1|\mathbf{y})=0$, then Eq. (5.22) is the model that we should use, and the Kalman gain here would treat the two measurements as independent and not mix them to estimate the location of the sound. If the probability of a common source is somewhere between 0 and 1, then a rational thing to do would be to use this probability to weigh each of our two estimates for the location of the sound. To explain this in detail, let us begin by computing $p(\mathbf{x},\mathbf{y})$. If there is a common source, then from Eq. (5.21) we have:

$$p(\mathbf{x},\mathbf{y}) = N\left(\begin{bmatrix}\boldsymbol{\mu}\\ C_1\boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}P & PC_1^T\\ C_1P & C_1PC_1^T + R\end{bmatrix}\right) \tag{5.26}$$

This implies that if there was a common source, our best estimate is the usual mixing of the two sources. From Eq. (5.20) we have:

$$E\left[\mathbf{x}|\mathbf{y}, z=1\right] = \boldsymbol{\mu} + PC_1^T\left(C_1PC_1^T + R\right)^{-1}\left(\mathbf{y} - C_1\boldsymbol{\mu}\right) \tag{5.27}$$

On the other hand, if there is not a common source, our best estimate is:

$$E\left[\mathbf{x}|\mathbf{y}, z=0\right] = \boldsymbol{\mu} + PC_2^T\left(C_2PC_2^T + R\right)^{-1}\left(\mathbf{y} - C_2\boldsymbol{\mu}\right) \tag{5.28}$$

In general then, our estimate of $\mathbf{x}$ should be a mixture of the above two estimates, weighted by the probability of a common source:

$$\begin{aligned} &\Pr\left(z=1|\mathbf{y}\right) = a \\ &E\left[\mathbf{x}|\mathbf{y}\right] = aE\left[\mathbf{x}|\mathbf{y}, z=1\right] + \left(1-a\right)E\left[\mathbf{x}|\mathbf{y}, z=0\right] \end{aligned} \tag{5.29}$$

We have plotted an example of this estimate of $\hat{x}_s$ (location of the speaker) in Fig. 5.2D. We assumed that the actual location of the sound (i.e., the speaker) was always at zero ($y_s = 0$), while the LED (y_v) was located at various displacements. We see that when the two sources are near each other, i.e., when y_v values are small, the estimate of the location of sound $\hat{x}_s$ is highly influenced by the location of the LED. However, as the LED moves farther away, the estimate of the speaker location returns back to zero. When there is a large discrepancy between the two measurements y_v and y_s, there is little chance that they are coming from a single source, and so the system does not combine the two measures.

We can now see that in many of the previous experiments in which people were presented multiple cues and they performed 'optimally' by combining the cues (some of these experiments were reviewed in the previous chapter, as in Fig. 4.7 and Fig. 4.9), they were doing so because the disparity between the cues was small. If the disparity is large, it is illogical to combine the two cues. The single-source models in those experiment are special cases of the more general causal inference model (Fig. 5.2) (Kording et al., 2007).

In summary, if we believe that our sensors are reporting the consequences of a common event, then our brain combines information from our various sensors. This belief regarding a common source is itself driven by the disparity between the measurements (i.e., their temporal and spatial agreement).

5.3 The influence of priors

When at the coffee shop the attendant hands you a cup full of tea, your brain needs to guess how heavy the drink is. This guess need to be accurate, otherwise you would have trouble grasping the cup (activating your finger muscles in a way that they does not let it slip out of your hand),

and holding it steady (activating your arm muscles so the cup does not rise up in the air or fall down). The only cue that you have is the size information provided by vision. Fortunately, the other piece of information is the prior experience that you have had with cups of tea. The fact that most people have little trouble holding cups that are handed to them in coffee shops suggests that they are making accurate estimates of weight. How are they making these guesses?

A useful place to start is by stating in principle how people should make guesses, whether it be regarding weight of a cup of tea, or something else. Consider the cup of tea problem. Suppose we label the weight of the cup of tea as x. We have some prior belief about the distribution of these weights, $p(x)$, i.e., how much cups of tea weigh in general. We have some prior belief about the relationship between visual property (size) of a tea cup s and the weight of tea that it can hold, $p(s|x)$. And we have some prior belief about the distribution of tea cup sizes $p(s)$. Then our guess about weight of this particular cup of tea should be based on the posterior distribution:

$$p(x|s) = \frac{p(s|x)\,p(x)}{p(s)} \tag{5.30}$$

In Eq. (5.30), what we are doing is transforming a prior belief $p(x)$ about the state of something (weight of a cup) into a posterior belief, after we made an observation or measurement (in this case s, the size of the cup that we saw). Let us do a thought experiment to consider how prior beliefs should affect people's guesses about weights of cups. In America, people are familiar with the rather large cups that are used for serving soft drinks and other refreshments. For example, in some convenient stores there are drinks that are called 'super big gulp', and they hold something like 1.2 liters of soda. [Single serve bottles in which fruit juice is sold in America tend to labeled as 'family-size' in Europe.] The prior distribution $p(x)$ for someone who frequents such places in America would be skewed toward large weights. In contrast, someone from another country in which big gulps are not available would have $p(x)$ skewed toward smaller masses. If we now take these two people to a lab and present them with a regular sized soft drink cup, upon visual inspection the subject for whom $p(x)$ is skewed toward heavier weights (the American fellow) should estimate the weight of the cup to be heavier than the subject for whom $p(x)$ is skewed toward lighter weights (the European chap). The belief about the larger weight

should be reflected in the larger force that the American fellow will use to pick up the cup. That is, the prior belief should affect the prediction.

Indeed, prior beliefs do play a very strong role in how people interact with objects in everyday scenarios. For example, consider the task of using your fingers to pick up a small object as compared to picking up a slightly larger object. To pick up an object, you will need to apply a grip force (so the object does not slip out of your fingers) and a load force (so you can lift the object), as shown in Fig. 5.3A. Suppose you walk into a lab and are given an instrumented device like that shown in Fig. 5.3A. This device is attached to either a small box or a large box. You should apply larger grip and load forces to the larger object. This is indeed what Andrew Gordon, Roland Johansson, and colleagues (1991) observed when they presented volunteers with three boxes that *weighed exactly the same*, but were of different sizes. People applied a larger grip force and a larger load force to lift the larger box (Fig. 5.3B). The result was the familiar scenario in which you go to pick up a bottle that you think is full, but is actually empty: the hand accelerates up faster than you intended.

Another way to explore prior beliefs about physics of objects is with regard to how objects move in a gravitational field: objects fall with a constant acceleration of $g = 9.8$ m/s^2. For example, when a ball is released from rest and falls toward your hand, your prediction regarding when it will reach you will determine when you will open your hand. Almost all of us spend our entire lives here on earth, so presumably our brain has formed an internal model of falling objects. Let us briefly sketch this internal model. Suppose the state (position and velocity) of the ball is labeled as $\mathbf{x}(t)$, our measurement of that state is labeled $\mathbf{y}(t)$, and our goal is to predict the future state $\mathbf{x}(t+\Delta)$. In a 1g environment, the state of the ball can be modeled as:

$$\begin{bmatrix} x(t+\Delta) \\ \dot{x}(t+\Delta) \end{bmatrix} = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ g\Delta \end{bmatrix} + \boldsymbol{\varepsilon}_x \tag{5.31}$$

Written in a more compact way, we have:

$$\begin{aligned} \mathbf{x}(t+\Delta) &= A\mathbf{x}(t) + \mathbf{b} + \boldsymbol{\varepsilon}_x & \boldsymbol{\varepsilon}_x &\sim N(0, Q) \\ \mathbf{y}(t) &= \mathbf{x}(t) + \boldsymbol{\varepsilon}_y & \boldsymbol{\varepsilon}_y &\sim N(0, R) \end{aligned} \tag{5.32}$$

If we assume that the noises $\boldsymbol{\varepsilon}_x$ and $\boldsymbol{\varepsilon}_y$ are Gaussian, then we can use the Kalman framework to estimate the state of the ball. We start with a prior belief, described by mean $\hat{\mathbf{x}}(t)$ and variance

$P(t)$. We use this prior belief to predict the expected value of what our sensors should be measuring:

$$\hat{\mathbf{y}}(t) = \hat{\mathbf{x}}(t) \tag{5.33}$$

The difference between what we expected and what we measured allows us to update our estimate of the current state and predict the future. The expected values of our predictions are:

$$\begin{aligned} \hat{\mathbf{x}}(t|t) &= \hat{\mathbf{x}}(t) + K\left(\mathbf{y}(t) - \hat{\mathbf{y}}(t)\right) \\ \hat{\mathbf{x}}(t+\Delta|t) &= A\hat{\mathbf{x}}(t|t) + \mathbf{b} \end{aligned} \tag{5.34}$$

In Eq. (5.34), K is the Kalman gain, which depends on our prior uncertainty (variance):

$$K = P(t)\left(P(t) + R\right)^{-1} \tag{5.35}$$

The variance of our posterior probability distribution is:

$$\begin{aligned} P(t|t) &= (I - K)P(t) \\ P(t+\Delta|t) &= AP(t|t)A^T + Q \end{aligned} \tag{5.36}$$

Now if you are an astronaut in an orbiting ship, this 1g internal model should bias your predictions about fall of objects in space. In particular, you should predict that the falling ball will reach you earlier than in reality. Joe McIntyre, Mirka Zago, Alan Berthoz, and Francesco Lacquaniti (2001) tested this idea by having astronauts catch balls in both 0g and 1g. An example of a ball's trajectory in 1g and 0g is shown in Fig. 5.3C. Suppose that the ball starts with a non-zero initial velocity. In 1g, the ball accelerates. However, in 0g the ball velocity remains constant (black lines in Fig. 5.3C). If one uses a 1g internal model to predict the state of the ball in 0g, one would predict that the ball will reach the hand sooner than in reality. That is, the prior belief about the behavior of the ball will produce an earlier preparation for ball hitting the hand. McIntyre et al. (2001) quantified this reaction to the ball motion by recording EMG activity from hand and elbow muscles. When the data were aligned to the moment of ball impact on the hand, they saw that the astronauts in 0g prepared the hand much sooner than in 1g, suggesting that they expected the ball to hit their hand earlier. Interestingly, this pattern continued for the 15 days that the astronauts were in space. That is, the 15 days of being in space was not enough to significantly alter the 1g internal model. Although different from the optimal response in 0g, the anticipatory behavior was not considered by the astronauts' brain as an error requiring correction.

(The ability of a good baseball pitcher to strike out a batter relies to a great extent on the prior belief that batters have regarding gravity and how it should affect the ball during its flight. In a

fastball, the ball has backspin, giving it lift so that it falls slower than expected, whereas in a curve ball, the ball has topspin, giving it downward force so that it falls faster than expected. The force caused by rotation of the ball is called Magnus force, and Isaac Newton himself studied it on a tennis ball.)

These two examples of picking up objects and predicting state of falling balls demonstrate that our brain relies on prior experience to make predictions. You do not need to read this book to know this point, as it is obvious. The more useful question is whether the process of prediction resembles Bayesian state estimation, as in Eqs. (5.30) and (5.34). To explore this question, an interesting experiment was performed by Harm Slijper, Janneke Richter, Eelco Over, Jeroen Smeets, and Maarten Frens (2009). Their idea was that the prior statistics of everyday movements might affect how people move a computer mouse to a given stimulus. That is, the prior statistics should bias the response to the stimulus. They installed 'spyware' software on the computers of a group of consenting volunteers and recorded their mouse movements on random days over a 50 day period. The distribution of movement amplitudes is shown in Fig. 5.4A. The majority of movements were 3mm or less. The distribution of movement directions, represented as an angle θ_e (angle of a line connecting the start to the endpoint), is plotted in Fig. 5.4B. The endpoint directions θ_e were clustered along the primary axes, i.e., most of the movements had endpoints that were up/down or left/right with respect to start position. Slijper et al. (2009) noticed that while some of the movements were straight, many of the movements had an initial angle θ_i that was somewhat different than θ_e. This is illustrated by the cartoon drawing in Fig. 5.4D. They wondered whether this difference $\theta_i - \theta_e$ was due to the distribution of movements that people made. Their idea was that the movements that were straight were the ones that people tended to repeat a lot, whereas movements that were not straight were 'attracted' toward the nearby movement directions that were repeated a lot.

When a movement is not straight, there is a difference between θ_i and θ_e. To explain why there was a difference between θ_i and θ_e for some movements but not others, Slijper et al. (2009) posed the problem in the following way. Suppose that given a desired endpoint at direction θ_e, the probability of moving the mouse in an initial direction θ_i is specified by $p\left(\theta_i \middle| \theta_e\right)$. This distribution can be written as:

$$p(\theta_i|\theta_e) = \frac{p(\theta_e|\theta_i)\,p(\theta_i)}{p(\theta_e)} \tag{5.37}$$

The prior probability of moving in an initial direction is specified by $p(\theta_i)$, and is plotted in Fig. 5.4C. The most frequent initial movement directions are along the left/right axis. That is, the prior has a large peak at 0^o and 180^o degrees. Next, they assumed that the likelihood $p(\theta_e|\theta_i)$ was simply a normal with mean at θ_i and variance of a few degrees. Now suppose that we consider making a movement to an endpoint θ_e at 10^o. The prior distribution $p(\theta_i)$ has a very large peak at $\theta_i = 0$. Even though the likelihood $p(\theta_e|\theta_i)$ has its peak at $\theta_e = \theta_i$, making it so that the maximum likelihood estimate is simply $\theta_e = \theta_i$ (i.e., the movement should be straight to the target), the strong prior at $\theta_i = 0$ will bias the posterior probability. Intuitively, we can see that the initial angle θ_i would be biased toward the large peak at 0^o in the prior distribution $p(\theta_i)$. That is, $p(\theta_i|\theta_e)$ would have its expected value somewhere between 0^o and 10^o. This makes the error in movement direction $\theta_i - \theta_e$ negative. The expected value of Eq. (5.37), $E\left[\theta_i - \theta_e|\theta_e\right]$ is plotted as the dashed line in Fig. 5.4D. This theoretical prediction matched reasonably well with the actual error $\theta_i - \theta_e$, plotted as the solid line in Fig. 5.4D.

It is instructive to approach this problem a bit more rigorously because by doing so we can consider probability distribution of a random variable defined on the circle. The circular normal distribution is:

$$p(\theta) = \frac{1}{2\pi I_0(\kappa)} \exp\left(\kappa \cos(\theta - \mu)\right) \tag{5.38}$$

In Eq. (5.38), the function $I_0(\kappa)$ is a normalization constant, called the modified Bessel function of order zero. The mean of the distribution in Eq. (5.38) is at μ and the variance increases with decreasing κ, as shown with an example in Fig. 5.5A. We can approximate the prior probability $p(\theta_i)$, i.e., the data in Fig. 5.4C, as a sum of four circular normal distributions with means at μ_1, μ_2, μ_3, and μ_4, with $\mu_i = (i-1)\frac{\pi}{2}$, and a uniform distribution, normalized for the sum to have an integral of one:

$$p(\theta_i) = \frac{1}{m}\left[b + \sum_{j=1}^{4} \frac{a_j}{2\pi I_0(\kappa)} \exp\left(\kappa \cos\left(\theta_i - \mu_j\right)\right) \right] \tag{5.39}$$

Our model of the prior $p(\theta_i)$ is plotted in Fig. 5.5B. Let us consider what happens when we intend to move to endpoint $\theta_e = 0.1$ radians (about 6°). What will be the initial movement direction θ_i? The prior $p(\theta_i)$ has a large peak at $\theta_i = 0$. The posterior distribution $p(\theta_i | \theta_e = 0.1)$ is approximately a Gaussian, with its peak at around 3°, which is the expected value of the posterior. The initial movement direction is skewed toward the prior.

In summary, our prior experience with everyday objects like tea cups, balls, and computer devices, produce internal models of physics that appear to strongly affect how we interact with these objects. These internal models act as priors that bias our ability to use current observations. If the internal models are correct (as in the distribution of weight of tea cups), they aid in control because they allow us to correctly estimate property of the current object. However, if the internal models are incorrect (as in the 1g physics being applied by the astronaut to a 0g environment), then we make mistakes in our estimations. This implies that priors have to continuously change with experience. That is, our internal models need to continuously learn from observations. We will pick up this topic in the next chapter when we consider the problem of adaptation.

5.4 The influence of priors on cognitive guesses

If we extend our interest a bit outside of motor control, we find evidence that in situations in which people make a 'cognitive' guess about an ordinary thing, their guess appears to be consistent with a Bayesian framework. Let us consider the problem of guessing the lifespan x of a person (i.e., how many years someone will live), given that they are now t years old. Thomas Griffiths and Joshua Tenenbaum (2006) asked a large group of undergraduates to guess the lifespan of someone who is now 18, 39, 61, 83, or 96 years old (each student made a guess only about a single current age t). Their results are shown in Fig. 5.6A. The students guessed that the 18 and 39 year olds would probably live to be around 75, which is about the mean lifespan of a male in the US. The 61 year old will probably live a little bit longer, around 76 years, but the 83 year old will likely live to around the age of 90 and the 96 year old will likely live to around the age of 100. The interesting thing about these guesses is the shape of the function that specifies

the guess about lifespan as a function of current age (the line in Fig. 5.6A): the line starts out flat, and then rises with a slope that is always less than one. Importantly, according to the students who took this survey, the 96 year old has less time to live than the 83 year old.

This pattern of guessing of course makes a lot of sense. However, Griffiths and Tenenbaum showed that it is exactly how you should guess if you have prior beliefs about life spans that resemble reality (as shown in Fig. 5.6B). Let us go through the mathematics and build a model of this guessing process. Suppose that we model the relative frequency of life spans using a Gaussian function:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right) \qquad \mu = 75, \quad \sigma = 15 \tag{5.40}$$

The above distribution is plotted in Fig. 5.6C. [In using this estimate, we are ignoring the fact that the actual distribution of life spans cannot be Gaussian, as the variable x can only be positive and cannot be greater than some finite maximal value. Furthermore, the actual distribution includes a large number of infants who died near birth, which means that the actual distribution (Fig. 5.6B) is skewed toward dying young.] Further suppose that we model the conditional probability of someone currently being t years old, given that their lifespan is x years old:

$$p(t|x) = \frac{1}{x} \text{ if } x \geq t, \text{ 0 otherwise} \tag{5.41}$$

Eq. (5.41) is our likelihood. Here, it implies that if the lifespan is 70 years, then the likelihood of currently being at any age less than or equal to 70 is simply $1/70$, but of course 0 for values larger than 70. Finally, we model the probability of someone being currently at any age t as:

$$\begin{aligned} p(t) &= \int_0^\infty p(t|x)p(x)\,dx \\ &= \int_t^\infty p(t|x)p(x)\,dx \end{aligned} \tag{5.42}$$

Eq. (5.42) is called the marginal probability. It describes the age distribution of people who are alive today, as shown in Fig. 5.6D. The integral's boundaries in Eq. (5.42) are at t and ∞ (rather than 0 and ∞) because of the constraint on x in Eq. (5.41). The posterior probability becomes:

$$\text{if } x \geq t \quad p(x|t) = \frac{\frac{1}{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}{\int_t^\infty p(t|x)p(x)\,dx} \tag{5.43}$$

If $x < t$, then of course $p(x|t) = 0$. Now suppose that we know that a person is 30 years old. The posterior probability $p(x|t=30)$ is plotted in Fig. 5.6E. What should be our guess about this person's lifespan? Say that we guess $\hat{x} = 35$. The probability of this person living 35 years or less is:

$$\Pr(x \leq 35|t=30) = \int_{30}^{35} p(x|t=30)dx \tag{5.44}$$

The probability of this person living more than 35 years is:

$$\Pr(x > 35|t=30) = \int_{36}^{\infty} p(x|t=30)dx \tag{5.45}$$

By looking at the area under the curve in Fig. 5.6E for $p(x|t=30)$, it should be clear that the probability of living 35 years or longer is a lot higher than the probability of living 35 years or less. Therefore, $\hat{x} = 35$ is a bad guess. The best guess $\hat{x}$ that we could make is one that makes the probability of living less than $\hat{x}$ equal to the probability of living longer than $\hat{x}$. The $\hat{x}$ that we are looking for is the median of the distribution $p(x|t=30)$. If the median value is labeled as m, then

$$\Pr(x \leq m|t) = \Pr(x \geq m|t) = \int_{-\infty}^{m} p(x|t)dx = \frac{1}{2} \tag{5.46}$$

Because $p(x|t=30)$ is quite similar to a Gaussian, peak of the density as well as its mean and median all correspond to the same value, which is a bit less than 75 (i.e., the mean of our prior). However, if the current age is 50, the posterior probability no longer resembles a Gaussian. Rather, $p(x|t=50)$ is a truncated Gaussian with a median that is no longer at the peak of the density. Regardless, the median of $p(x|t=50)$ is still quite close to 75, and this is still our best guess. However, if our subject is 70 years old, then the median (around 80) is now quite far from the peak of the posterior distribution. This captures the intuition that if someone has already lived to be 70, then they are more likely to live beyond the average life span than not. The median for

the posterior distribution as a function of current age *t* is plotted with the solid line in Fig. 5.6F, illustrating a form similar to the guesses that people had made (Fig. 5.6A).

The exact form of the function in Fig. 5.6F is strongly dependent on the shape of the prior $p(x)$. For example, suppose that instead of the Gaussian prior with a broad standard deviation that we assumed in Fig. 5.6C we had chosen a narrow Gaussian with the same mean but standard deviation $\sigma = 5$ years (rather than 15). Based on this prior, the best guess for lifespan of someone who is currently 80 years old is around 81, which is inconsistent with the guesses that people made.

This kind of data is of course not conclusive evidence that people are Bayesian estimators, because it reflects the 'wisdom of the masses' rather than individuals (as the data in Fig. 5.6A was averaged guesses from a group of people) (Mozer et al., 2008). Yet, it is consistent with the assumption that when people make guesses about every day questions, they are doing so by relying on a prior internal model (Fig. 5.6C) that resembles reality (Fig. 5.6B), and performing a computation that is consistent with a Bayesian process.

This conclusion appears to be challenged by a simple game that the reader can try with a group of friends. The game is known as the Monty-Hall problem. One of its variants is as follows. You are in a room with three closed doors, labeled d1, d2 and d3. Behind one of the doors is a one million dollars prize. You have to guess which door it is, and you do this in two steps. First, you form an initial hypothesis, for example, d1. There is an oracle in the room, who knows where the prize is. You tell the oracle your initial choice and ask her to reveal one of the losing doors among those that you did not select. In this case, it could either be d2 or d3. Suppose that the oracle declares that d3 does not have the coveted prize. At this point you must make your final choice. There are two possibilities: either you stick with the original choice (d1) or you switch to the only remaining alternative (d2). These are the choices, and the question is: "Which is the best strategy? Sticking with the initial choice, switching, or, it really cannot be decided because they are equally likely to win or lose?" Almost invariably the most popular answer is the last. And the argument appears to be compelling. You have two options, a winning option and a losing option. So, each must have a 50-50 chance to win. The correct answer: sticking with d1 has a 1/3 probability to win, while switching has a 2/3 probability to win. So, switching is the right thing to do. We leave the calculation of these probabilities, using Bayes' rule, as an exercise. Instead we ask: does the failure to answer correctly reveal that the human mind is non-Bayesian?

To see how this is not the case, you may try a simple variant of the Monty-Hall game. The original version has only $N = 3$ doors. Let us try again with a bigger number, say $N = 1000$. The problem is as follows. There are 1000 doors, labeled $d1, d2, \ldots, d1000$. Behind one of these lies a million dollar prize. Step 1: choose a door (for example d1). The game has 999 losing doors and now you ask the oracle to eliminate 998 of these. So, now you are left with two doors, d1 and – say – d376. What would you do? Facing with this variant, most people have no trouble recognizing that switching is the best strategy. This is simply because in this case, the prior is much stronger. In the classic Monty-Hall the prior probability of being correct was 1/3. Now it is 1/1000. When you had to make your initial choice, your expectation of being defeated is much stronger, as when you play the national lottery. It seems reasonable to assume that this feeling has a stronger and more evident persistence when you are faced with the final choice, compared to a situation where the difference between ignoring and considering the prior is $\frac{1}{2} - \frac{1}{3}$. So, in a way the outcome of the test, in its original and modified form provides more support to the mind's ability to deal with Bayes's rule. However, the "strength" of the prior plays a major role in the final decision making.

5.5 Behaviors that are not Bayesian: the rational and the irrational

While there are many behaviors in which it appears that the brain acts as a Bayesian state estimator, integrating prior beliefs with observations to estimate state of the body or the world, there are some curious behaviors that do not fit the theory. In these behaviors, the behaviors are perplexing, and seemingly illogical. It is interesting to ask why people behave in this way. Let us consider some of the more prominent examples.

Suppose that you were presented with two objects. They have the same shape and color, for example two yellow cubes, made of the same material. But the one is small and the other large. Despite their different sizes, the objects weigh the same (the larger cube is hollow and some of the core material was removed). You pick up one of the objects, place it back down, and then pick up the other object. It is very likely that you will feel that the small object weighed more than the large object. This is called the *size-weight illusion*, something that was first reported over a century ago (Charpentier, 1891). Your belief that the small object weighed more is in fact

the opposite of what a Bayesian state estimator should do. To see why, let us consider this problem in detail.

When we first see an object, the visual input y_v allows the brain to make an estimate of its weight w (Fig. 5.7A). We think this is true because the grip and load forces that people use to pick up objects are affected by what the object looks like (Fig. 5.3A): if the object looks large, we expect it to be heavier than an object that looks small. Generally, as volume of an object increases, so does its weight. The slope of this relationship depends on the materials in the object. For example, if the object is made of aluminum, the weight-volume relationship grows faster than if it is made of balsa wood (Fig. 5.7B). The visual property indicates the class of weight-volume relationships that we should expect, which is the basis with which we form prior beliefs about the weights of objects. For example, if we think the object is aluminum, then the relationship between the hidden state w and the volume y_v that we see is:

$$y_v = c_{al} w + \varepsilon_v \tag{5.47}$$

Whereas if it looks like the object is made of balsa wood,

$$y_v = c_b w + \varepsilon_v \qquad c_b > c_{al} \tag{5.49}$$

Before we see the object, we have some prior belief $\hat{w}^{(1|0)}$ (the distribution of weights of objects in general) and $\hat{c}^{(1|0)}$ (the distribution of weight-volume slopes in general). When we see the object (gather information about its volume y_v), we form a posterior estimate $\hat{c}^{(1|1)}$ and $\hat{w}^{(1|1)}$. That is, given its visual properties, we form a belief about what it is made of and how much it weighs. Propagating this forward in time, we have the prior estimate $\hat{w}^{(2|1)}$ that predicts the force that we should feel via our arm's Golgi tendon organs y_g as we pick up the object:

$$y_g = dw + \varepsilon_g \tag{5.50}$$

Now suppose that we see two objects, one small and one large. They both look like they are made of aluminum, but actually the larger one is made of balsa wood, and the two are exactly the same weight. When we pick up the large object, we expect it to be heavy, and there will be a prediction error, i.e. a difference between the expected force $\hat{y}_g$ and the observed force y_g, as $\hat{y}_g > y_g$. Because y_g is less than what we expected, the posterior $\hat{w}^{(2|2)}$ should be smaller than the prior $\hat{w}^{(2|1)}$. This is shown in Fig. (5.7C) by the gray (observation) and black (posterior) circles. Now suppose that we are given a small object that is made of aluminum and weighs the

same as the large object that we just picked up. When we pick up the small object, we will have little or no prediction errors, resulting in a weight estimate that is close to our prior (shown by the black circle in Fig. 5.7C). If we now estimate which object weighs more, we should clearly guess that the larger object weighs more. But this is not what happens. In fact, people consistently report that the smaller object weighs more. This judgment is irrational from a Bayesian estimation perspective, but is in fact the way our brain works. What is going on?

Look at the data in Fig. 5.3B. Clearly, people are gripping the larger object with greater force and are applying a larger load force to pick it up. On the first try, the motor system certainly seems to believe that the larger object weighs more. If after this first try the brain really believes that the smaller object weighs more (as this is what subjects tell you after they pick up the objects), then on the second try the grip forces should be higher for the smaller object. But this is not what happens! Figure 5.7D shows the force data from 20 consecutive attempts to pick up small and large objects that weigh the same. Clearly, the smaller object is never experiencing the larger load force. In fact, by the 8^{th} or 10^{th} trial the forces are about the same, which makes sense as the two objects really do weigh the same. The remarkable fact is that the illusion that the smaller object weighs more (as verbalized by subjects) persists even after the motor system continues to demonstrate that it believes that the two objects weigh the same, an observation that was nicely quantified by Randy Flanagan and Michael Beltzner (Flanagan and Beltzner, 2000).

In summary, these results show that our motor system (as assayed by the forces that we produce with our hands) never believes that the small object weighs more than the larger one. However, apparently our motor system is not responsible for what we verbalize, because we consistently say that "the smaller object feels heavier." It appears that the brain does not have a single estimate of an object's weight. There appears to be two such estimates: one that is used by our 'declarative' system to state (in words) how much it thinks the object weighs, and one that is used by our 'motor' system to state (in actions) how much it thinks the object weighs. In these experiments, the rational part of the brain is the one that programs the motor commands, while the seemingly irrational one is the one that verbalizes what it thinks the objects weigh.

More recent experiments have shed some light on the mechanism that our brain uses to verbalize opinions about objects and their weights. Apparently, this declarative system also relies on a volume-weight prior belief, but this belief changes much more slowly than the one used by the motor system (it still remains unclear how the prior is integrated with observations). To explore

this issue, Randy Flanagan, Jennifer Bittner, and Roland Johansson (2008) trained people on a set of objects that had an unusual property: the larger the volume, the smaller the weight (Fig. 5.8A). People lifted these objects hundreds of times a day for up to 12 days. At the end of the first day, the experimenters gave the subjects a small and a large object and asked them to indicate their relative weight. The two objects weighed the same, but the subjects reported that the smaller object weighed more (Fig. 5.8B). This is the usual size-weight illusion that we have seen before. However, after a few more days of lifting objects, the illusion subsided and by the 11^{th} day it had reversed direction so that they now perceived the larger object to weigh slightly more.

During this period of training, the prior that the motor system used for estimating weight from volume changed. On the first trial of the first day, people expected the small object to be light, and so they produced a small load force (early trial, Fig. 5.8C). By the 8^{th} trial, the load force had substantially increased (late trial, Fig. 5.8C). On the second day, the motor system remembered this unusual, inverted relationship between volume and weight: from the very first trial, the motor system produced a larger force for the small object than the larger object (Fig. 5.8D). Therefore, a few trials of training were sufficient to teach the motor system that this class of objects had an unusual property that increased volume produced reduced weight. The declarative system too relied on a prior model, one in which weight increased with volume. However, this model appeared to change much more slowly, as it took 11 days before the illusion reversed. Therefore, if the perceptual system is not acting irrationally, it follows that the experiential structure that form the basis for its internal model is different from the experiential structure associated with motor learning.

5.6 Multiple prior beliefs

It is curious indeed that one should believe that a small object feels heavier than a large object, and yet consistently act as if believing that the weight of the small object is the same as the large object. Perhaps in our brain there are distinct and sometimes conflicting beliefs about the properties of single objects. Perhaps depending on how our brain is queried, we express one belief or the other. An elegant experiment by Tzvi Ganel, Michal Tanzer, and Melvyn Goodale (2008) lends support to this counter-intuitive conjecture.

In the experiment (Fig. 5.9A), two lines of unequal size were presented on a screen. The background was manipulated to give cues suggesting object 1 to be closer than object 2. As a

result, most people would estimate object 2 to be taller than object 1. In fact, object 2 was about 5% shorter than object 1, as shown in the figure without the illusory background. On each trial, people were instructed to pick up the taller or the shorter object. When they were instructed to pick up the taller object, in about 90% of the trials they picked the shorter object, and similarly in about 90% of the trials people picked the taller object when they were instructed to pick the shorter one. That is, the background clearly produced a strong illusion.

As the subjects reached toward their selected object, the authors recorded the distance between the fingers. By choosing which object to pick, the subjects expressed their belief about which object was taller. By moving their fingers apart during the reach, they expressed their belief about the height of the object. Interestingly, they found that the distance between the fingers was not affected by the illusion: the aperture was small when picking up the short object, despite the fact that subjects were picking up that object because they thought it was the taller of the two objects (Fig. 5.9B). Similarly, the aperture was large when picking up the tall object, despite believing that it is the shorter object. Control experiments in which the visual feedback from the object and hand were removed confirmed this result. The motor commands that controlled the fingers in the task of picking up the object were not fooled by the visual cues that caused the illusion.

The same people were then asked to use their fingers to show their estimate of the size of the objects. With the illusory background in place, people were asked to estimate size of the shorter object. To convey their decision regarding which object they believed to be shorter, they moved their hand 5 cm to the right of the object that they chose and then split their fingers apart to show their estimate of its size. As before, people chose the tall object when they were instructed to estimate size of the shorter object. However, now they had their fingers apart by a smaller amount than when they were asked to estimate the size of the taller object. That is, in all cases, their perception of which object was smaller was affected by the illusory background. However, when they were asked to pick up the object, they moved their fingers apart in a way that suggested they were not fooled by the visual cues. In contrast, when they were asked to move their fingers apart so to estimate the size of the object, they were fooled. Finally, when the illusory background was removed and the two objects were displayed on a normal background (middle plot, Fig. 5.9A), the grip sizes in the grasp trials and in estimation trials accurately reflected the relative object sizes.

One way to make sense of this data is to imagine that when we respond to "pick up the object", our actions are based on beliefs that are formed in parts of our brain that are distinct from beliefs that are used to respond to "show me the size of the object." Perhaps these beliefs are distinct because the various brain regions focus on distinct parts of the available sensory information. Melvyn Goodale and David Milner (Goodale and Milner, 1992) have proposed that the pathway that carries visual information from the visual areas in the occipital lobe to the parietal lobe (dorsal pathway) and the pathway that carries information from the occipital lobe to the temporal lobe (ventral pathway) build fundamentally distinct estimates of the object properties in the visual scene. In a sense, the actions that we perform based on the belief of the ventral pathway can be different than actions that we perform based on belief of the dorsal pathway. They have argued that when we pick up the object, we are relying on internal models in the dorsal pathway. This pathway is less affected by the background. When we use our hand to show an estimate of the size of the object, we are relying on internal models in the ventral pathway. This pathway is more affected by the background.

Summary

In Bayesian estimation, the objective is to transform a prior belief about a hidden state by taking into account an observation, forming a posterior belief. The Kalman gain is a weighting of the difference between the predictions and observations, which when added to a prior, transforms it to the expected value of the posterior. Here, we linked Bayesian estimation with the Kalman gain by showing that the posterior belief $\hat{\mathbf{x}}^{(n|n)}$ is the expected value of $p\left(\mathbf{x}\middle|\mathbf{y}^{(n)}\right)$, and the uncertainty of our posterior belief $P^{(n|n)}$ is the variance of $p\left(\mathbf{x}\middle|\mathbf{y}^{(n)}\right)$.

When our various sensory organs produce information that are temporally and spatially in agreement, we tend to believe that there was a single source that was responsible for our observations. In this case, we combine the readings from the sensors to estimate the state of the source. On the other hand, if our sensory measurements are temporally or spatially inconsistent, then we view the events as having disparate sources (Fig. 5.2A), and we do not combine the sources. The probability of a common source depends on the temporal and spatial alignment of our various sensory measurements, and this probability describes how we will combine our various observations.

Prior beliefs do play a very strong role in how people interact with objects in everyday scenarios. We expect larger things to weigh more than smaller things. We expect objects to fall at an acceleration of 1g. When objects behave differently than we expected, we combine our observations with our prior beliefs in a manner that resembles Bayesian integration. Our guesses about everyday things like how long someone is likely to live is also consistent with a Bayesian process that depends on a prior belief.

The motor system appears to be rational in the sense that it estimates properties of objects by combining prior beliefs with measurements to form posterior beliefs in a Bayesian way. However, as the size-weight illusion demonstrates, our verbal estimate of an object's relative weight (i.e., whether it is heavier or lighter than another object) is not the same as the motor system's estimate. The verbal estimate appears to rely on a separate internal model, one that changes much more slowly than the internal model that the motor system relies upon. It is possible that our brain has multiple internal models that describe properties of a single object, and depending on how we are asked to interact with that object, we may rely on one or the other model. The distinct pathways that carry visual information in the parietal and temporal lobes may be a factor in these distinct internal models. This may explain the fact that visual cues that produce perceptual illusions about an object's properties often do not affect the motor system's abilities to interact with that object.

Figure Legends

Figure 5.1. People combine visual and auditory information if they believe that the two sensors were driven by a common spatial source. **A**) Volunteers were placed in the center of a semi-circle and heard an auditory stimulus followed by a light from one of the LEDs. The onset of the two cues was separated in time by 200-800ms. They then pointed to the location of the sound and pressed a switch if they thought that the light and sound came from the same location. **B**) Probability of perceiving a common source as a function of the temporal and spatial disparity between the sound and visual cues. **C**) As the probability of a common source increased, the perceived location of sound $\hat{x}_s$ was more strongly biased by the location of light x_v. (From (Wallace et al., 2004) with permission.)

Figure 5.2. Estimating the state when there are two potential generative models. A) The visual and sound sensor may be driven by a common source, or by two different sources. B) The probability of a common source, given the sensory measurements. This plot is Eq. (5.23). When the two measurements y_s and y_v are close to each other, the probability of a common source is nearly one. When they are far from each other, the probability is close to zero. C) As the spatial disparity between the two measurements increases, it is less likely that one is observing the consequences of a common source. D) The estimate of the location of the sound $\hat{x}_1$ when the sound is heard at position zero $y_s = 0$ but the light is observed at various displacements y_v. When y_v and y_s are near each other, estimated location of the sound is affected by the observed location of the light.

Figure 5.3. The effect of prior beliefs during interactions with everyday objects. **A**) Volunteers were asked to use their fingers to lift up a small, medium, or a large box. The instrumented device measured grip and load forces. The three boxes were the same weight. **B**) People tended to produce the smaller grip and load forces for the smallest box, resulting in large lift velocities for the largest box. (From (Gordon et al., 1991) with permission). **C**) The ball starts with a non-zero velocity from a given height and falls in 0g or 1g gravity. In the 0g scenario (i.e., in space), the recently arrived astronaut will use a 1g internal model to predict the ball's trajectory, expecting it to arrive earlier (dashed line). **D**) EMG activity from arm muscles of an astronaut in 0g and 1g. In 0g, the arm muscles activate sooner, suggesting that the astronaut expected the ball to arrive sooner than in reality. (From McIntyre et al. (2001) with permission.)

Figure 5.4. The prior statistics of everyday movements affect how people move a computer mouse to a given stimulus. **A**) The distribution of movement amplitudes as measured over a multi-day period. **B**) The distribution of movement directions, represented as an angle θ_e. This is the angle of a line connecting the start to the endpoint. **C**) While some of the movements were straight, many of the movements had an initial angle θ_i that was somewhat different than θ_e. The prior probability distribution of θ_i is shown. **D**) The expected value of Eq. (5.37), $E\left[\theta_i - \theta_e \middle| \theta_e\right]$ is plotted as the dashed line. The measured value $\theta_i - \theta_e$ is plotted as a solid line. (From Slijper et al. (2009) with permission.)

Figure 5.5. Modeling the data in Fig. 5.4. **A**) A normal distribution for a random variable defined on a circle. The mean of the distribution in Eq. (5.38) is at μ and the variance increases with decreasing κ. For the example in Fig. 5.4, we assume that $p\left(\theta_e \middle| \theta_i\right)$ is a normal with mean at θ_i and variance of a few degrees. **B**) The data in Fig. 5.4C, i.e., the prior probability $p\left(\theta_i\right)$, approximated as sum of four normal distributions and a uniform distribution, as in Eq. 5.39. C) $p\left(\theta_i \middle| \theta_e\right)$. When the target is at 0.1 radians (about 6°), as indicated by the dashed line, the initial movement direction θ_i is likely to be toward 3°, which is the expected value of the posterior.

Figure 5.6. The problem of guessing the lifespan x of a person (i.e., how many years someone will live), given that they are now t years old. A) Guesses from a group of students about lifespan, given current age. B) The distribution of lifespan for a person born in America, i.e., $p\left(x\right)$. C) A simplified model of lifespan distribution, Eq. (5.40). D) Age distribution of people alive today, as in Eq. (5.42). E) The posterior probability $p\left(x \middle| t = 30\right)$, $p\left(x \middle| t = 50\right)$, etc. The median for each distribution is marked by an arrow. F) The median for the posterior distribution $p\left(x \middle| t\right)$ as a function of current age *t* is plotted with the solid line. This indicates the best guess regarding lifespan for a prior probability $p\left(x\right)$ as shown in part C. If the prior probability $p\left(x\right)$ is narrower than that shown in part C but has the same mean, the posterior (dashed line) is altered.

Figure 5.7. The size-weight illusion. **A**) A generative model to estimate the weight of an object. y_v refers to observations from the visual sensors and y_g refers to observations from the golgi-

tendon (force) sensors. **B**) The prior belief regarding the relationship between weight, volume, and material that the object is made of. The term c describes the slope of the weight-volume relationship. When we see two objects that look like are made of aluminum, we expect the larger one to weigh more. C) Suppose that the larger object is actually made of wood, and weighs the same as the smaller object. When we pick up the larger object, the measured weight is smaller than we expected. A Bayesian estimator would believe that the larger object weighs somewhere in between what it predicted (heavy weight), and what is observed (light weight). In all cases, this posterior estimate of weight should be larger than for the smaller object. D) While people 'feel' that the small object weighs more than the large object, the motor system in fact uses a greater amount of load force rate and grip force rate to pick up the larger object. After about 8-10 trials, the force rates for the larger object converge to the small object, consistent with the fact that the two objects weigh the same. Despite this, people still 'feel' that the smaller object weighs more, and this feeling lasts for up to hundreds of trials. (From (Flanagan and Beltzner, 2000), with permission.)

Figure 5.8. People were trained to pick up a set of objects that had an unusual property: the larger the volume, the smaller the weight. **A**) Experimental set up. People lifted and placed these objects hundreds of times a day for 12 or more days. **B**) At the end of the first day, the experimenters gave the subjects a small and a large object and asked them to indicate their relative weight. The two objects weighed the same, but the subjects reported that the smaller object weighed more. By the 11^{th} day the illusion had reversed direction so that they now perceived the larger object to weigh slightly more. **C**) On the first trial of the first day, people expected the small object to be light, and so they produced a small load force (early trial). By the 8^{th} trial, the load force had substantially increased. The black dashed vertical lines mark the time of initial peak in load-force rate, and the horizontal dashed lines mark the load force at the time of initial peak in load-force rate. The gray vertical lines mark the time of liftoff. **D**) Load force at the time of the initial peak in load-force rate for the small (filled circle, heavy object) and the medium (open circle, light object) objects. Each point represents the average across participants, for 5 consecutive trials. The motor system learned that the smaller object weighed more, and remembered this from day to day. (From (Flanagan et al., 2008) with permission.)

Figure 5.9. Visual illusions affect perception but not action. **A**) The background was manipulated so that line 2 appears to be longer than line 1. In fact, line 1 is about 5% longer than line 2. People were instructed to reach and pick up the shorter or the longer line. **B**) In about

90% of the trials, people reached for the shorter line when instructed to pick up the longer line, and vice versa. However, grasp size during the act of picking up the object was not fooled by the visual illusion (bars on the left). In contrast, when subjects were asked to estimate the size of the objects, they were fooled by the visual illusion (bars on the right). **C**) Control experiments without an illusory background. (From (Ganel et al., 2008) with permission.)

Reference List

Charpentier A (1891) Analyse experimentale quelques elements de la sensation de poids [Experimental study of some aspects of weight perception]. Arch Physiol Normales Pthologiques 3:122-135.

Flanagan JR, Beltzner MA (2000) Independence of perceptual and sensorimotor predictions in the size-weight illusion. Nat Neurosci 3:737-741.

Flanagan JR, Bittner JP, Johansson RS (2008) Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. Curr Biol 18:1742-1747.

Ganel T, Tanzer M, Goodale MA (2008) A double dissociation between action and perception in the context of visual illusions: opposite effects of real and illusory size. Psychol Sci 19:221-225.

Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends Neurosci 15:20-25.

Gordon AM, Forssberg H, Johansson RS, Westling G (1991) Visual size cues in the programming of manipulative forces during precision grip. Exp Brain Res 83:477-482.

Griffiths TL, Tenenbaum JB (2006) Optimal predictions in everyday cognition. Psychol Sci 17:767-773.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS ONE 2:e943.

McIntyre J, Zago M, Berthoz A, Lacquaniti F (2001) Does the brain model Newton's laws? Nat Neurosci 4:693-694.

Mozer MC, Pashler H, Homaei H (2008) Optimal predictions in everyday cognition: the wisdom of individuals or crowds? Cognitive Science 32:1133-1147.

Slijper H, Richter J, Over E, Smeets J, Frens M (2009) Statistics predict kinematics of hand movements during everyday activity. J Mot Behav 41:3-9.

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004) Unifying multisensory signals across time and space. Exp Brain Res 158:252-258.