4. Forming beliefs: state estimation theory

In a cocktail party, one of the toughest jobs (at least from a motor control standpoint) belongs to the fellow who brings the tray with the drinks. As he holds the tray and you pick up the glass, he needs to compensate for the reduced mass on the tray and not spill the remaining drinks. To convince yourself that that is indeed a tough task, try the following experiment. Put out your hand with the palm side up and place a moderately heavy book on it. Now have a friend stand in front of you and have him or her pick the book up off your hand. You will note that despite the fact that you can see the friend reaching and picking up the book, you cannot hold your hand perfectly steady—it invariably moves up when the book is removed. Now replace the book on your palm and go ahead and pick it up yourself. When you pick up the book, the hand that used to hold the book remains perfectly still.

This simple experiment suggests that when you send commands to your arm to pick up the book, your brain predicts the exact moment that the mass will be removed off the resting arm and the exact mass of the book, and then reduces the activity in the muscles that are producing force to hold up the book. In contrast, when someone else picks up the book, you have to rely on your sensory system (rather than your predictions). Because of the inherent delay in the sensory system, you are invariability late in reducing the activity of those same muscles. Therefore, the brain appears to predict that lifting the book by one arm has consequences on the other arm, and compensates before arrival of sensory feedback. (This is the reason why you should let the waiter pick up the glass and hand it to you.)

The general idea is that our body is a multi-segmented structure in which motion of any one segment has consequences on the stability of other segments. To maintain stability, the brain needs to be able to predict how motor commands to one segment affect the states of all other segments. For example, consider an experiment by Paul Cordo and Lewis Nashner (1982) in which people were placed in a standing posture and were asked to pull on a door knob that was attached to a rigid wall (Fig. 4.1). Pulling on the knob involves activation of the biceps muscle. However, doing so would not only result on a pulling force on the knob, but also a pulling force on your body. If you're not careful, you will end up hitting your forehead on the wall. To maintain stability, you need to activate muscles that resist bending (i.e., flexion) of the ankle and knee joints. Indeed, Cordo and Nashner (1982) found that people activated their hamstring (a muscle that produces an extension torque on the ankle) and gastrocnemius (produces extension

torque on the knee) muscles just before they activated their biceps muscle (Fig. 4.1A). That is, the brain stabilizes the leg just as it pulls on the knob.

This may seem like a 'hard-wired' reflex, but it is not. Consider what happens when a rigid bar is placed in front of the chest, preventing the body from swaying forward. In this case, a pull on the door knob would not cause a sway of the body no matter what you do with your knee and ankle muscles. Indeed, with the rigid bar in place, people no longer activated the hamstring and gastrocnemius muscles as they pulled on the knob (Fig. 4.1B).

These data suggest that as our brain plans and generates motor commands, it also predicts the sensory consequences and acts on the predicted consequences. When standing upright without a rigid bar to lean against, pulling on the knob will flex the knee and ankles. If we can predict that this is the sensory consequence of pulling on the knob, we can act on it by activating muscles that resist this flexion (the extensors). When we have the rigid bar to lean against, the same pull on the knob will no longer flex the knee and ankles. If we can predict this, we need to do nothing, which seems consistent with the data in Fig. 4.1.

More direct evidence for the idea that the brain predicts the sensory consequences of motor commands comes from the work of Rene Duhamel, Carol Colby, and Michael Goldberg (1992). They trained monkeys to fixate a light spot and make a saccade to it whenever it jumped from one location to another. They recorded from cells in the posterior parietal cortex (PPC), an area in the brain in which cells have receptive fields that depend on the location of the stimulus with respect to the fixation point. Fig. 4.2A illustrates response of a cell that had its receptive field to the upper right of fixation. When the stimulus was turned on in the receptive field of this cell (circular dashed line), the cell responded after about 60ms, which is a typical delay period. Next, they had the monkey make a saccade by turning off the fixation point at right and re-displaying it at left (Fig. 4.2B). They placed the stimulus in the upper right part with respect to the 2nd fixation point. In this way, when the animal was looking at the 1st fixation point (the one on the right), there was nothing in the cell's receptive field. When the animal looked at the 2nd fixation point (the one on the left), the stimulus fell in the cell's receptive field. However, if the brain predicts the sensory consequences of the motor commands, then one of the consequences of the command to move the eyes from one fixation to another is that after the completion of the movement, the stimulus will fall in the cell's receptive field. And so, the cell might fire in anticipation of this event, rather than in passive response to the sensory input. Indeed, the cell's response was around

the start of the saccade, rather than after its completion (Fig. 4.2B, note that saccades take about 50ms to complete). Therefore, some part of the brain predicted that as a consequence of the motor commands that move the eyes, a light stimulus would fall in the receptive field of this cell, and it began to fire in apparent anticipation.

4.1 Why predict sensory consequences of motor commands?

Why should the brain predict the sensory consequences of motor commands? In the experiment where you lifted the book off your hand, the clear advantage of making sensory predictions is that the brain does not have to wait for the sensory measurements to know that it needs to shut-off the muscles that are holding up the book. The delay in sensory measurements is long enough that it can cause stability problems. Relying on predictions, rather than delayed measurements, allows one to overcome this delay. However, in the case of the saccade shown in Fig. 4.2, it may be unclear why the brain should predict that the sensory consequences: why should a visually responsive cell be activated in anticipation of the sensory stimulus, as well as in response to that sensory stimulus? After all, the visual stimulus will appear in the receptive field shortly after the saccade. Why predict this event?

One possibility is that our perception, i.e., our ability to estimate the state of our body and the external world, is always a combination of two streams of information: one in which our brain predicts what we should sense, and one in which our sensory system reports what was sensed. The advantage of this is that if our expectations or predictions are unbiased (i.e., their mean is not different from the 'true' state), then our perception (and the decisions that are made based on that perception) will be better than if we had to rely on sensory measurements alone. In a sense, our perception will be more accurate (e.g., less variable), if we combine what we predicted with what our sensory system measured. This improved accuracy in perception is a fundamental advantage of making predictions about the sensory consequences of motor commands.

Although this may seem like a fairly new idea, it was first proposed in the 11th century by Ibn al-Haytham, an Arab scientist (known in Europe as Alhazen), in his *Book of Optics*. He was considering the moon illusion, the common belief that the moon looks larger when it is near the horizon. Aristotle and Ptolemy had thought that this was due to a magnification caused by Earth's atmosphere (refraction theory). However, this is not the case. If you were to measure the size of the moon by taking a picture, you would measure a width that is 1.5% smaller at the

horizon than straight up in the sky. That is, the image that falls on your retina is actually smaller when you are looking at the moon on the horizon (this is because the moon is actually farther away from you by half of the Earth's diameter when it is at the horizon). So despite the fact that the moon at the horizon produces a visual image that is smaller than when it is overhead, we perceive it to be bigger at the horizon. Ibn al-Haytham argued that the moon looked bigger at the horizon because perception was occurring in the brain, and not in the eyes. He recognized that our perception of size depends not only on the size of the image on our retina, but also on our estimate of the object's distance. At the horizon, the brain has cues like trees and buildings to judge distance of objects, whereas in the sky above, these cues are missing. He argued that presence of these cues affected the brain's estimate of distance of the moon, making us perceive it to be much farther away at the horizon than overhead. Perhaps we 'see' the moon to be larger at the horizon because we believe it is much farther away at the horizon than overhead.

(The term 'belief' is used loosely here. As we will see in the next chapter, different parts of our brain may have differing beliefs about size of a single object. For example, we may verbally indicate that object A is bigger than object B, reflecting the belief of our perceptual system, but when we go to pick up object B, move our fingers apart more for it than when we go to pick up object A, reflecting the belief of our motor system. The root cause of these apparently different beliefs about the property of a single object is poorly understood. It is possible that it has something to do with the fact that each part of the brain may specialize in processing a different part of the sensory data. For example, by focusing on the background of an image, some part of our visual system may get fooled and form an illusion, whereas by focusing on the foreground of an image, another part of our visual system may not form this illusion. So there may not be a single 'belief' in our brain about some issue, but multiple beliefs. The belief that you express may depend on what part of your brain was queried.)

In 1781 Immanuel Kant in his theory of *idealism* claimed that our perceptions are not the result of a physiological process in which, for example, the eyes faithfully transmit visual information to the brain, but rather, our perceptions are a result of a psychological process in which our brain combines what it already thinks, believes, knows, wants to see, with the sensory information to form a perception (Gilbert, 2006). He wrote "The understanding can intuit nothing, the senses can think nothing. Only through their union can knowledge arise." (I. Kant, Critique of Pure Reason, trans. N. K. Smith, 1781; New York: St. Martin's Press, 1965, p. 93.)

If we follow this line of reasoning and return to our example of predicting the sensory consequences of the motor commands that move the eyes in a saccade (Fig. 4.2B), we might guess that during the post-saccadic period, the brain combines what it predicted with what it currently senses. The combination of the two streams of information would allow it to sense the world better than if it only had the sensory information stream. That is, it would be able to 'see' the stimulus better because it had two sources of information about it, rather than just one.

Siavash Vaziri, Jörn Diedrichsen, and Shadmehr (2006) tested this idea. In their control condition (Fig. 4.3A), subjects fixated a visual target and reached to it. In their static condition, the target appeared in the peripheral vision. The ability to estimate the location of objects that appear in peripheral vision is poor, and so the standard deviation of the reach endpoints was higher for these stimuli in the peripheral vision (Fig. 4.3B). Therefore, in this experiment reach variability was a proxy for the goodness with which the brain could estimate the location of the visual target. When the target was in the periphery, subjects could not localize it as well as when they looked straight at it. Now, what if one could predict the location of the target? Based on our theory, the brain should be able to localize it better, and this should result in less variable reaches to that target.

In the remap condition, subjects looked at the target, and then looked away before they reached to the target. When they looked away, the target had disappeared. Presumably, during the saccade the brain predicts the new location of the target with respect to fixation (this is called remapping). So after the saccade completes, it can rely on the predicted target position to reach. The endpoint variance in the remapped condition was much less than in the static condition (Fig. 4.3B). The crucial test was in the combined condition in which subjects looked away from the target, and in the post-saccadic period the target reappeared for a brief period of time. If the brain combined what it predicted about target position with the actual sensory feedback, then reach endpoint variance in the combined condition should be better than both the remap and the static conditions (Fig. 4.3B). Indeed, this was the case, consistent with the idea that the brain predicted the sensory consequences of the saccade (remapping of the target) and then combined this prediction with the post-saccadic visual information to estimate the actual position of the target.

Therefore, by predicting the sensory consequences of motor commands, the brain not only can over-come delay in sensory feedback, but perhaps more importantly, it can actually sense the world better than is possible form sensory feedback alone. The latter comes about when our brain

combines what it has predicted with what it has measured. A diagram that summarizes the idea of predicting the sensory consequences of motor commands is provided in Fig. 4.4. An internal model that predicts the sensory consequences of motor commands is known as a *forward model*.

4.2 Disorders in predicting the sensory consequences of motor commands

The central idea is that our perception is based on a combination of two streams, one that arises from the motor system (predicting the sensory consequences), and the other that arises from the sensory system (measuring the sensory consequences). If our brain could not accurately predict sensory consequences of our motor commands, then we would not be able to sense the world around us in a normal way. An example of this is patient RW, a 35 year old man who was described by Thomas Haarmeier, Peter Thier, Marc Repnow, and Dirk Petersen (Haarmeier et al., 1997). RW suffered a stroke in a region covering parts of the parietal and occipital cortex, centered on an area that contains the vestibular cortex, a location in which cells are sensitive to visual motion. RW complained of vertigo only when his eyes tracked visual objects, but not when his eyes were closed. He explained that when he was watching his son run across a field (a condition in which his eyes smoothly moved to follow his son), he would see the boy running, but he would also perceive the rest of the visual scene (e.g., the trees) smoothly moving in the opposite direction.

Haarmeier et al. conjectured that when RW moved his eyes, his brain was unable to predict the sensory consequences of the oculomotor commands. As his eyes moved to follow his son, the trees moved in the opposite direction on his retina. The healthy brain predicts that moving the eyes will have the sensory consequence of shifting the image of the stationary world on the retina. We do not perceive this shifting image as real motion of the world because we predict it to be a consequence of motion of our eyes. In RW, perhaps his vertigo was a symptom of his brain's inability to predict such sensory consequences.

To test this conjecture, Haarmeier et al. (Haarmeier et al., 1997) had RW sit in front of a computer monitor and keep his eyes on a moving cursor (Fig. 4.5). As the cursor moved smoothly from left to right, random dots were displayed for 300ms. In some trials the random dots would stay still, and in other trials the dots would move to the right or left with a constant speed. On each trial they asked RW and some healthy volunteers to guess whether the random dots were moving to the left or right. From the response that they recorded the authors estimated

the speed of motion of the random dots for which subjects sensed it to be stationary. For healthy volunteers, the speed of subjective stationarity of the random dots was near zero, no matter what the speed of the moving cursor that they were looking at. That is, regardless of whether the eyes moved quickly or slowly, healthy people perceived a stationary collection of dots as stationary (the unfilled circles in Fig. 4.5C). However, RW saw the collection of dots as being stationary only when the dots moved at the same speed as the eyes (the filled symbols in Fig. 4.5B). That is, for RW an object was stationary only if during the movement of the eye, its image remained stationary on the retina.

You do not need to have a brain lesion to get a feel for what RW sees when he moves his eyes. Take a camera and aim it at a runner and try to move (i.e., 'pan') so that the image of the runner stays at the center of the picture. As you are moving the camera, take a picture. That picture will show a sharply focused runner but a blurry background that appears to be moving in the opposite direction. However, when you are following the runner with your naked eyes, the background appears perfectly still. The reason is because your brain predicts the background image shift that should take place on the retina as you move your eyes, and combines it with the actual shift. By combining the observed and predicted images, the parts that agree must have been stationary, and parts that disagree must have moved.

In 1996, Chris Frith put forth the hypothesis that in schizophrenia, the symptoms of delusions, e.g., auditory hallucinations in which the patient hears voices, or has beliefs about someone else guiding their actions, are potentially related to a disorder in the brain's ability to interpret its own inner voice or covert action (Frith, 1996). Chris Frith, Sarah Blakemore, and Daniel Wolpert (Frith et al., 2000) formalized this idea in terms of a problem in which the brain had an incorrect forward model. This would result in the patients mistaking the sensory consequences of their own actions as consequences of actions of others, i.e., a misattribution of the cause of sensory events.

An interesting test of this idea was performed by Axel Lindner, Peter Thier, Tilo Kircher, Thomas Haarmeier, and Dirk Leube (Lindner et al., 2005). They essentially repeated the experiment that they had earlier performed on patient RW on a group of schizophrenic patients. In this experiment, subjects watched a red dot that moved from left to right at a constant velocity (Fig. 4.6). They then presented a moving background for 200ms during the sweep of the target and asked the subject to report the direction of the background motion. By varying the background

speed, they determined the velocity that produced about equal perception of rightward or leftward motion. At this velocity of perceptual stationarity, they compared the speed of the background image (which ideally should be zero) with the speed of the eye movements (plotted in bar graph in Fig. 4.6). In this figure, 100% implies that the subjects saw the background as stationary when in fact it was stationary. Healthy subjects had a less than 100% performance, and this was attributed to the very limited time for which the background was presented. However, for the schizophrenic patients who suffered from delusional symptoms, the performance was significantly worse, i.e., they had a harder time compensating for self-induced image motion. When these patients moved their eyes, they attributed more of the observed motion on their retina to movement of the external world than healthy controls.

Of course, schizophrenia is a complex disease for which accurately estimating the sensory consequences of self-generated actions may be a minor component (if this is not the case, then why is patient RW not schizophrenic?). Nevertheless, it is curious that for those patients who suffer from delusional symptoms, there is a tendency to have a motor disorder in perceiving self-generated sensory consequences.

4.3 Combining predictions with observations

To combine two streams of information, one needs to apply a weighting to each stream. In principle, the weight should be higher for the more reliable information source. In the experiment shown in Fig. 4.3, the two sources of information are the remapped target (i.e., predicted target location), and the post-saccadic target (i.e., observed target location). Vaziri et al. (2006) manipulated the reliability of the post-saccadic information by presenting the target for either a short or long period of time. The idea was that the longer the information was available, its reliability would increase, and so the weight that the brain might assign to it should increase. Indeed, with increased post-saccadic target exposure in the combined condition, endpoint variance decreased, suggesting that the brain increased the weighting assigned to the observed sensory information source.

The basic idea that emerges is that our estimate of the state of the world is a combination of two sources of information: what we predicted, and what we observed. Konrad Körding and Daniel Wolpert (Kording and Wolpert, 2004) varied the reliability of these two hypothetical sources of information and tested the idea that perception was a weighted combination of the two sources.

They first trained subjects to reach to a goal location by providing them feedback via a cursor on a screen (the hand was never visible). As the finger moved from the start position, the cursor disappeared. Halfway to the target, the cursor re-appeared briefly (Fig. 4.7A). However, its position was, on average, 1 cm to the right of the actual finger position, but on any given trial the actual displacement was chosen from a Gaussian distribution. The objective was to produce a movement that placed the cursor inside the target.

If you were a subject for this task, you might start by moving the finger straight to the target. You would note that in the middle of the movement the cursor appears about 1 cm to the right, so you would correct by moving the finger a little to the left. After some practice, you'd learn that when you produce motor commands that move the hand slightly to the left, you should, on average, see the cursor at about straight ahead. Because the location of the cursor is probabilistic, the confidence that you have about this predicted sensory consequence of your motor command should be described by the variance of the Gaussian distribution that describes the displacement of the cursor (Fig. 4.7B, top row). In this way, the experiment controlled the confidence that the brain should have in predicting the sensory consequences of its motor commands.

To control the confidence that the brain should have regarding sensory measurements, they added noise to the display of the cursor: the cursor was displayed as a cloud of dots. This induced uncertainty. On some trials the cursor was shown clearly so the uncertainty regarding its position was low. In other trials the uncertainty was high as the cursor was hidden in a cloud of noise. The idea was that on a given trial, when the subject observes the cursor position midway to the target, they will correct based on two sources of information: what they observed on this trial, and what they predicted regarding where the cursor should be. For example, if on a given trial they see the cursor at 2 cm (Fig. 4.7B, middle row), they should combine this observation with their prediction (Fig. 4.7B, top row), to form a belief about cursor position that is somewhere between 2 cm (observed) and 1 cm (predicted). This 'belief' would depend on how much weight (or uncertainty) they assign to the observed and predicted sources of information. If the observed cursor is in a noisy cloud, they should rely more on their prediction (σ_∞ line, Fig. 4.7A). If the observed cursor is clear, they should rely more on their observation (σ_0 line, Fig. 4.7A). The weighting of the two sources should be inversely related to the variance of the distributions. Indeed, Körding and Wolpert's (2004) experimental data was consistent with this theoretical framework.

In summary, the data suggests that as the brain programs motor commands, it also predicts the sensory consequences. Once the sensory system reports its measurements, the brain combines what it had predicted with the measurements to form a 'belief' that represents its estimate of the state of the world. Our actions are not simply based on our current sensory observations. Rather, our actions are based on an integration of the sensory observations with our predictions. In engineering, this is called estimation theory.

4.4 State estimation: the problem of hiking in the woods

The problem of estimating the state of our body (or state of something else) has two components. The first is associated with learning to accurately predict what the sensory measurements of that state should be – this is our *prior* estimate of state. The second is associated with combining the measured quantities with the predicted one to form a *posterior* estimate of state. The first problem is one of model building, i.e., describing an internal model that predicts what our sensors should be telling, called a forward model. The second problem is one of integration, i.e., describing how to form an estimate of state based on the two sources of information, the prediction and the observation. Our goal here is to build a mathematical framework in which we can describe this problem. Once the framework is in place, we will use it to try and account for some behaviors in people and other animals.

To illuminate the basic problem that the brain faces, let us consider the following example. Say that you are hiking in the woods and are concerned about getting lost. To help with navigation, you have bought two Global Positioning System (GPS) devices. One of the devices is European made and uses satellites operated by the European Union. The other is made to work with the US satellites. Therefore, you can assume that each uses an independent set of measurements as it provides you with a probability distribution regarding your position (Fig. 4.8). Device *a* reports that your position is at coordinates $\mathbf{y}_a = (+1, 0)$, with a probability distribution specified by a Gaussian with covariance R_a (a 2x2 matrix). This is the location for which its probability distribution has a peak (i.e., the mean). The second device reports that your position is at coordinate $\mathbf{y}_b = (+4, 0)$, with a probability distribution with covariance R_b. Interestingly, your most likely location is not somewhere between the two centers of probability distributions.

Rather, given the uncertainties associated with each device, as depicted by the distributions shown in Fig. 4.8, your most likely location is probably at coordinates (+2.5, -1.5).

To see why this is the case, suppose that the state that we wish to estimate (our position) is described by a 2x1 vector $\mathbf{x}$, and the reading from each device is described by 2x1 vectors $\mathbf{y}_a$ and $\mathbf{y}_b$. To estimate our position, we need to put forth a hypothesis regarding how the devices' readings are related to our position. Suppose that our measurement, denoted by 4x1 vector $\mathbf{y} = \begin{bmatrix} \mathbf{y}_a & \mathbf{y}_b \end{bmatrix}$, is related to a hidden state (our position) $\mathbf{x}$ by the following equation:

$$\mathbf{y} = C\mathbf{x} + \boldsymbol{\varepsilon} \tag{4.1}$$

This is our internal model. It describes our belief about how the data that we observe is related to the hidden state that we wish to estimate. In our internal model, we believe that the devices are unbiased. Therefore, we set

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} I_{2x2} \\ I_{2x2} \end{bmatrix} \tag{4.2}$$

where I is a 2x2 identity matrix. We also believe that the noises inherent in the measurements of the devices are independent, zero mean, and distributed as a Gaussian:

$$\begin{aligned} \boldsymbol{\varepsilon} &\sim N(0, R) \\ R &= \begin{bmatrix} R_a & 0 \\ 0 & R_b \end{bmatrix} \end{aligned} \tag{4.3}$$

where R_a and R_b are 2x2 symmetric matrices. The expected value of $\mathbf{y}$, written as $E(\mathbf{y})$, is $E(\mathbf{y}) = C\mathbf{x}$. The variance of $\mathbf{y}$, written as $\text{var}(\mathbf{y})$, is $\text{var}(\mathbf{y}) = C\,\text{var}(\mathbf{x})C^T + R$. The probability distribution of $\mathbf{y}$ is specified by a Gaussian:

$$\mathbf{y} \sim N\left(C\mathbf{x}, C\,\text{var}(\mathbf{x})C^T + R\right). \tag{4.4}$$

This implies that if we 'knew' our position $\mathbf{x}$ with certainty, i.e. if $\text{var}(\mathbf{x}) = 0$ (or alternatively, we stayed still and kept on taking measurements), the measurements from the devices would have the following distribution:

$$p(\mathbf{y}|\mathbf{x}) = N(C\mathbf{x}, R). \tag{4.5}$$

This equation is called a *likelihood*. It describes the probability distribution of an observation given that the thing that we want to estimate (the hidden state $\mathbf{x}$) is at a particular value. To find our most likely position, we find the value for $\mathbf{x}$ that maximizes this likelihood:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^4 |R|}} \exp\left[-\frac{1}{2} (\mathbf{y} - C\mathbf{x})^T R^{-1} (\mathbf{y} - C\mathbf{x}) \right]. \tag{4.6}$$

It is convenient to take the log of the above expression, arriving at

$$\ln p(\mathbf{y}|\mathbf{x}) = -2\ln(2\pi) - \frac{1}{2}\ln|R| - \frac{1}{2}(\mathbf{y} - C\mathbf{x})^T R^{-1} (\mathbf{y} - C\mathbf{x}). \tag{4.7}$$

To find the location $\mathbf{x}$ for which this quantity is maximum, we find the location at which its derivative is zero:

$$\frac{d}{d\mathbf{x}} \ln p(\mathbf{y}|\mathbf{x}) = \left(C^T R^{-1} \mathbf{y} - C^T R^{-1} C\mathbf{x} \right)$$

$$\hat{\mathbf{x}} = \left(C^T R^{-1} C \right)^{-1} C^T R^{-1} \mathbf{y} \tag{4.8}$$

If we note that $R^{-1} = \begin{bmatrix} R_a^{-1} & 0 \\ 0 & R_b^{-1} \end{bmatrix}$ and $C^T R^{-1} = \begin{bmatrix} R_a^{-1} & R_b^{-1} \end{bmatrix}$, then we can rewrite Eq. (4.8) as:

$$\hat{\mathbf{x}} = \left(R_a^{-1} + R_b^{-1} \right)^{-1} \left(R_a^{-1} \mathbf{y}_a + R_b^{-1} \mathbf{y}_b \right) \tag{4.9}$$

Eq. (4.9) describes our maximum likelihood estimate of $\mathbf{x}$. Now if we simply stay still and keep taking measurements, our readings $\mathbf{y}$ will keep changing. These changes are due to the inherent noise in the devices, and will produce changes in our estimate $\hat{\mathbf{x}}$. Therefore, $\hat{\mathbf{x}}$ is a random variable with a distribution. Its expected value is specified by Eq. (4.8) and its variance is:

$$\mathrm{var}(\hat{\mathbf{x}}) = \left(C^T R^{-1} C \right)^{-1} C^T R^{-1} \mathrm{var}(\mathbf{y}) R^{-T} C \left(C^T R^{-1} C \right)^{-T} \tag{4.10}$$

Assuming that we stay still and do not move around, then $\mathrm{var}(\mathbf{y}) = R$. Because R is symmetric (as are all variance-covariance matrices), the above equation simplifies to:

$$\mathrm{var}(\hat{\mathbf{x}}) = \left(C^T R^{-1} C \right)^{-1} \tag{4.11}$$

Note that $C^T R^{-1} = \begin{bmatrix} R_a^{-1} & R_b^{-1} \end{bmatrix}$ simplifies the above equation to:

$$\mathrm{var}(\hat{\mathbf{x}}) = \left(R_a^{-1} + R_b^{-1} \right)^{-1} \tag{4.12}$$

The result in Eq. (4.9) indicates that our most likely location is one that weighs the reading from each device by the inverse of the device's probability covariance. In other words, we should

discount the reading from each device according to the inverse of each device's uncertainty. Using Eq. (4.9) and Eq. (4.12), we have drawn the mean and variance of our maximum likelihood estimate in Fig. 4.8 (the distribution in the middle of the figure). It is quite unlikely that you are somewhere between the centers of left and right distributions because $p(\mathbf{y}|\mathbf{x})$ is quite low there. The most likely location, it turns out, is a bit south of the center of each distribution, at (+2.5, -1.5). Another important point to note is that the estimate of your position has a tighter distribution than either of the two sensors, i.e., it has a 'smaller' variance. Therefore, when you combined these two pieces of information, your result was an estimate that had a lower uncertainty than either of the two initial measures. It is better to have two GPSs than one.

4.5 Optimal integration of sensory information by the brain

It turns out that this framework is quite relevant to how the brain perceives the environment and processes sensory information. After all, we have multiple sensors. For example, when we examine an object, we do so with both our hands and our eyes. Marc Ernst and Marty Banks (2002) were first to demonstrate that when our brain makes a decision about a physical property of an object, it does so by combining various sensory information about that object in a way that is consistent with maximum likelihood state estimation.

Ernst and Banks began by considering a hypothetical situation in which one has to estimate the height of an object. Suppose that you use your index and thumb to hold an object. Your somatosensory/haptic system reports its height. Let us represent this information as a random variable y_h (a scalar quantity) that has as a distribution described by a Gaussian with variance σ_h^2. Similarly, your visual system provides you with information y_v, which has a variance σ_v^2. Our internal model is:

$$\mathbf{y} = \mathbf{c}x + \boldsymbol{\varepsilon} \tag{4.13}$$

where x is the true height of the object, $\mathbf{y} = \begin{bmatrix} y_h & y_v \end{bmatrix}^T$, $\mathbf{c} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ (which implies that your sensors are unbiased), $\boldsymbol{\varepsilon} \square N(0, R)$, and $R = \begin{bmatrix} \sigma_h^2, 0; 0, \sigma_v^2 \end{bmatrix}$. According to Eq. (4.9), your maximum likelihood estimate of the height of the object has the following distribution:

$$E(\hat{x}) = \frac{1/\sigma_h^2}{1/\sigma_h^2 + 1/\sigma_v^2} y_h + \frac{1/\sigma_v^2}{1/\sigma_h^2 + 1/\sigma_v^2} y_v$$
$$\text{var}(\hat{x}) = \frac{1}{1/\sigma_h^2 + 1/\sigma_v^2} \quad (4.14)$$

If the noise in the two sensors is equal, then the weights that you apply to the sensors (the coefficients in front of y_h and y_v in Eq. 4.14) are equal as well. This case is illustrated in the left column of Fig. 4.9. On the other hand, if the noise in y_h is larger than y_v, your uncertainty for the haptic measure is greater and so you apply a smaller weight to its reading (illustrated in the right column of Fig. 4.9). If one was to ask you to report the height of the object, of course you would not report your belief as a probability distribution. To estimate this distribution, Ernst and Banks acquired a psychometric function, shown in the lower part of the graph in Fig. 4.9. To acquire this function, they provided their subjects a standard object of height 5.5cm. They then presented a second object of variable length and asked whether it was taller than the first object. If the subject represented the height of the standard object with a maximum likelihood estimate with a distribution of Eq. (4.14), then the probability of classifying the second object as being taller is simply the cumulative probability distribution of $\hat{x}$. This is called a psychometric function, and is shown in the lower row of Fig. 4.9. The point of subject equality (PSE) is the height at which the probability function is at 0.5. Note that this point shift toward the estimate from vision when one is more certain about the visual measurement (right column of Fig. 4.9).

The task at hand is to test whether the brain combines haptic and sensory information in a way that is consistent with maximum likelihood. To test the theory, the first step is to estimate the noise in the haptic and visual sensors. To do so, Ernst and Banks considered a situation in which both the standard and the second stimulus were of the same modality. For example, a robot presented a virtual object (standard, of length μ_1) and then a second object (of length $\mu_1 + \Delta$). [People held the handle of the robot and when they moved it, they felt a stiff surface of a given length. So there was no real object there, just a sensation associated with running your hand along the surface of an object.] The subject responded by choosing one that had a longer height. To make this decision, suppose that the subject's haptic sense represented the objects as y_1 and y_2, where $y_1 \sim N(\mu_1, \sigma_h^2)$ and $y_2 \sim N(\mu_1 + \Delta, \sigma_h^2)$. The estimate of the difference between

the two objects is $\hat{\Delta} = y_2 - y_1$ and this random variable is a Gaussian with the following distribution: $\hat{\Delta} \sim N\left(\Delta, 2\sigma_h^2\right)$. The probability of picking the second object as being taller is:

$$\Pr\left(y_2 > y_1\right) = \Pr\left(\hat{\Delta} > 0\right) = \int_0^\infty N\left(x; \Delta, 2\sigma_h^2\right) dx \tag{4.15}$$

The term $N\left(x; \Delta, 2\sigma_h^2\right)$ represents a normal distribution of random variable x with mean Δ and variance $2\sigma_h^2$. To compute this integral, we use the "error function":

$$\text{erf}(x) = \frac{2}{\pi}\int_0^\infty e^{-x^2} dx$$

$$\int_{-\infty}^{x} N\left(t; \mu, \sigma^2\right) dt = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

And so we have:

$$\Pr\left(\hat{\Delta} > 0\right) = 1 - \frac{1}{2}\left(1 + \text{erf}\left(\frac{-\Delta}{2\sigma_h}\right)\right) \tag{4.16}$$

For example, suppose that the second object is 3mm longer than the standard object, i.e., $\Delta = 3$. From Eq. (4.16), we would predict that if $\sigma_h^2 = 1$ (resulting in the two distributions plotted in Fig. 4.10A), then the subject should pick object 2 as being larger with 98% probability. In contrast, if $\sigma_h^2 = 2$, then the subject should pick object 2 as being larger with 69% probability, and so on. The condition for which $\Delta = 3$ and $\sigma_h^2 = 1$ is plotted in Fig. 4.10A. The resulting distribution for $\hat{\Delta}$ is plotted in Fig. 4.10B. The probability of Eq. (4.16) is plotted as a function of Δ, i.e., the difference in the lengths of the two objects, for various σ_h^2 in Fig. 4.10C. As the noise in the haptic sensor increases, the subject has more difficulty dissociating the two objects at a given length difference.

In Eq. (4.16), notice that if we set $\Delta = \sqrt{2}\sigma_h$, then the probability of picking object two as being larger is always 84%. So Ernst and Banks presented second objects of various Δ and found the one for which the subject was 84% correct (call it Δ^*). From the results in Fig. 4.11A we see that this object was about 5mm longer than the standard, i.e., $\Delta^* \approx 5$. Therefore, the noise in the haptic sensory pathway must be:

$$\sigma_h^2 = \frac{1}{2}\left(\Delta^*\right)^2$$

A similar procedure was performed to estimate the noise in the visual pathway. Two objects were presented visually and as Fig. 4.11A suggests (0% noise curve), 84% probability of correct choice was for $\Delta^* \approx 2.5$. Therefore, $\sigma_h^2 \approx 4\sigma_v^2$. In the final step of the experiment, Ernst and Banks used these noise parameters to predict how subjects would estimate the height of an object when the visual and haptic senses were both present (as in Fig. 4.9). For example, subjects were presented with a standard object for which the haptic information indicated a height of μ_1 and visual information indicated a height of $\mu_1 + \Delta$. Eq. (4.14) predicted that subjects would assign a weight of around 0.8 to the visual information and around 0.2 to the haptic information. To estimate these weights, they presented a second object (for which the haptic and visual information agreed) and ask which one was taller. The probability of the second object being taller is plotted in Fig. 4.11B, and the weight assigned to each sensory modality is plotted in Fig. 4.11C. The observed weights (Fig. 4.11C) agree quite well with the predicted values. Note that for the 0% visual noise condition, the curve is shifted toward the mean of the visual information. One can also predict the variance of the estimate (Eq. 4.14), which specifies the rate of rise in the psychometric curve. A proxy for this rate is a 'discrimination threshold', defined as the difference between the point of subject equality and the height of the second stimulus when it is judged to be taller than the standard stimulus 84% of the time. As the variance of the estimate increases, the rate of rise decreases, and the discrimination threshold increases. Fig. 4.11D shows that the predicted value nicely matches the observed value. They next repeated these steps for conditions in which noise was added to the visual display. This reduced the weighting of the visual information, making the subjects rely more on the haptic information. The predicted values continued to match the measured quantities.

Therefore, the results showed that the brain combined visual and haptic information in a way that was similar to a maximum-likelihood estimator: a weight was assigned to each sensory modality and this weight was a function of the uncertainty in that modality.

4.6 Uncertainty

How does uncertainty come about? It has something to do with the precision with which one knows some bit of information. In particular, it seems to have something to do with the history of

how that information was acquired. In our hiking in the wood example, the uncertainty of each GPS device had a particular shape because there were correlations along the two dimensions in which each device was making measurements. Perhaps these correlations came about because of the position of the satellites, giving rise to the shape of the probability distribution. Let us explore this idea a bit further in a scenario in which we control how we acquire information about a quantity that we wish to estimate.

In Fig. 4.12 we have drawn points from three normal distributions. In each distribution, the vector $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ is drawn from a Gaussian: $\mathbf{x} \sim N(\boldsymbol{\mu}, R)$, with mean zero $\boldsymbol{\mu} = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ and covariance R. This covariance is defined as:

$$\begin{aligned} R &= E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \right] \\ &= \begin{bmatrix} \text{var}[x_1] & \text{cov}[x_1, x_2] \\ \text{cov}[x_2, x_1] & \text{var}[x_2] \end{bmatrix} \end{aligned} \tag{4.16}$$

In Fig. 4.12, the distribution on the left has a covariance matrix in which x_1 and x_2 are negatively correlated (the off-diagonal elements of the covariance matrix are negative). This roughly corresponds to our GPS device *a* in Fig. 4.8. We can now infer that the measurements that device *a* was taking along the east/west (x_1) and north/south (x_2) dimensions were showing negative correlations: as the measurement along one dimension increased, the measurement along the other dimension decreased. So when device *a* reported its measurement, it was uncertain along the northwest-southeast dimension, but very certain along the southwest-northeast dimension.

In Fig. 4.12 the distribution at right has a covariance in which x_1 and x_2 are positively correlated. This roughly corresponds to our device *b*, indicating that for this device, as the measurement along one dimension increased, so did the measurement along the other dimension. The overlap between the left-most distribution and right-most distribution (assuming that the centers are not at the same location) would be maximum (i.e., most number of points likely to overlap) at somewhere along their diagonals.

Our simple example illustrates the idea that in order for any system to combine two pieces of information about a single quantity, it needs to have a measure of uncertainty about each quantity.

That is, it needs to 'know' the two pieces of information in terms of their probability distributions. If the distribution is Gaussian, all we need to know are the mean and covariance.

Now to illustrate how these correlations, i.e., uncertainties, arise during the process of acquiring information, let us consider the classic problem in linear regression. On each trial *i*, you measure a scalar quantity $y^{(i)}$. However, your measurements are noisy. The 'true' value on that trial is $y^{*(i)}$ and your measure is corrupted by a Gaussian noise that has zero mean and σ^2 variance:

$$y^{(i)} = y^{*(i)} + \varepsilon \qquad \varepsilon \square N\left(0, \sigma^2\right) \tag{4.17}$$

This means that the probability to find a value of $y^{(i)}$ within a given interval of the true value is

$$\Pr[y^{*(i)} + \varepsilon_1 \le y^{(i)} \le y^{*(i)} + \varepsilon_2] = \frac{1}{\sigma\sqrt{2\pi}} \int_{\varepsilon_1}^{\varepsilon_2} e^{-\frac{\varepsilon^2}{\sigma^2}}\, d\varepsilon \tag{4.18}$$

We assume that the term $y^{*(i)}$ is linearly related to a known quantity expressed as a N-dimensional vector, $\mathbf{x}^{(i)}$:

$$y^{*(i)} = \mathbf{w}^T \mathbf{x}^{(i)} \tag{4.19}$$

In the language of systems theory, $\mathbf{x}^{(i)}$ is an input vector and $y^{*(i)}$ is the corresponding output generated by the linear transformation. In this case the unknown multiplicative term $\mathbf{w}^T$ is a 1xN matrix – i.e. a row vector, mapping an N-dimensional vector into a single number. According to this simple model, noise is only affecting the observation of the output variable $y^{(i)}$. More complex situations may involve noise in the input vector $\mathbf{x}^{(i)}$ and in the vector $\mathbf{w}$. Our objective is to estimate the vector $\mathbf{w}$ from a data set that includes pairs of input/output measures: $D = \left(\left\{\mathbf{x}^{(1)}, y^{(1)}\right\}, \left\{\mathbf{x}^{(2)}, y^{(2)}\right\}, \cdots, \left\{\mathbf{x}^{(n)}, y^{(n)}\right\}\right)$. Given our model in Eqs. (4.17) and (4.19), we can write the probability distribution of observing $y^{(i)}$, given input $\mathbf{x}^{(i)}$ and parameters $\mathbf{w}$ and σ^2:

$$p\left(y^{(i)} \middle| \mathbf{x}^{(i)}\right) = N\left(\mathbf{w}^T \mathbf{x}^{(i)}, \sigma^2\right) \tag{4.20}$$

That is, we would expect $y^{(i)}$ to be normally distributed with mean $\mathbf{w}^T \mathbf{x}^{(i)}$ and variance σ^2. The probability density function for observing the specific data that we were given is simply the joint probability density of all data points:

$$p\left(y^{(1)},\cdots,y^{(n)}\middle|\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(n)},\mathbf{w},\sigma\right)=\prod_{i=1}^{n}p\left(y^{(i)}\middle|\mathbf{x}^{(i)},\mathbf{w},\sigma\right)$$
$$=\prod_{i=1}^{n}\frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)}-\mathbf{w}^T\mathbf{x}^{(i)}\right)^2\right) \quad (4.21)$$
$$=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)}-\mathbf{w}^T\mathbf{x}^{(i)}\right)^2\right)$$

The quantity in Eq. (4.21) is a likelihood. It describes how likely it is that given our model and its parameters, we would observe the specific collection of data $\left\{y^{(1)},\ldots,y^{(n)}\right\}$. Let us refer to our likelihood as $L\left(\mathbf{w},\sigma\right)$. The parameters that we would like to estimate are $\mathbf{w}$ and σ. If we have no other information about the parameters that we wish to estimate (i.e., no priors), then the best estimate of these parameters are those that maximize the likelihood. We start by observing that the exponential function is monotonic in its argument. Therefore, the exponential is at a maximum when its argument is also at a maximum. Because in the Normal distribution of Eq. (4.20) the argument is always negative, the maximum of the likelihood function is attained when the argument of the exponential reaches a minimum. To put into a formula, we take the natural logarithm of the function that we wish to maximize:

$$\log L\left(\mathbf{w},\sigma\right)=\log\prod_{i=1}^{n}p\left(y^{(i)}\middle|\mathbf{x}^{(i)},\mathbf{w},\sigma\right)$$
$$=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)}-\mathbf{w}^T\mathbf{x}^{(i)}\right)^2-\frac{n}{2}\log\left(2\pi\sigma^2\right) \quad (4.22)$$

and look for its maximum. To further simplify things, let us write the first sum in the second line of the above equation in matrix form. Suppose we use vector $\mathbf{y}$ and matrix X to refer to collection of $y^{(i)}$ and $\mathbf{x}^{(i)}$ (the vector $\mathbf{x}^{(i)}$ becomes a row in the matrix X), and then rewrite Eq. (4.22). This requires a little manipulation, but the final result is worth it. We begin by collecting all the $\mathbf{x}^{(i)}$ vectors into a matrix, one vector per row:

$$X=\begin{bmatrix}x_1^{(1)} & x_2^{(1)} & \cdots & x_N^{(1)}\\ x_1^{(2)} & x_2^{(2)} & \cdots & x_N^{(2)}\\ \cdots & \cdots & \cdots & \cdots\\ x_1^{(n)} & x_2^{(n)} & \cdots & x_N^{(n)}\end{bmatrix} \quad (4.23)$$

and all the output data $y^{(i)}$, in a column vector

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} \tag{4.24}$$

Then, the vector

$$\begin{aligned} \mathbf{y} - X\mathbf{w} &= \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} - \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_N^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_N^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_N^{(n)} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_N \end{bmatrix} \\ &= \begin{bmatrix} y^{(1)} - \sum_{k=1}^{N} w_k \; x_k^{(1)} \\ y^{(2)} - \sum_{k=1}^{N} w_k \; x_k^{(2)} \\ \cdots \\ y^{(n)} - \sum_{k=1}^{n} w_k \; x_N^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} - w^T \cdot \mathbf{x}^{(1)} \\ y^{(2)} - w^T \cdot \mathbf{x}^{(2)} \\ \cdots \\ y^{(n)} - w^T \cdot \mathbf{x}^{(n)} \end{bmatrix} \end{aligned} \tag{4.25}$$

has a Euclidean norm

$$\left\| \mathbf{y} - X \cdot \mathbf{w} \right\|^2 = \left(\mathbf{y} - X \cdot \mathbf{w} \right)^T \left(\mathbf{y} - X \cdot \mathbf{w} \right) = \sum_{i=1}^{n} \left(y^{(i)} - w^T \mathbf{x}^{(i)} \right)^2 . \tag{4.26}$$

Note that this provides a way to re-write the sum on the left side of Eq. (4.22). With this substitution, we obtain:

$$\begin{aligned} \log L\left(\mathbf{w}, \sigma\right) &= -\frac{1}{2\sigma^2} \left(\mathbf{y} - X\mathbf{w} \right)^T \left(\mathbf{y} - X\mathbf{w} \right) - \sum_{i=1}^{n} \log\left(2\pi\sigma^2 \right)^{1/2} \\ &= -\frac{1}{2\sigma^2} \left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\mathbf{w} - \mathbf{w}^T X^T \mathbf{y} + \mathbf{w}^T X^T X\mathbf{w} \right) \\ &\quad -\sum_{i=1}^{n} \left(\log\left(2\pi\right)^{1/2} + \log\sigma \right) \end{aligned} \tag{4.27}$$

To find $\mathbf{w}$ at which this function is maximum, we find its derivative and set it equal to zero:

$$\begin{aligned} \frac{d \log L\left(\mathbf{w}, \sigma\right)}{d\mathbf{w}} &= -\frac{1}{2\sigma^2} \left(-2X^T\mathbf{y} + 2X^T X\mathbf{w} \right) = 0 \\ X^T X\mathbf{w} &= X^T\mathbf{y} \\ \hat{\mathbf{w}} &= \left(X^T X \right)^{-1} X^T\mathbf{y} \end{aligned} \tag{4.28}$$

We observe that this solution for $\mathbf{w}$ corresponds to a maximum of the log likelihood, because the second derivative (or Hessian) of $\log L$:

$$\frac{d^2 \log L(\mathbf{w},\sigma)}{d\mathbf{w}^2} = -\frac{1}{\sigma^2} X^T X \tag{4.29}$$

is negative definite, if X is full row rank (i.e. if $rank(X) = N$). This result is interpreted intuitively by observing that the matrix $\left(X^T X\right)^{-1} X^T$ is the left inverse of the input data matrix, X. Therefore the maximum likelihood estimation of $\mathbf{w}$ is analogous to a ratio of the output to the input. We derive the same solution by looking for the parameters that minimize the square difference from the observed outputs and the outputs that are obtained by assuming the model of Eq. (4.19). Thus, we have found a correspondence between the maximum likelihood estimation and the least-squares estimation, in this linear case with Gaussian noise.

Eq. (4.28) represents our maximum likelihood estimate of $\mathbf{w}$. Similarly, to find the maximum likelihood estimate of noise parameter σ we take the derivative of Eq. (4.27) with respect to σ and set it equal to zero.

$$\begin{aligned} \frac{dl}{d\sigma} &= \frac{1}{\sigma^3}(\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w}) - \sum_{i=1}^{n} \frac{1}{\sigma} = 0 \\ &= \frac{1}{\sigma^2}(\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w}) - n = 0 \\ \hat{\sigma}^2 &= \frac{1}{n}(\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w}) \end{aligned} \tag{4.30}$$

However, it is important to note that because **y** is a random variable whose value is corrupted by random noise, our estimates $\hat{\mathbf{w}}$ and $\hat{\sigma}$ are also random variables. To see why this is true, consider how we got our data set D: for some known set of inputs $\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(n)}$, we made a set of measurements $y^{(1)},\cdots,y^{(n)}$ that were generated by the "true" parameter vector $\mathbf{w}$ but were corrupted with random noise $\varepsilon^{(1)},\cdots,\varepsilon^{(n)}$. If someone gave the same inputs again to the system, it would not generate exactly the same measurements as before. That is, given a set of inputs $\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(n)}$ in data set $D^{(1)}$, we can estimate an optimal $\hat{\mathbf{w}}$ as in Eq. (4.28). But if the same inputs are used for generating a second data set $D^{(2)}$, our estimate $\hat{\mathbf{w}}$ would generally be different from $D^{(1)}$. Indeed if we do this over and over, we would see that $\hat{\mathbf{w}}$ has a distribution

that depends upon the statistics of the noise on the data. We can compute this distribution as follows:

$$\begin{aligned}\hat{\mathbf{w}} &= \left(X^T X\right)^{-1} X^T \mathbf{y} \\ &= \left(X^T X\right)^{-1} X^T \mathbf{y}^* + \left(X^T X\right)^{-1} X^T \boldsymbol{\varepsilon} \\ &= \mathbf{w} + \left(X^T X\right)^{-1} X^T \boldsymbol{\varepsilon}\end{aligned} \tag{4.31}$$

In Eq. (4.31), the vector $y*$ is the vector of "true" outputs $\left[y*^{(1)}, \quad y*^{(2)}, \quad \ldots, \quad y*^{(n)}\right]^T$ and the term $\boldsymbol{\varepsilon}$ is the corresponding vector random variable whose elements are the noises ε in Eq. (4.17). From Eq. (4.31) we can compute the probability distribution of $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} \sim N\left(\mathbf{w}, \text{var}\left(\left(X^T X\right)^{-1} X^T \boldsymbol{\varepsilon}\right)\right) \tag{4.32}$$

The term 'var' inside the parenthesis is the variance-covariance matrix of the distribution. We compute it as follows:

$$\begin{aligned}\text{var}\left(\left(X^T X\right)^{-1} X^T \boldsymbol{\varepsilon}\right) &= \left(X^T X\right)^{-1} X^T \,\text{var}(\boldsymbol{\varepsilon})\, X \left(X^T X\right)^{-T} \\ &= \left(X^T X\right)^{-1} X^T I\sigma^2 X \left(X^T X\right)^{-T} \\ &= \sigma^2 \left(X^T X\right)^{-1}\end{aligned} \tag{4.33}$$

So, the probability distribution of our estimate $\hat{\mathbf{w}}$ has a mean equal to the 'true' value $\mathbf{w}$, but a covariance that depends both on the measurement noise σ^2 and the specific inputs $\mathbf{x}^{(i)}$ (recall that the rows of matrix *X* consists of $\mathbf{x}^{(i)}$). Intuitively, the variance of the vector $\mathbf{w}$ is simply the ratio of the output variance to the input covariance. Because we have assumed that the input variable is noiseless, then a large excursion in the input vector value - corresponding to higher covariance values, would lead to a diminished uncertainty on the estimate $\hat{\mathbf{w}}$. Larger variability of the input corresponds to a broader exploration of the domain over which the transformation from **x** to **y** is defined. Basically, this means that when we are given input data that ranges over a large region, we are more certain about our estimate $\hat{\mathbf{w}}$. Thus, it is not surprising that the estimate of $\mathbf{w}$ becomes more accurate. In contrast, large amount of output variability can only reduce our confidence on the estimate of $\mathbf{w}$.

Now, returning to the question at the beginning of this section: where does uncertainty in our estimate $\hat{\mathbf{w}}$ come from? From Eq. (4.33) we note that some of the uncertainty comes from the fact that our measurements $y^{(i)}$ were affected by noise (with variance σ^2). If this noise was large, we would be more uncertain about our estimate. But this point is trivial. More interestingly, the uncertainty regarding $\hat{\mathbf{w}}$ also depends on the history of inputs $\mathbf{x}^{(i)}$, which are the elements of the matrix *X*. This is a very important idea and one that is worth thinking about.

To see how the history of inputs $\mathbf{x}^{(i)}$ affects the probability distribution of $\hat{\mathbf{w}}$, let us try some example data sets. Suppose that we are given five data points, as shown in the table at the top part of Fig. 4.13A. Each row is a data point, where $\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} \end{bmatrix}^T$ and $y^{(i)} = y^{*(i)} + \varepsilon$ (the noise is not explicitly specified in the table, but it has a normal distribution). What we need to estimate is $\mathbf{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix}^T$ in our model $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \varepsilon$. We notice that in our data, x_1 is 'on' most of the time, x_2 is 'on' once, and x_1 and x_2 are never 'on' together. So when we go and estimate w_1 and w_2, we should be able to estimate w_1 with a lot of confidence, but perhaps not so for w_2. If we generate lots of data sets from this table (that is, we keep $\mathbf{x}^{(i)}$ as specified in this table and generate $y^{(i)} = y^{*(i)} + \varepsilon$ by adding random noise), and find $\hat{\mathbf{w}}$ for each data set, we will end up with a distribution shown in Fig. 4.13A. The distribution demonstrates that we can be pretty certain of $\hat{w}_1$ but we will be fairly uncertain for $\hat{w}_2$. This is simply a reflection of the amount of information we received for x_1 and x_2 (we got more information, in some vague sense, about x_1 than x_2). Furthermore, the distribution has zero covariance between $\hat{w}_1$ and $\hat{w}_2$, which is a reflection of the fact that x_1 and x_2 were never 'on' together. In the case of the data set in Fig. 4.13B, x_1 and x_2 are mostly 'on' together. The resulting $\hat{\mathbf{w}}$ will have the same mean as in the data in Fig. 4.13A (that is, the distribution is centered at +0.5, +0.5). However, now all that we can say with certainty is that as $\hat{w}_1$ increases, $\hat{w}_2$ should decrease. This is reflected in the negative covariance of the distribution.

Therefore, in linear regression the probability distribution associated with the parameter that we are trying to estimate is implicitly a reflection of the statistics of inputs, that is, the data that we were given to form our estimate. In a sense, the covariance of the probability distribution keeps a

record of this data. In the hiking problem, the uncertainty of each device was due to the history of its measurements. Similarly in the regression problem, the uncertainty of our estimate is due to the history of our measurements. In the next chapter we will consider the problem of learning and we will see that when animals learn, they too appear to keep a measure of uncertainty about their estimates, and this uncertainty is a reflection of the history of the stimuli that they observed.

4.7 State estimation and the Kalman filter

In the hiking problem, we used the uncertainties of each device to combine the two measurements. The problem is basically the same when you want to combine a predicted value for a quantity with the measured value for that same quantity. For example, the evidence in Fig. 4.7 suggests that when the brain estimates the position of the hand-held cursor, it does so by optimally combining the value that it predicted with the value that it measured in such a way as to render minimal the variance of this estimate. The result is a 'belief' or estimate that is a weighted combination of the predicted and measured values.

Unfortunately, our framework is still quite weak as it suffers from two flaws. First, we cannot incorporate our prior knowledge into the problem of estimation. For example, if we have hiked in the region before, we have some idea of where we might be. It is not clear how to combine this information with our readings from the GPS devices. Second, when we make a movement, predictions and measurements are not quantities that occur once; they are continuous streams of information. We need to continuously combine our predictions with observations, form beliefs, and then update our predictions for the future.

To give you a simple example of this, consider the problem of lifting one of those insulated coffee cups that they sell for taking your coffee with you for a drive. These cups have a lid that makes it impossible to see how much liquid is inside. Suppose that at a coffee shop, the person behind the counter just placed such a cup on the counter, so you believe it to be full, but it is in fact empty. When you go to lift it, your hand will jerk upward. Your prediction about the cup's weight was larger than the amount reported by the proprioception from your arm. Your belief about the actual weight will shift from your prediction toward the observation. As the hand keeps moving upward, this belief will converge to the mean value reported by the sensory apparatus. The rate of this convergence will have a lot to do with how strongly the brain believed in its own

predictions, i.e., the uncertainty of the prediction. Said in simple terms, if you have a belief that is very certain, it will take a lot of evidence (i.e., data) to change your mind.

To modify our framework so we can approach this problem, it is useful to consider again the problem of regression, but now in a trial-by-trial basis rather than in a batch form in which all the data were given to us at once. Suppose that on trial *i*, we are given input $\mathbf{x}^{(i)}$ and we use our current estimate $\hat{\mathbf{w}}^{(i)}$ to predict that the output should be $\hat{y}^{(i)} = \mathbf{x}^{(i)T}\hat{\mathbf{w}}^{(i)}$. However, we observe the quantity $y^{(i)}$ instead. So we have a difference between what we predicted and what we observed. Note that earlier we wrote it differently as $\hat{y}^{(i)} = \hat{\mathbf{w}}^{(i)T}\mathbf{x}^{(i)}$. If two vectors have only real components, then the scalar product is symmetric, since $\hat{\mathbf{w}}^{(i)T}\mathbf{x}^{(i)} = \mathbf{x}^{(i)T}\hat{\mathbf{w}}^{(i)}$. The first form, $\hat{y}^{(i)} = \mathbf{x}^{(i)T}\hat{\mathbf{w}}^{(i)}$ will turn out to be convenient in the following derivation. We need to combine the two pieces of information in order to update our estimate:

$$\hat{\mathbf{w}}^{(i+1)} = \hat{\mathbf{w}}^{(i)} + \mathbf{k}^{(i)}\left(y^{(i)} - \mathbf{x}^{(i)T}\hat{\mathbf{w}}^{(i)}\right) \tag{4.34}$$

It seems rational that the term $\mathbf{k}^{(i)}$ (our 'sensitivity to prediction error') should somehow reflect our uncertainty about our estimate. This term is a column vector with as many entries as the parameter vector $\hat{\mathbf{w}}^{(i)}$. The quantity in parenthesis in Eq. (4.34) is a single number, the prediction error. If $\mathbf{x}^{(i)}$ was oriented along a dimension for which $\hat{\mathbf{w}}^{(i)}$ was uncertain then we should learn a lot from the prediction error. If on the other hand $\mathbf{x}^{(i)}$ was oriented along a dimension for which our estimate of $\hat{\mathbf{w}}^{(i)}$ was quite certain, then perhaps our prediction error would simply be due to measurement noise, and we should basically ignore it. Of course, this is a simplification. A large error in a dimension in which we have great certainty would actually be likely to induce a large change in the estimate. However, this would be a rather "catastrophic" or discontinuous change. Whereas here, we are considering only smooth gradual changes. The term $\mathbf{k}^{(i)}$ in Eq. (4.34) is called a Kalman gain. It was named after Rudolph Kalman, an electrical engineer and mathematician working at the Research Institute for Advanced Study in Baltimore, Maryland. He discovered a principled way to set this sensitivity (Kalman, 1960), reflecting both the uncertainty in the prediction and the uncertainty in the measurement.

Let us change slightly the notation and refer to our prior knowledge about parameter **w** on trial *n* as $\hat{\mathbf{w}}^{(n|n-1)}$. That is, our estimate of **w** on trial *n*, given the past $n-1$ trials, is $\hat{\mathbf{w}}^{(n|n-1)}$. After

we make an observation (i.e. we measure $y^{(n)}$), we form a new or "posterior" estimate. We can rewrite Eq. (4.34) using this new terminology:

$$\hat{\mathbf{w}}^{(n|n)} = \hat{\mathbf{w}}^{(n|n-1)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\hat{\mathbf{w}}^{(n|n-1)}\right) \tag{4.35}$$

We now express the value of the output $y^{(n)}$ in terms of the true value of **w** and its associated uncertainty. This is done by inserting Eqs. (4.17) and (4.19) into (4.35):

$$\begin{aligned}\hat{\mathbf{w}}^{(n|n)} &= \hat{\mathbf{w}}^{(n|n-1)} + \mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}\mathbf{w} + \varepsilon^{(n)} - \mathbf{x}^{(n)T}\hat{\mathbf{w}}^{(n|n-1)}\right) \\ &= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)\hat{\mathbf{w}}^{(n|n-1)} + \mathbf{k}^{(n)}\varepsilon^{(n)} + \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\mathbf{w}\end{aligned} \tag{4.36}$$

Let us define $P^{(n|n-1)}$ to be the variance (i.e., uncertainty) of our prior estimate $\hat{\mathbf{w}}^{(n|n-1)}$:

$$P^{(n|n-1)} \equiv \mathrm{var}\left(\hat{\mathbf{w}}^{(n|n-1)}\right) \tag{4.37}$$

Similarly, we define $P^{(n|n)}$ to be the variance of our posterior estimate $\hat{\mathbf{w}}^{(n|n)}$:

$$P^{(n|n)} \equiv \mathrm{var}\left(\hat{\mathbf{w}}^{(n|n)}\right) \tag{4.38}$$

From Eq. (4.36), we can write our posterior variance as a function of the prior variance:

$$\begin{aligned}P^{(n|n)} &= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)}\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)^T + \mathbf{k}^{(n)}\,\mathrm{var}\left(\varepsilon^{(n)}\right)\mathbf{k}^{(n)T} \\ &= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)}\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)^T + \sigma^2\mathbf{k}^{(n)}\mathbf{k}^{(n)T}\end{aligned} \tag{4.39}$$

[The last term in Eq. (4.36) depends on **w**, which is not a random variable and has zero variance.]

Now our problem is to set the term $\mathbf{k}^{(n)}$ in such a way so that our posterior estimate $\hat{\mathbf{w}}^{(n|n)}$ is as certain as possible, i.e., it is minimum variance. Therefore, our problem is to find $\mathbf{k}^{(n)}$ so to minimize $P^{(n|n)}$. However, $P^{(n|n)}$ is a matrix, so we need to clarify what we mean by 'minimizing' it. This requires us to first define the size or "norm" of a matrix. The norm of a matrix must be a positive number whose value is zero only when the matrix is zero. There are several definitions that fit this criterion. One, for example, is the largest singular value. However, it would be difficult to use it for our purpose. Another kind of 'norm' is simply the trace of a matrix. This is clearly not a valid norm for all matrices, because a nonzero matrix can have a zero trace. However, the variance matrix can only have positive or zero terms along the main diagonal. And it is easy to verify that if all the terms along the diagonal are zero, i.e. if the trace

is zero, the whole variance matrix must also be zero [because the diagonal elements are σ_1^2, σ_2^2, etc., and the off-diagonal terms are $\rho_{1,2}\sigma_1\sigma_2$, $\rho_{2,1}\sigma_1\sigma_2$ etc., where $\rho_{1,2}$ is the correlation between $\hat{w}_1$ and $\hat{w}_2$]. Therefore the trace is a very simple way to measure the size of a variance matrix. If we minimize the trace of the matrix $P^{(n|n)}$, we minimize the sum of the diagonal elements. The diagonal elements of $P^{(n|n)}$ are variances of the individual elements of the vector $\hat{\mathbf{w}}$. By minimizing the diagonal elements, it may seem that one is ignoring the covariance terms. But this is not really the case, because the covariance terms are proportional to the square root of the individual variances. Kalman's approach was to set the term **k** so to minimize the trace of the posterior uncertainty. Multiplying out the terms in Eq. (4.39) and then finding their trace results in the following:

$$\begin{aligned} tr\left[P^{(n|n)}\right] = tr\left[P^{(n|n-1)}\right] - tr\left[P^{(n|n-1)}\mathbf{x}^{(n)}\mathbf{k}^{(n)T}\right] - tr\left[\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\right] \\ + tr\left[\mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\right] \end{aligned} \tag{4.40}$$

The trace is a linear operator with some nice properties that we can use to simplify Eq. (4.40). For example, $tr(A) = tr(A^T)$. Therefore:

$$\begin{aligned} tr\left[P^{(n|n)}\right] = tr\left[P^{(n|n-1)}\right] - 2tr\left[\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\right] \\ + tr\left[\mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\right] \end{aligned} \tag{4.41}$$

Noting that the term $\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2$ is a scalar quantity and that $tr[aA] = a\ tr[A]$, we have:

$$\begin{aligned} tr\left[P^{(n|n)}\right] = tr\left[P^{(n|n-1)}\right] - 2tr\left[\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\right] \\ + \left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)tr\left[\mathbf{k}^{(n)}\mathbf{k}^{(n)T}\right] \end{aligned} \tag{4.42}$$

One last important step can be taken by observing that the trace of the external product of two vectors (a column vector multiplying a row vector) is simply the dot product of the same vectors. Noting that the second and third terms of Eq. (4.42) contain traces of such external products, we can further simplify the expression:

$$tr\left[P^{(n|n)}\right] = tr\left[P^{(n|n-1)}\right] - 2\mathbf{k}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\mathbf{k}^{(n)} \tag{4.43}$$

To minimize this expression, we find its derivative with respect to **k**. We have:

$$\frac{d}{d\mathbf{k}^{(n)}}tr\left[P^{(n|n)}\right] = -2P^{(n|n-1)}\mathbf{x}^{(n)} + \left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\left(2\mathbf{k}^{(n)}\right) \tag{4.44}$$

We set Eq. (4.44) equal to zero and solve for **k**, resulting in:

$$\mathbf{k}^{(n)} = \frac{P^{(n|n-1)}}{\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)}\mathbf{x}^{(n)} \tag{4.45}$$

Because the second derivative of Eq. (4.43) is positive definite:

$$\frac{d^2}{d\mathbf{k}\mathbf{k}^T}tr\left[P^{(n|n)}\right] = 2\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right) \tag{4.46}$$

the expression in Eq. (4.45) corresponds to a minimum.

Note that the measurement uncertainty as expressed by the variance of the output is:

$$\mathrm{var}\left(y^{(n)}\right) = \mathrm{var}\left(\mathbf{w}^{(n|n-1)T}\mathbf{x}^{(n)} + \varepsilon\right) = \mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2 \tag{4.47}$$

Therefore, the optimum sensitivity to prediction error, or Kalman gain, is a ratio between our prior uncertainty on the parameter that we are trying to estimate (numerator of Eq. 4.45), and our measurement uncertainty (the denominator). If we are uncertain about what we know (numerator is 'large' relative to the denominator), we should learn a lot from the prediction error. If we are uncertain about our measurement (denominator is large relative to the numerator), we should ignore the prediction error. [You can perhaps already guess that the Kalman gain formula in Eq. (35) is a very useful model of biological learning, which is something that we will explore in detail in the next chapter.]

Our final step is to formulate the posterior uncertainty, i.e., the variance of $\mathbf{w}^{(n|n)}$. We insert Eq. (4.45) into Eq. (4.39), and after a bit of simplification[1] we arrive at the posterior uncertainty:

$$P^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)} \tag{4.48}$$

Since the term $\mathbf{k}^{(n)}\mathbf{x}^{(n)T}$ is positive definite, Eq. (4.48) implies that typically, our uncertainty declines as we make more observations.

In Fig. 4.14 we have summarized the problem that we just solved. We begin with the assumption that there was a hidden variable $\mathbf{w}$ that linearly interacts with a known quantity $\mathbf{x}^{(n)}$ on each trial to produce the measurement $y^{(n)}$. To estimate $\mathbf{w}$, we have a prior estimate $\hat{\mathbf{w}}^{(1|0)}$ and prior uncertainty $P^{(1|0)}$. We compute $\mathbf{k}^{(1)}$ using Eq. (4.45) and the posterior uncertainty using Eq. (4.48) and then update our estimate using the prediction error (Eq. 4.35). Because we assume that the variable $\mathbf{w}$ did not change from trial to trial, we compute the prior estimate for the next trial by simply setting it to the equal to the posterior estimate of the last trial:

$$\begin{aligned}\hat{\mathbf{w}}^{(n+1|n)} &= \hat{\mathbf{w}}^{(n|n)} \\ P^{(n+1|n)} &= P^{(n|n)}\end{aligned} \tag{4.49}$$

What if we believe that the variable $\mathbf{w}$ will not stay constant from trial to trial? For example, in Fig. 4.15 we have assumed that $\mathbf{w}$ changes from trial to trial. This change is simply governed by a matrix *A* (assumed to be known), and random noise $\boldsymbol{\varepsilon}_w$. Now our *generative model* (that is, the model that we assume is responsible for generating our data) becomes:

---

[1]

$$\begin{aligned}P^{(n|n)} &= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)}\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)^T + \sigma^2\mathbf{k}^{(n)}\mathbf{k}^{(n)T} \\ &= P^{(n|n-1)} - 2\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)} + \mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)}\mathbf{k}^{(n)T} + \mathbf{k}^{(n)}\sigma^2\mathbf{k}^{(n)T} \\ &= P^{(n|n-1)} - 2\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)} + \mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T} \\ &= P^{(n|n-1)} - 2\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)} + \mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{x}^{(n)T}\frac{P^{(n|n-1)}}{\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)} \\ &= P^{(n|n-1)} - 2\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)} + \mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)} \\ &= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)}\end{aligned}$$

$$\begin{aligned} \mathbf{w}^{(n+1)} &= A\mathbf{w}^{(n)} + \boldsymbol{\varepsilon}_w^{(n)} \qquad \boldsymbol{\varepsilon}_w \sim N(0,Q) \\ y^{(n)} &= \mathbf{x}^{(n)T}\mathbf{w}^{(n)} + \varepsilon_y^{(n)} \qquad \varepsilon_y \sim N\left(0,\sigma^2\right) \end{aligned} \tag{4.50}$$

In Eq. (4.50), the noise $\boldsymbol{\varepsilon}_w$ in the state update equation reflects our uncertainty about how $\mathbf{w}$ might change from trial to trial. The effect of this assumption is to alter how we project our estimates $\hat{\mathbf{w}}$ from the posterior estimate of one trial to the prior estimate of the next trial:

$$\begin{aligned} \hat{\mathbf{w}}^{(n+1|n)} &= A\hat{\mathbf{w}}^{(n|n)} \\ P^{(n+1|n)} &= AP^{(n|n)}A^T + Q \end{aligned} \tag{4.51}$$

The algorithm for solving this version of our problem is summarized in Fig. 4.15.

4.8 Combining predictions with delayed measurements

We now have an algorithm to continuously combine our prior predictions with observations, form posterior beliefs, and then form predictions for the future observations. We have the tools in place to test a simple prediction of our framework: when the brain estimates a value but this value differs from a measured value, the rate at which belief is corrected and converges onto the measured value will depend on the uncertainty of the predictions - the higher the certainty of prediction, the slower the convergence. In another words, the more confident you are about your predictions, the more evidence you will need to change your mind.

Jun Izawa and Shadmehr (Izawa and Shadmehr, 2008) tested this idea by having people move a handheld LED to the center of a target (Fig. 4.16A). People could view the handheld LED at all times, but the position of the target was made uncertain by displaying it as a 'blob' (a display in which pixel intensity was described by a Gaussian). This noise affected the certainty with which the subject could estimate the target position. For example, in trials in which the target's position was certain, reaction time (time to start the reach) was shorter and endpoint variance was small. In a fraction of the trials, as the movement was unfolding the target changed position by a small amount. When it changed position, the blob's characteristics also changed (e.g., the distribution that describes the blob became tighter, i.e., easier to estimate its center). According to the framework presented earlier, the brain's belief about target position could not change instantaneously to the new position reported by the sensory system. Rather, the rate at which the belief would converge to this new position should depend on the uncertainty of the prior belief, controlled by the characteristics of the 1st target. For example, if the first target had low

uncertainty but the second target has medium uncertainty, the rate of change in hand position should be slow. However, if the first target had high uncertainty, the rate of change in hand position toward the same second target should be fast.

One can describe the estimation problem with the generative model shown in Fig. 4.16B. The states that we wish to estimate include position of the target, and position and velocity of the handheld LED, all represented by vector $\mathbf{x}$. Our observation $\mathbf{y}$ is a noisy, time delayed version of these states. With our motor commands $\mathbf{u}$ to the arm we manipulate the position of the handheld LED. The state update equation becomes:

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} + B\mathbf{u}^{(k)} + \varepsilon^{(k)} \tag{4.52}$$

To represent the fact that our observation is a delayed version of the states, a simple technique is to extend the state vector by its copies. Suppose that an event at time step *k* was sensed by the first stage of sensory processing and then transmitted to the next stage at time $k+\delta$, where in this case $\delta = 10$ ms. This information continues to propagate and becomes 'observable' with a delay of Δ. Thus, we can represent the delay by first extending the state vector $\mathbf{x}_e^{(k)} = \left[\mathbf{x}_0^{(k)}, \mathbf{x}_\delta^{(k)}, \mathbf{x}_{2\delta}^{(k)}, \cdots, \mathbf{x}_\Delta^{(k)}\right]^T$ and then allowing the system to observe only the most delayed state $\mathbf{x}_\Delta^{(k)}$. The sparse matrix *C* takes care of this last step:

$$\begin{aligned} \mathbf{x}_e^{(k+1)} &= A_e\mathbf{x}_e^{(k)} + B\mathbf{u}^{(k)} + \varepsilon_e^{(k)} \\ \mathbf{y}^{(k)} &= C\mathbf{x}_e^{(k)} + \omega^{(k)} \end{aligned} \tag{4.53}$$

The basic prediction of this model is that when our sensory system reports that some Δ time ago the target moved to a new position, our estimate of that position does not change immediately to this new position. Rather, our belief converges to the observed position with a rate that depends on our uncertainty about our current estimate. A simulation that shows this idea is provided in Fig. 4.16C. When the 1st target is displayed, uncertainty about its position rapidly declines. The rate of decline is fastest for the target with the smallest variance (blob variance is S, M, or L). If we set an arbitrary threshold on our uncertainty, we can see that crossing this threshold is earlier for the small variance target, potentially explaining the faster reaction time. As the movement unfolds, the target jumps to a new location. If the 2nd target has a larger variance, uncertainty will increase after a delay period Δ (for example, condition S-M or M-L). Let us compare the S-M condition with the L-M condition. Reaction to the target jump is most influenced by the prior uncertainty at the time of the target jump. The S-M condition has a low uncertainty when one senses the target jump, and therefore a small Kalman gain. The L-M condition has a high

uncertainty and a larger Kalman gain. The large Kalman gain will produce a rapid reaction to the information regarding the changed target position. This is shown in the simulated hand acceleration traces in Fig. 4.16D. Indeed, when the 1st target position had a small uncertainty and the 2nd target had medium uncertainty (S-M condition, Fig. 4.16D typical subject), people corrected the handheld LED's trajectory gradually toward the 2nd target. However, if the 1st target's uncertainty was large, reaction to that same medium uncertainty 2nd target (L-M condition, Fig. 4.16D) was quite strong: people corrected the hand path vigorously.

4.9 Hiking in the woods in an estimation framework

Let us now return to the problem of combining multiple sources of information about a hidden variable--the hiking in the woods problem--and recast it in the Kalman framework. Our objective is to show that the Kalman gain is precisely the weights in Eq. (4.9) that we earlier assigned to the two devices. That is, when our only source of information is from our observation, and we have no prior beliefs, then the Kalman gain is the maximum likelihood estimate.

A graphical representation of the generative model for the hiking in the woods problem is shown in Fig. 4.17. The hidden variable $\mathbf{x}$ is observed via two independent measurements $\mathbf{y}_a$ and $\mathbf{y}_b$. Our generative model takes the form:

$$\begin{aligned} \mathbf{x}^{(n+1)} &= A\mathbf{x}^{(n)} + \boldsymbol{\varepsilon}_x^{(n)} \qquad \boldsymbol{\varepsilon}_x \sim N\left(\mathbf{0}, Q\right) \\ \mathbf{y}^{(n)} &= C\mathbf{x}^{(n)} + \boldsymbol{\varepsilon}_y^{(n)} \qquad \boldsymbol{\varepsilon}_y \sim N\left(\mathbf{0}, R\right) \end{aligned} \tag{4.54}$$

We begin with a prior estimate $\mathbf{x}^{(1|0)}$ and our uncertainty $P^{(1|0)}$. The general form of the Kalman gain and posterior uncertainty is:

$$\begin{aligned} \mathbf{k}^{(n)} &= P^{(n|n-1)}C^T\left(CP^{(n|n-1)}C^T + R\right)^{-1} \\ P^{(n|n)} &= \left(I - \mathbf{k}^{(n)}C\right)P^{(n|n-1)} \end{aligned} \tag{4.55}$$

Suppose that we have no idea where we are, i.e., $P^{(1|0)} = \infty$. In this case, we cannot compute the Kalman gain from Eq. (4.55). However, we can proceed by first computing the posterior estimate (second line in Eq. 4.55) in terms of the prior estimate, and then rewrite the Kalman gain (first line in Eq. 4.55) in terms of the posterior rather than prior uncertainty. We begin by expressing the posterior uncertainty in terms of the prior:

$$\begin{aligned} P^{(n|n)} &= P^{(n|n-1)} - \mathbf{k}^{(n)} C P^{(n|n-1)} \\ &= P^{(n|n-1)} - P^{(n|n-1)} C^T \left(C P^{(n|n-1)} C^T + R \right)^{-1} C P^{(n|n-1)} \end{aligned} \tag{4.56}$$

We can simplify the second line of Eq. (4.56) by using the matrix inversion lemma. This lemma states that:

$$\left(Z - X Y^{-1} X^T \right)^{-1} = Z^{-1} + Z^{-1} X \left(Y - X^T Z^{-1} X \right)^{-1} X^T Z^{-1} \tag{4.57}$$

Let us set $-Z^{-1} = P^{(n|n-1)}$, $X = C^T$, and $Y = R$. We can rewrite the second line in Eq. (4.56) as:

$$P^{(n|n)} = -Z^{-1} - Z^{-1} X \left(Y - X^T Z^{-1} X \right)^{-1} X^T Z^{-1} \tag{4.58}$$

Therefore we have:

$$-P^{(n|n)} = \left(Z - X Y^{-1} X^T \right)^{-1} = \left(-\left(P^{(n|n-1)} \right)^{-1} - C^T R^{-1} C \right)^{-1}$$

which we can simplify to:

$$\left(P^{(n|n)} \right)^{-1} = \left(P^{(n|n-1)} \right)^{-1} + C^T R^{-1} C \tag{4.59}$$

Eq. (4.59) explains that if our prior uncertainty is infinite, then the inverse of our posterior uncertainty is simply $C^T R^{-1} C$. Now, let us express the Kalman gain in terms of the posterior uncertainty. We begin by multiplying both sides of the first line in Eq. 4.55 by the term in the parenthesis:

$$\mathbf{k}^{(n)} \left(C P^{(n|n-1)} C^T + R \right) = P^{(n|n-1)} C^T \tag{4.60}$$

We next multiply both sides of the above expression by R^{-1} and then after a little rearrangement arrive at:

$$\begin{aligned} \mathbf{k}^{(n)} &= P^{(n|n-1)} C^T R^{-1} - \mathbf{k}^{(n)} C P^{(n|n-1)} C^T R^{-1} \\ &= \left(P^{(n|n-1)} - \mathbf{k}^{(n)} C P^{(n|n-1)} \right) C^T R^{-1} \end{aligned} \tag{4.61}$$

The term in the parenthesis in the above expression is simply the posterior uncertainty, allowing us to express the Kalman gain for the system of Eq. (4.54) in this way:

$$\mathbf{k}^{(n)} = P^{(n|n)} C^T R^{-1}. \tag{4.62}$$

If we insert our prior $P^{(1|0)} = \infty$ into Eq. (4.59), our posterior becomes

$$P^{(1|1)} = \left(C^T R^{-1} C\right)^{-1} \tag{4.63}$$

The Kalman gain becomes:

$$\mathbf{k}^{(1)} = \left(C^T R^{-1} C\right)^{-1} C^T R^{-1} \tag{4.64}$$

Our estimate of the hidden variable (our position) is:

$$\mathbf{x}^{(1|1)} = \mathbf{x}^{(1|0)} + \mathbf{k}^{(1)}\left(\mathbf{y}^{(1)} - C\mathbf{x}^{(1|0)}\right) \tag{4.65}$$

Assuming that our prior $\mathbf{x}^{(1|0)}$ was zero, the above expression reduces to:

$$\mathbf{x}^{(1|1)} = \left(C^T R^{-1} C\right)^{-1} C^T R^{-1} \mathbf{y}^{(1)} \tag{4.66}$$

This expression is our maximum likelihood estimate in Eq. (4.8). Furthermore, the variance of our estimate, as expressed in Eq. (4.63), is the variance of our maximum likelihood estimate in Eq. (4.11). Therefore, if we are naïve in the sense that we have no prior knowledge about the state that we wish to estimate, then a weighted combination of the two sources of information is the optimal solution. On the other hand, if we also have a prior, e.g., we have hiked this path before and have some idea of where we might be, then the Kalman framework gives us the tools to weigh in this additional piece of information.

From a practical point of view, a useful result that we derived in this section is with regard to the prior uncertainty. Often when we try to estimate something we are at a loss as to what the prior uncertainty should be. If we are completely naïve, then Eq. (4.63) tells us what the uncertainty will be after the first data point is observed. A common technique is to set the prior uncertainty to this value.

4.10 Signal dependent noise

Thus far we have considered processes that have additive Gaussian noise. With this kind of noise, the variance of the variable that we are observing is independent of its mean. If you think about it, this is a bit odd: the signal varies by some amount and the variance is the same whether the signal is small or large. Biological systems seem to have a different kind of noise: variance varies with the size of the signal. For example, Richard Schmidt and colleagues (Schmidt et al., 1979) asked volunteers to make rapid reaching movements to a target and measured the

variability of the endpoint (Fig. 4.18). They found that for a given target distance, the smaller durations (i.e., faster movements) were associated with larger endpoint variance. As movement duration decreased, the force required to make that movement would of course increase. Richard Schmidt hypothesized that the noise in a muscle was likely dependent on the force developed by the muscle (Schmidt, 1991). He wrote: "Movement's inaccuracy increases as movement time decreases, primarily because of the increased noise involved in the stronger muscle contractions." That is, noise associated with the motor commands was likely an increasing function of the magnitude of those commands.

We can examine the noise properties of muscles in an experiment in which volunteers produced a force by pushing with their thumb on a transducer while viewing the resulting force on a video monitor. Kelvin Jones, Antonia Hamilton, and Daniel Wolpert (2002) measured this force and focused on the standard deviation of this variable during a 4 sec period in which the visual feedback was eliminated (shown in the second column of Fig. 4.19A). They found that the standard deviation of the force grew roughly linearly as a function of mean force (Fig. 4.19B). The change in the standard deviation of force may have been because of the transduction of neural signal into force in the muscle, or from the neural signal itself. To resolve this question, Jones et al. electrically stimulated the thumb flexor muscle (shown in the third and fourth column of Fig. 4.19A). They did not observe an increase in variance with increased stimulation strength. This result suggested that the noise originated from the neural motor-commands, and was not due to the muscle and how it produced force.

The term *signal dependent noise* refers to a process in which the standard deviation of the noise depends on the mean of the signal. For example, the force produced by muscle i may be related to its input u_i via a noise process that has the following form:

$$f_i = u_i\left(1 + c_i\phi_i\right) \qquad \phi_i \square N\left(0,1\right) \tag{4.67}$$

The term c_i indicates the rate at which the noise grows with the signal. That is, the variance of force increases as a function of the signal u_i:

$$\text{var}\left[f_i\right] = c_i^2 u_i^2 \tag{4.68}$$

And so the standard deviation of force increases linearly with u_i, with slope c_i.

How would we estimate the state of a system if it suffers from signal dependent noise? The Kalman framework that we have been using will need to be modified. Suppose that we have a system in which we produce motor commands $\mathbf{u}$ (e.g., force in the muscles), and this affects the state of our system $\mathbf{x}$ (e.g., position, velocity, of our body), resulting in sensory feedback $\mathbf{y}$ (e.g., proprioception, vision). Suppose that the system is of the form:

$$\begin{aligned}\mathbf{x}^{(k+1)} &= A\mathbf{x}^{(k)} + B\left(\mathbf{u}^{(k)} + \boldsymbol{\varepsilon}_u^{(k)}\right) + \boldsymbol{\varepsilon}_x^{(k)} \\ \mathbf{y}^{(k)} &= H\left(\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_s^{(k)}\right) + \boldsymbol{\varepsilon}_y^{(k)}\end{aligned} \tag{4.69}$$

where $\boldsymbol{\varepsilon}_x$ and $\boldsymbol{\varepsilon}_y$ are zero mean Gaussian noise:

$$\begin{aligned}\boldsymbol{\varepsilon}_x &\sim N\left(\mathbf{0}, Q_x\right) \\ \boldsymbol{\varepsilon}_y &\sim N\left(\mathbf{0}, Q_y\right)\end{aligned} \tag{4.70}$$

and $\boldsymbol{\varepsilon}_u$ and $\boldsymbol{\varepsilon}_s$ are zero mean signal dependent noise terms, meaning that noise depends on the motor commands $\mathbf{u}$, and state $\mathbf{x}$, respectively:

$$\boldsymbol{\varepsilon}_u^{(k)} \equiv \begin{bmatrix} c_1 u_1^{(k)} \phi_1^{(k)} \\ c_2 u_2^{(k)} \phi_2^{(k)} \\ \vdots \\ c_n u_n^{(k)} \phi_n^{(k)} \end{bmatrix} \quad \boldsymbol{\varepsilon}_s^{(k)} \equiv \begin{bmatrix} d_1 x_1^{(k)} \mu_1^{(k)} \\ d_2 x_2^{(k)} \mu_2^{(k)} \\ \vdots \\ d_m x_m^{(k)} \mu_m^{(k)} \end{bmatrix}$$
$$\phi \sim N(0,1) \qquad \mu \sim N(0,1) \tag{4.71}$$

The signal dependent motor noise $\boldsymbol{\varepsilon}_u$ affects the state $\mathbf{x}$ and the signal dependent sensory noise $\boldsymbol{\varepsilon}_s$ affects the observation $\mathbf{y}$. It is useful to express the signal dependent noise terms as a linear function of $\mathbf{u}$ and $\mathbf{x}$. To do so, we can define:

$$C_1 \equiv \begin{bmatrix} c_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \ddots \end{bmatrix} \quad C_2 \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & \ddots \end{bmatrix}$$
$$D_1 \equiv \begin{bmatrix} d_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \ddots \end{bmatrix} \quad D_2 \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & \ddots \end{bmatrix} \tag{4.72}$$

and so we have:

$$\begin{aligned}\boldsymbol{\varepsilon}_u^{(k)} &= \sum_{i=1}^{m} C_i \mathbf{u} \phi_i^{(k)} \\ \boldsymbol{\varepsilon}_s^{(k)} &= \sum_{i=1}^{n} D_i \mathbf{x} \mu_i^{(k)}\end{aligned} \tag{4.73}$$

In Eq. (4.73), m is the size of the vector $\mathbf{u}$ and n is the size of the vector $\mathbf{x}$. And so we can re-write the system equations as:

$$\begin{aligned}\mathbf{x}^{(k+1)} &= A\mathbf{x}^{(k)} + B\mathbf{u}^{(k)} + \boldsymbol{\varepsilon}_x^{(k)} + B\sum_i C_i\mathbf{u}^{(k)}\phi_i^{(k)} \\ \mathbf{y}^{(k)} &= H\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_y^{(k)} + H\sum_i D_i\mathbf{x}^{(k)}\mu_i^{(k)}\end{aligned} \tag{4.74}$$

On trial k, we have a prior belief $\hat{\mathbf{x}}^{(k|k-1)}$, with uncertainty $P^{(k|k-1)}$, and we make an observation $\mathbf{y}^{(k)}$. To update our belief, we have:

$$\hat{\mathbf{x}}^{(k|k)} = \hat{\mathbf{x}}^{(k|k-1)} + K^{(k)}\left(\mathbf{y}^{(k)} - H\hat{\mathbf{x}}^{(k|k-1)}\right) \tag{4.75}$$

How should we set $K^{(k)}$? As before, we will set it in such a way as to minimize the trace of the posterior uncertainty $P^{(k|k)}$. However, when we do this, we will see that the Kalman gain will now depend on both the state $\mathbf{x}$ and the motor commands $\mathbf{u}$.

Re-writing Eq. (4.75), we have:

$$\begin{aligned}\hat{\mathbf{x}}^{(k|k)} &= \hat{\mathbf{x}}^{(k|k-1)} + K^{(k)}\left(H\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_y^{(k)} + H\sum_i D_i\mathbf{x}^{(k)}\mu_i^{(k)} - H\hat{\mathbf{x}}^{(k|k-1)}\right) \\ &= \left(I - K^{(k)}H\right)\hat{\mathbf{x}}^{(k|k-1)} + K^{(k)}\left(H\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_y^{(k)} + H\sum_i D_i\mathbf{x}^{(k)}\mu_i^{(k)}\right)\end{aligned} \tag{4.76}$$

The variance of our posterior estimate is:

$$\begin{aligned}P^{(k|k)} = &\left(I - K^{(k)}H\right)P^{(k|k-1)}\left(I - K^{(k)}H\right)^T + K^{(k)}Q_yK^{(k)T} \\ &+\sum_i KHD_i\mathbf{x}^{(k)}\mathbf{x}^{(k)T}D_i^TH^TK^T\end{aligned} \tag{4.77}$$

Eq. (4.77) can be written as:

$$\begin{aligned}P^{(k|k)} = &P^{k|k-1} - 2P^{k|k-1}H^TK^{(k)T} \\ &+K^{(k)}\left(HP^{k|k-1}H^T + Q_y + \sum_i HD_i\mathbf{x}^{(k)}\mathbf{x}^{(k)T}D_i^TH^T\right)K^{(k)T}\end{aligned} \tag{4.78}$$

The derivative of the trace of the above equation with respect to $K^{(k)}$ is:

$$\begin{aligned}\frac{d}{dK^{(k)}}tr\left[P^{(k|k)}\right] = &-2P^{(k|k-1)}H^T \\ &+2K^{(k)}\left(HP^{(k|k-1)}H^T + Q_y + \sum_i HD_i\mathbf{x}^{(k)}\mathbf{x}^{(k)T}D_i^TH^T\right)\end{aligned} \tag{4.79}$$

Setting the above expression to zero and solving for $K^{(k)}$ we have:

$$K^{(k)} = P^{(k|k-1)}H^T\left(HP^{(k|k-1)}H^T + Q_y + \sum_i HD_i\mathbf{x}^{(k)}\mathbf{x}^{(k)T}D_i^TH^T\right)^{-1} \tag{4.80}$$

Notice that because of signal dependent noise, the Kalman gain is a function of the state $\mathbf{x}$. (In practice, we replace the term $\mathbf{x}^{(k)}\mathbf{x}^{(k)T}$ in Eq. (4.80) with our estimate $E\left[\mathbf{x}^{(k)}\mathbf{x}^{(k)T}\right]$). In fact, when we compute the uncertainties, we see that the Kalman gain is also a function of the motor commands. Substitute the expression in Eq. (4.80) in Eq. (4.78), and we have:

$$P^{k|k} = P^{k|k-1}\left(I - H^T K^{(k)T}\right) \tag{4.81}$$

The prior uncertainty in step $k+1$ becomes:

$$P^{(k+1|k)} = AP^{(k|k)}A^T + Q_x + \sum_i BC_i\mathbf{u}^{(k)}\mathbf{u}^{(k)T}C_i^T B^T \tag{4.82}$$

Therefore, the state uncertainty increases with the size of the motor commands, and the Kalman gain decreases with the size of the state.

An implication of our derivation is that if we are pushing a large mass (producing relatively large motor commands), then we will have a larger uncertainty regarding the consequences of these commands (as compared to pushing a small mass with a smaller amount of force). As a result, when we are producing large forces we should rely more on the sensory system and our observations and less on our predictions.

Summary

As our brain plans and generates motor commands, it also predicts the sensory consequences and acts on the predicted consequences. One clear advantage of making sensory predictions is that the brain does not have to wait for the sensory measurements to know that it needs to intervene. The delay in sensory measurements is long enough that it can cause stability problems. Relying on predictions, rather than delayed measurements, allows one to over-come this delay. A second advantage of making sensory predictions is that the brain can combine its predictions with the sensory measurements (when they arrive). The combination of the two streams of information would allow the brain to sense the world better than if it only had the sensory information stream. That is, we are able to 'see' the stimulus better because we have two sources of information about it, rather than just one. These two streams also allow the brain to separate the sensory data that is a result of self-generated motion from data that is due to external events. If the process of state estimation is damaged due to disease, then the result may be symptoms of delusions, i.e., an inability to assign agency.

To combine two streams of information, one needs to apply a weight to each stream. In principle, the weight should be higher for the more reliable information source. State estimation theory provides a framework in which to describe this problem. In this framework, the objective is to estimate the state of our body or the world around us from our sensory measurements. This problem has two components. The first is associated with learning to accurately predict the future, i.e., what the sensory measurements should be – this is our *prior* estimate of state. The second is associated with optimally combining the measured quantities with the predicted one to form a *posterior* estimate of state. Kalman first solved this problem by finding a posterior estimate that for a linear system with Gaussian noise produced a posterior estimate that had the minimum variance.

Figure 4.1. Subject was instructed to pull on a knob that was fixed on a rigid wall. **A**) EMG recordings from arm and leg muscles. Before biceps is activated, the brain activates the leg muscles to stabilize the lower body and prevent sway due to the anticipated pulling force on the upper body. **B**) When a rigid bar is placed on the upper body, the leg muscles are not activated when biceps is activated. (From (Cordo and Nashner, 1982) with permission).

Figure 4.2. Effect of eye movement on the memory of a visual stimulus. In the top panel, the filled circle represents the fixation point, the asterisk indicates the location of the visual stimulus, and the dashed circle indicates the receptive field a cell in the LIP region of the parietal cortex. **A**) Discharge to the onset and offset of a visual stimulus in the cell's receptive field. Abbreviations: H. eye, horizontal eye position; Stim, stimulus; V. eye, vertical eye position. **B**) Discharge during the time period in which a saccade brings the stimulus into the cell's receptive field. The cell's discharge increased before the saccade brought the stimulus into the cell's receptive field. (From (Duhamel et al., 1992) with permission).

Figure 4.3. Subject reached to the location of a visual stimulus. **A**) In the Control condition, subject fixated the target stimulus for a brief period of time, the target disappeared, and after a delay period a reach was made. In the Static condition, subject fixated a secondary stimulus when the target stimulus appeared to one side. In the Remap condition, the subject fixated the target stimulus, then made a saccade to a secondary stimulus (at which point the target stimulus was erased). In the Combined condition, the subject fixated the target stimulus, made a saccade to a secondary stimulus, and then during the delay period was again shown the target stimulus after completion of the saccade. **B**) The standard deviation of the reach, i.e., a measure of uncertainty about the location of the target stimulus. Optimal refers to a weighted combination of static and remap conditions. (From (Vaziri et al., 2006) with permission).

Figure 4.4. Motor commands change the states of our body and the environment around us. These states are transduced by our sensory system and become the measured sensory consequences of our motor commands. As we generate motor commands, our brain also predicts the sensory consequences via an internal model that is called a 'forward model'. By combining the predicted and measured sensory consequences, we form a belief about the states of our body and the environment. This belief is a reflection of both our predictions and our observations.

Figure 4.5. Subject looked at a moving cursor while a group of dots appeared on the screen for 300ms. In some trials the dots would remain still (**A**) while in other trials they would move together left or right with a constant speed (**B**). Subject indicated the direction of motion of the dots. From this result, the authors estimated the speed of subjective stationarity, i.e., the speed of dots for which the subject perceived them to be stationary. **C**) The unfilled circles represent performance of control subjects. Regardless of the speed of the cursor, they perceived the dots to be stationary only if their speed was near zero. The filled triangles represent performance of subject RW. As the speed of the cursor increased, RW perceived the dots to be stationary if their speed was near the speed of the cursor. (From (Haarmeier et al., 1997) with permission).

Figure 4.6. Disorders of agency in schizophrenia relate to an inability to compensate for sensory consequences of self-generated motor commands. In a paradigm similar to that shown in Fig. 4.5, volunteers estimated whether during motion of a cursor the background moved to the right or left. By varying the background speed, at each cursor speed the experimenters estimated the speed of perceptual stationarity, i.e., the speed of background motion for which the subject saw the background to be stationary. They then computed a compensation index as the difference between speed of eye movement and speed of background when perceived to be stationary, divided by speed of eye movement. The subset of schizophrenic patients who had delusional symptoms showed a greater deficit than control in their ability to compensate for sensory consequences of self-generated motor commands. (From (Lindner et al., 2005) with permission.)

Figure 4.7. Response to a visual stimulus depends on both the noise in the stimulus and the brain's predictions about that stimulus. **A**) In a reaching task, feedback about the current position of the finger was provided for only a brief time during the movement. As the finger moved from the starting circle, cursor (indicating current finger position) was extinguished. Halfway to the target, feedback was briefly provided. The position of this feedback was displaced from actual finger position by a random amount, with a mean of 1 cm. The quality of the feedback was controlled: the feedback was either clear (σ_0), or with different degrees of blur (σ_M and σ_L), or withheld (σ_∞). The paths illustrate typical trajectories for a displacement of 2 cm. **B**) Top sub-plot: For 1000 trials, subjects trained with the illustrated distribution of lateral shifts, i.e., a Gaussian with a mean of 1 cm. This constitutes the prior probability of the displacement. Middle sub-plot: A diagram of various probability distributions associated with the current measurement. This distribution is shown for the clear and the two blurred feedback conditions for a trial in

which the true shift was 2 cm. Bottom sub-plot: The estimate of displacement for an optimal observer that combines the prior with the evidence. **C**) The lateral deviation of the cursor at the end of the reach as a function of the imposed lateral shift for a typical subject. The horizontal line at 0 would indicate full compensation to the observed error. The dash line would indicate complete denial of the observed error. The solid line is the Bayesian model with the level of uncertainty fitted to the data. For example, when the feedback was clear, i.e., σ_0, the subject compensated almost fully, nearly hitting the target in all conditions. When the feedback was uncertain, i.e., σ_L, the subject missed the target in most conditions. **D**) The inferred priors for each subject and condition. The true distribution is shown in part B. (From (Kording and Wolpert, 2004) with permission.)

Figure 4.8. Device *a* and device *b* provide independent estimates of a hidden variable (position on a map). Each device has a Gaussian noise property with a covariance $R_a = [1,-1;-1,3]$ and $R_b = [1,1;1,3]$. The ellipses describe the region centered on the mean of the distribution that contains 10%, 25%, and 50% of the data under the distribution. The maximum likelihood estimate of the hidden variable is marked by the distribution at the center.

Figure 4.9. Maximum likelihood integration of two sensory modalities. Visually and haptically specified heights of an object differ by Δ. On the left columns, the visual and haptic variances are equal. The mean of the combined probability density is equal to the mean of the visual and haptic densities. The variance of the combined density is half of the visual (or haptic) density. If the judgment of relative height is based on the combined density, the psychometric function is the cumulative Gaussian (bottom left) with the point of subjective equality (PSE) equal to the average of the visual and haptic heights. In the right column of figures, the noise in the haptic sensor is four times the visual noise. The psychometric function is shifted so that the PSE is closer to the visual height. (From (Ernst and Banks, 2002) with permission.)

Figure 4.10. Procedure for determining the noise in the haptic sensory modality. **A**. Two objects that differ in size by amount $\Delta = 3$ are presented and a subject holds each object to estimate its size. The estimate of size for each object, $y_1 \square N(\mu_1, \sigma_h^2)$ and $y_2 \square N(\mu_1 + \Delta, \sigma_h^2)$ are shown (here, we assumed $\sigma_h^2 = 1$). **B.** The estimate of the difference between the two objects is

$\hat{\Delta} = y_2 - y_1$ and this random variable is a Gaussian with the following distribution: $\hat{\Delta} \sim N\left(\Delta, 2\sigma_h^2\right)$. The probability of picking the second object as being larger, $\Pr\left(\hat{\Delta} > 0\right)$, is the integral of $p\left(\hat{\Delta}\right)$ from zero to infinity. **C.** The probability $\Pr\left(\hat{\Delta} > 0\right)$ rises faster as a function of Δ when sensory noise σ_h^2 is small. When $\Delta = \sqrt{2}\sigma_h$, $\Pr\left(\hat{\Delta} > 0\right) = 0.84$, as indicated by the dashed line.

Figure 4.11. Experimental data from an experiment on haptic and visual sensory integration. **A.** Within-modality experimental results. Proportion of trials in which the second stimulus was perceived as taller than the standard stimulus is plotted against the height of the second stimulus. Four noise levels were considered for the visual condition. **B.** Visual-haptic discrimination. **C.** Predicted and experimental weights and PSEs. The shaded area represented predicted weights from within-modality discrimination. The height of the shaded area represents predicted errors given the standard errors of within modality discrimination. **D.** Combined and within modality discrimination thresholds. Thresholds are from the psychometric function in parts A and B. Dashed line represents haptic-alone threshold. (From (Ernst and Banks, 2002) with permission.)

Figure 4.12. Samples from three Gaussian distributions.

Figure 4.13. Distribution of parameter **w** in a linear regression problem. In all three cases the mean value of the distribution is [0.5, 0.5], but the covariance of the distribution depends on the specific data that were used to form the estimate.

Figure 4.14. Estimation of a hidden variable via Kalman filter. In this example, the hidden state **w** is assumed to be invariant. The open circles indicate hidden variables. The gray circles indicate observed variables.

Figure 4.15. Estimation of a hidden variable via Kalman filter. In this example, the hidden state **w** is assumed to change from trial to trial.

Figure 4.16. The rate at which belief converges onto the measured value depends on the uncertainty of the prior belief. **A**) Subjects reached to a blob-like target, with visual characteristics that was described by a Gaussian of small, medium, or large variance. The

objective was to place the handheld LED in the center of the target. **B**) A generative model describing the relationship between input (motor commands), state (position of the target and hand), and observation (delayed measurement of state). **C**) Model predictions. Uncertainty of the state of the target as a function of time. Movement starts earlier for a small variance target. When the target jumps, the prior uncertainty is specified by the 1st target and uncertainty converges to the variance of the 2nd target. **D**) Example data. When the target jumped, the hand path was corrected to the new target. However, for a given second target (medium uncertainty here), the rate of correction was largest when the 1st target had the largest uncertainty. (From (Izawa and Shadmehr, 2008) with permission.)

Figure 4.17. A graphical representation of the generative model for the hiking in the woods problem. In this problem, we have two sensors $\mathbf{y}_a$ and $\mathbf{y}_b$ that provide us with independent measures of a hidden state $\mathbf{x}$.

Figure 4.18. Endpoint variability of rapid aiming movements (shown as standard deviation of the endpoint position) as a function of movement duration and amplitude. Reduced duration increases endpoint variability. (Data from (Schmidt et al., 1979)).

Figure 4.19. The standard deviation of noise grows with mean force in an isometric task. Participants produced a given force with their thumb flexors. In one condition (labeled "voluntary"), the participants generated the force, whereas in another condition (labeled "NMES") the experimenters stimulated the muscles artificially to produce force. To guide force production, the participants viewed a cursor that displayed thumb force, but the experimenters analyzed the data during a 4-s period in which this feedback had disappeared. **A**. Force produced by a typical participant. The period without visual feedback is marked by the horizontal bar in the 1st and 3rd columns (top right) and is expanded in the 2nd and 4th columns. **B**. When participants generated force, noise (measured as the standard deviation) increased linearly with force magnitude (with a slope of ~1. Abbreviations: NMES, neuromuscular electrical stimulation; MVC, maximum voluntary contraction. (From Jones et al. (2002) with permission.)

Reference List

Cordo PJ, Nashner LM (1982) Properties of postural adjustments associated with rapid arm movements. J Neurophysiol 47:287-302.

Duhamel JR, Colby CL, Goldberg ME (1992) The updating of the representation of visual space in parietal cortex by intended eye movements. Science 255:90-92.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Frith C (1996) Neuropsychology of schizophrenia, what are the implications of intellectual and experiential abnormalities for the neurobiology of schizophrenia? Br Med Bull 52:618-626.

Frith CD, Blakemore S, Wolpert DM (2000) Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. Brain Res Brain Res Rev 31:357-363.

Gilbert D (2006) Stumbling on happiness. New York: Vintage Books.

Haarmeier T, Thier P, Repnow M, Petersen D (1997) False perception of motion in a patient who cannot compensate for eye movements. Nature 389:849-852.

Izawa J, Shadmehr R (2008) Online processing of uncertain information in visuomotor control. J Neurosci 28:11360-11368.

Jones KE, Hamilton AF, Wolpert DM (2002) Sources of signal-dependent noise during isometric force production. J Neurophysiol 88:1533-1544.

Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Engineering 82 (series D):35-45.

Kording KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. Nature 427:244-247.

Lindner A, Thier P, Kircher TT, Haarmeier T, Leube DT (2005) Disorders of agency in schizophrenia correlate with an inability to compensate for the sensory consequences of actions. Curr Biol 15:1119-1124.

Schmidt RA (1991) Motor learning and performance: from principles to practice. Champaign, IL: Human Kinetics Books.

Schmidt RA, Zelaznik H, Hawkins B, Frank JS, Quinn JT, Jr. (1979) Motor-output variability: a theory for the accuracy of rapid motor acts. Psychol Rev 47:415-451.

Vaziri S, Diedrichsen J, Shadmehr R (2006) Why does the brain predict sensory consequences of oculomotor commands? Optimal integration of the predicted and the actual sensory feedback. J Neurosci 26:4188-4197.