5 Uncertainty

As we had discussed at the beginning of the course, stochastic signals arise in three contexts: *uncertainty*, *noise*, and *randomness*. Uncertainty refers to lack of precise knowledge about some situation of interest, and it is often convenient to compensate for this lack using probabilistic models; noise refers to stochastic manifestations of some underlying physical causes; and, finally, randomness describes situations where we delibrately introduce stochastic effects into some otherwise deterministic system. In this lecture, we will take a closer look at uncertainty.

5.1 Conditioning: information-driven revision of probabilities

One of the most fundamental operations in probability theory is *conditioning*. The basics of conditional probability were covered in ECE 313. As a reminder, let A and B be two events on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where $\mathbf{P}[B] > 0$. Suppose that we are told that B has occurred; then the *conditional probability* of A given B is defined as

$$\mathbf{P}[A|B] \triangleq \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}. \tag{5.1}$$

This simple definition has a powerful interpretation: Recall that a generic element ω of the sample space Ω represents complete knowledge of some phenomenon of interest, which cannot be observed directly. Instead, we can ask "yes or no" questions of the form "is ω an element of A?" — or, equivalently, "has event A occurred?" The probability $\mathbf{P}[A]$ of A occurring can be interpreted in different ways: One is to imagine a long sequence of experiments, where the ith experiment generates an independent ω_i, and for each i write down whether A has occurred. Thus, if we perform many such experiments, and if each experiment is representative of the underlying phenomenon, then it is reasonable to take $\mathbf{P}[A]$ as the proportion of the experiments in which A has occurred. This is known as the *frequency interpretation* of probability. On the other hand, the number $\mathbf{P}[A]$ can come from other sources — for example, an expert in the type of phenomena described by $(\Omega, \mathcal{F}, \mathbf{P})$ can provide you with some information that you can use to arrive at $\mathbf{P}[A]$. In this case, there may not be any repeated experiments, and it is more proper to think about $\mathbf{P}[A]$ as your *degree of belief* in the occurrence of A based on your *a priori* knowledge (naturally, 0 corresponds to "impossible" and 1 to "certain to occur"). This is the *Bayesian interpretation* of probability. The term "Bayesian" is associated with the name of Rev. Thomas Bayes (for whom the Bayes' rule is named as well).

In either case, $\mathbf{P}[A]$ encapsulates our uncertainty about A based only on a priori knowledge (that comes either from objective experimentation or from a combination of objective and subjective judgments). Now, when we are told that B has occurred, this *new information* (or *evidence*) causes us to revise our assessment of the probability of A. Indeed, whereas before we knew nothing about the actual realized ω, aftere we were told that B has occurred, we know that A can only occur if ω belongs to both A and B. The probability of this is, by definition, given by $\mathbf{P}[A \cap B]$, i.e., the numerator on the right-hand side of (5.1), and $\mathbf{P}[B]$ in the denominator is just the normalizing

factor. Similarly, if we were told that B did not occur, then we would revise $\mathbf{P}[A]$ to

$$\mathbf{P}[A|\bar{B}] = \frac{\mathbf{P}[A \cap \bar{B}]}{\mathbf{P}[\bar{B}]},$$

where $\bar{B}$ denotes nonoccurrence of B, so $\mathbf{P}[\bar{B}] = 1 - \mathbf{P}[B]$. We can also flip this situation and compute the conditional probablity of B given that A has occurred:

$$\mathbf{P}[B|A] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[A]}.$$

Note that, as far as the computation of conditional probabilities goes, A and B play the same role, and in fact $\mathbf{P}[A]\mathbf{P}[B|A] = \mathbf{P}[B]\mathbf{P}[A|B]$. However, in many problems involving probabilistic uncertainty, some conditional probabilities, say $\mathbf{P}[B|A]$ and $\mathbf{P}[B|\bar{A}]$, may be known beforehand — this is the conditional probability of seeing evidence B given that situation A has or has not occurred. Then the conditional probability formula (5.1) is a computational device for *probabilistic inversion*: if we are told that B has occurred, we update the prior probability $\mathbf{P}[A]$ to the posterior probability

$$\mathbf{P}[A|B] = \frac{\mathbf{P}[B|A]\mathbf{P}[A]}{\mathbf{P}[B|A]\mathbf{P}[A] + \mathbf{P}[B|\bar{A}]\mathbf{P}[\bar{A}]}.$$

Note that all the quantities on the right-hand side are specified at the outset as part of our model. Similarly, if we are told that B did not occur, then we update $\mathbf{P}[A]$ to

$$\mathbf{P}[A|\bar{B}] = \frac{\mathbf{P}[\bar{B}|A]}{\mathbf{P}[\bar{B}|A]\mathbf{P}[A] + \mathbf{P}[\bar{B}|\bar{A}]\mathbf{P}[\bar{A}]}.$$

Again, all the quantities on the right-hand side are intrinsic to our model.

We can express this updating in terms of two binary-valued random variables X, Y on the common probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where X takes the value 1 if A has occurred and 0 otherwise, and Y takes the value 1 if B has occurred and 0 otherwise. We are interested in computing $\mathbf{P}[X = 1|Y = y]$ for $y \in \{0, 1\}$, where the interpretation is as follows: if B occurs, then $Y = 1$, and we update $\mathbf{P}[X = \cdot] \to \mathbf{P}[X = \cdot|Y = 1]$; if B does not occur, then $Y = 0$, and we update $\mathbf{P}[X = \cdot] \to \mathbf{P}[X = \cdot|Y = 0]$. To see how $\mathbf{P}[A] = \mathbf{P}[X = 1]$ and $\mathbf{P}[B|A] = \mathbf{P}[Y = 1|X = 1]$ can arise from a model, suppose that $Y = f(X, U)$ for some function f and for some random variable U independent of X. Then

$$\begin{aligned}\mathbf{P}[X = x, Y = y] &= \mathbf{P}[X = x, f(x, U) = y] \\ &= \mathbf{P}[X = x]\mathbf{P}[f(x, U) = y].\end{aligned}$$

Here, $\mathbf{P}[X = x]$ is the a priori probability that $X = x$, and $\ell(x, y) \triangleq \mathbf{P}[f(x, U) = y]$ is called the *likelihood* of observing $Y = y$ when $X = x$. Since $\mathbf{P}[Y = y|X = x] = \frac{\mathbf{P}[X=x, Y=y]}{\mathbf{P}[X=x]}$, we see that $\ell(x, y) = \mathbf{P}[Y = y|X = x]$. Let $\pi_x \triangleq \mathbf{P}[X = x]$. Then the posterior probabilities of $X = 0$ and $X = 1$ given the observation Y are actually random variables, given by

$$\mathbf{P}[X = 0|Y] = \frac{\ell(Y, 0)\pi_0}{\ell(Y, 0)\pi_0 + \ell(Y, 1)\pi_1}.$$

Note that, again, the right-hand side involves the likelihood ℓ and the prior probabilities π_0 and π_1 that are part of our model.

5.2 A glimpse of Bayesian filtering

5.2.1 Evolution of beliefs as a stochastic signal

We will start with a simple example. Suppose that someone hands you a coin and tells you that this coin is either "heavy" (i.e., the probability of HEADS is $\frac{1+\theta}{2}$) or "light" (i.e., the probability of HEADS is $\frac{1-\theta}{2}$), where the parameter $\theta \in (0,1)$ is fixed and known. How can you find out whether the coin is heavy or light? Suppose it's heavy. If we toss it n times, then there should be approximately $\frac{n(1+\theta)}{2}$ tosses when the coin came up HEADS. By the same token, if the coin is light, then we should expect to see approximately $\frac{n(1-\theta)}{2}$ HEADS in a sequence of n tosses. Of course, we don't know whether the coin is heavy or light, but we hope that we can find out by repeatedly tossing it. Moreover, we expect each toss to provide additional evidence about the coin's weight. We will now see how to formalize this intuition using Bayesian updating, and we will also see that we can represent the evolution of our belief that the coin is heavy based on the evidence so far as a stochastic signal.

According to the Bayesian viewpoint, we represent the true weight of the coin by a random variable X taking values in the set $\{0,1\}$. We interpret the event $\{X=0\}$ as the coin being light, and $\{X=1\}$ as the coin being heavy. In the language of Bayesian inference, the probability distribution of X is called the *prior distribution* (or simply the prior), meaning that these are the probabilities we assign to the weight of the coin prior to seeing any evidence from tossing the coin. Let $p = \mathbf{P}[X=1]$. If we have complete ignorance about the weight of the coin, it makes sense to take $p = \frac{1}{2}$, but there may be good reasons to take $p \neq \frac{1}{2}$. For example, if this is not your first time encountering such a situation, and the coin was heavy in, say, 60% of all of the previous occurrences, it makes sense to take $p = 0.6$. Now we toss the coin repeatdely, which generates a stochastic signal $Y = (Y_t)_{t\in\mathbb{N}}$ with each Y_t taking values in $\{0,1\}$. In order to make inferences, we need to first describe the joint distribution of X and Y.

We assume that, given the weight of the coin (or *conditioned* on the value of X), the outcomes of the tosses are independent of one another. That is, we can write

$$Y = (Y_t)_{t\in\mathbb{N}} \text{ are } \begin{cases} \text{i.i.d. Bern}(\frac{1+\theta}{2}), & \text{if } X=1 \\ \text{i.i.d. Bern}(\frac{1-\theta}{2}), & \text{if } X=0 \end{cases}.$$

More succinctly, we can write

$$X \sim \text{Bern}(p) \text{ and } Y = (Y_t)_{t\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \text{Bern}\left(\frac{1-(-1)^X\theta}{2}\right). \tag{5.2}$$

Eq. (5.2) shows that X and Y are dependent random objects. In order to write down their joint distribution, we introduce an imperative model. Specifically, let $U = (U_t)_{t\in\mathbb{N}}$ be a sequence of i.i.d. Unif$(0,1)$ random variables, which are also independent of X. Then we introduce a function f, such that

$$Y_t = f(X, U_t), \qquad t = 1, 2, \ldots.$$

Based on (5.2), this function takes the following form:

$$f(x,u) \triangleq \begin{cases} 0, & \text{if } 0 \le u < \frac{1+(-1)^x\theta}{2} \\ 1, & \text{if } \frac{1+(-1)^x\theta}{2} \le u < 1 \end{cases} \tag{5.3}$$

(exercise: verify that this choice of f delivers what we want). Now we can write down the joint distribution of X and $Y^t \triangleq (Y_1, \ldots, Y_t)$, the tuple of the outcomes of the first t tosses, for every t. Indeed, since the random variables $X, U_1, \ldots, U_t$ are independent, we have

$$\begin{aligned}\mathbf{P}[X = x, Y^t = y^t] &= \mathbf{P}[X = x, f(x, U_1) = y_1, \ldots, f(x, U_t) = y_t] \\ &= \mathbf{P}[X = x] \prod_{s=1}^{t} \mathbf{P}[f(x, U_s) = y_s]. \end{aligned} \tag{5.4}$$

It is not hard to show from (5.3) that, for $U \sim \text{Unif}(0,1)$ and for any $x, y \in \{0,1\}$,

$$\mathbf{P}[f(x,U) = y] = \left(\frac{1 + (-1)^x \theta}{2}\right)^{1-y} \left(\frac{1 - (-1)^x \theta}{2}\right)^{y}$$

Substituting this into (5.4), we get

$$\mathbf{P}[X = x, Y^t = y^t] = \mathbf{P}[X = x] \cdot \prod_{s=1}^{t} \left(\frac{1 + (-1)^x \theta}{2}\right)^{1-y_s} \left(\frac{1 - (-1)^x \theta}{2}\right)^{y_s}.$$

The conditional probability of seeing the outcomes $Y^t = y^t$ given $X = x$ is

$$\mathbf{P}[Y^t = y^t | X = x] = \prod_{s=1}^{t} \left(\frac{1 + (-1)^x \theta}{2}\right)^{1-y_s} \left(\frac{1 - (-1)^x \theta}{2}\right)^{y_s}. \tag{5.5}$$

In particular,

$$\mathbf{P}[Y^t = y^t | X = 1] = \prod_{s=1}^{t} \left(\frac{1 - \theta}{2}\right)^{1-y_s} \left(\frac{1 + \theta}{2}\right)^{y_s},$$

which simply reflects our modeling assumption that, conditioned on the coin being heavy, the outcomes of different tosses are i.i.d. Bernoulli random variables with bias $\frac{1+\theta}{2}$. Moreover, we also see that, for any t,

$$\mathbf{P}[Y_t = y_t | X = x] = \mathbf{P}[f(x, U_t) = y_t] = \left(\frac{1 + (-1)^x \theta}{2}\right)^{1-y_t} \left(\frac{1 - (-1)^x \theta}{2}\right)^{y_t}.$$

In the language of Bayesian inference, the conditional probability $\mathbf{P}[Y^t = y^t | X = x]$ is called the *likelihood* of observing the *evidence* $Y^t = y^t$ conditioned on $X = x$. We can now use Bayes' rule to compute the *posterior probability distribution* (or simply the posterior) of X given the evidence $Y^t = y^t$:

$$\mathbf{P}[X = x | Y^t = y^t] = \frac{\mathbf{P}[Y^t = y^t | X = x] \mathbf{P}[X = x]}{\mathbf{P}[Y^t = y^t]}.$$

Here, $\mathbf{P}[Y^t = y^t]$ is the probability of seeing the evidence $Y^t = y^t$, averaged over all possible values of X, and we can use (5.5) to evaluate it explicitly. To that end, let us define the quantity

$$\lambda \triangleq \frac{1+\theta}{1-\theta}.$$

Then, using (5.5), we get

$$\begin{aligned}
\mathbf{P}[Y^t = y^t] &= \mathbf{P}[X = 0, Y^t = y^t] + \mathbf{P}[X = 1, Y^t = y^t] \\
&= \mathbf{P}[Y^t = y^t | X = 0]\mathbf{P}[X = 0] + \mathbf{P}[Y^t = y^t | X = 1]\mathbf{P}[X = 1] \\
&= (1-p)\prod_{s=1}^{t}\left(\frac{1+\theta}{2}\right)^{1-y_s}\left(\frac{1-\theta}{2}\right)^{y_s} + p\prod_{s=1}^{t}\left(\frac{1-\theta}{2}\right)^{1-y_s}\left(\frac{1+\theta}{2}\right)^{y_s} \\
&= (1-p)\left(\frac{(1+\theta)}{2}\right)^t \prod_{s=1}^{t}\lambda^{-y_s} + p\left(\frac{(1-\theta)}{2}\right)^t \prod_{s=1}^{t}\lambda^{y_s}. \qquad (5.6)
\end{aligned}$$

Let $Z_t \triangleq Y_1 + \ldots + Y_t$ be the total number of HEADS in t tosses. Using this definition in (5.6), we obtain

$$\mathbf{P}[Y^t = y^t] = (1-p)\left(\frac{(1+\theta)}{2}\right)^t \lambda^{-z_t} + p\left(\frac{(1-\theta)}{2}\right)^t \lambda^{z_t}. \qquad (5.7)$$

Since X takes only two values, 0 or 1, it suffices to determine only the posterior probability $\mathbf{P}[X = 1 | Y^t = y^t]$ of the coin being heavy. In that case, we have

$$\begin{aligned}
\mathbf{P}[Y^t = y^t | X = 1] &= \prod_{s=1}^{t}\left(\frac{1-\theta}{2}\right)^{1-y_s}\left(\frac{1+\theta}{2}\right)^{1+y_s} \\
&= \left(\frac{1-\theta}{2}\right)^t \lambda^{z_t},
\end{aligned}$$

so, using this and (5.7), we get

$$\mathbf{P}[X = 1 | Y^t = y^t] = \frac{p(1-\theta)^t \lambda^{z_t}}{(1-p)(1+\theta)^t \lambda^{-z_t} + p(1-\theta)^t \lambda^{z_t}}. \qquad (5.8)$$

Let us take a moment to analyze the expression (5.8) in detail. The first thing we see is that the posterior probability of $X = 1$ (i.e., the coin being heavy) given $Y^t = y^t$ does not depend on the actual sequence of the first t tosses, but only on the number of HEADS z_t. Secondly, since $z_{t+1} = z_t + y_{t+1}$, we have

$$\begin{aligned}
\mathbf{P}[X = 1 | Y^{t+1} = y^{t+1}] &= \frac{p(1-\theta)^{t+1}\lambda^{z_{t+1}}}{p(1-\theta)^{t+1}\lambda^{z_{t+1}} + (1-p)(1+\theta)^{t+1}\lambda^{-z_t}} \\
&= \frac{(1-\theta)\lambda^{y_{t+1}} \cdot p(1-\theta)^t \lambda^{z_t}}{(1-\theta)\lambda^{y_{t+1}} \cdot p(1-\theta)^t \lambda^{z_t} + (1+\theta)\cdot(1-p)(1+\theta)^t \lambda^{-z_t}}. \qquad (5.9)
\end{aligned}$$

Let D_t denote the denominator of (5.8). Then we can rewrite (5.9) as follows:

$$\begin{aligned}\mathbf{P}[X=1|Y^{t+1}=y^{t+1}] &= \frac{(1-\theta)\lambda^{y_{t+1}}\cdot D_t\mathbf{P}[Y^t=y^t|X=x]}{(1-\theta)\lambda^{y_{t+1}}\cdot D_t\mathbf{P}[Y^t=y^t|X=x]+(1+\theta)\lambda^{-y_{t+1}}\cdot D_t(1-\mathbf{P}[Y^t=y^t|X=x])}\\ &= \frac{\mathbf{P}[Y^t=y^t|X=x](1-\theta)\lambda^{y_{t+1}}}{\mathbf{P}[Y^t=y^t|X=x](1-\theta)\lambda^{y_{t+1}}+\cdot(1-\mathbf{P}[Y^t=y^t|X=x])(1+\theta)\lambda^{-y_{t+1}}}. \qquad (5.10)\end{aligned}$$

In other words, when a new piece of evidence (i.e., Y_{t+1}) arrives, and we wish to update the posterior probability of $X=1$, we do not need to keep track of the past evidence Y^t (or even of the HEADS count Z_t), but only of the most recent posterior $\mathbf{P}[X=1|Y^t=y^t]$ — we see that $\mathbf{P}[X=1|Y^{t+1}=y^{t+1}]$ is a function of $\mathbf{P}[X=1|Y^t=y^t]$ and y_{t+1} only. What is more, let us consider the case of a single toss of the coin which heavy with probability p and light with probability $1-p$. In that case, the posterior probability of the coin being heavy is given by (5.8) with $t=1$:

$$\begin{aligned}\mathbf{P}[X=1|Y_1=y_1] &= \frac{p(1-\theta)\lambda^{y_1}}{(1-p)(1+\theta)\lambda^{-y_1}+p(1-\theta)\lambda^{y_1}}\\ &\triangleq g(p,y_1).\end{aligned}$$

Looking back at (5.10), we see that

$$\mathbf{P}[X=1|Y^{t+1}=y^{t+1}] = g\Big(\mathbf{P}[X=1|Y^t=y^t],y_{t+1}\Big),$$

which has the same functional form.

We can summarize all of this as follows: We start with a coin with prior probability p of being heavy and repeatedly toss it. Let $Y_1,Y_2,\ldots$ be the outcomes of the tosses. Define the stochastic signal $\Pi=(\Pi_t)_{t\in\mathbb{Z}_+}$ by $\Pi_0=p$ and

$$\begin{aligned}\Pi_t &= g(\Pi_{t-1},Y_t)\\ &\equiv \frac{\Pi_{t-1}(1-\theta)\lambda^{Y_t}}{\Pi_{t-1}(1-\theta)\lambda^{Y_t}+(1-\Pi_{t-1})(1+\theta)\lambda^{-Y_t}}. \qquad (5.11)\end{aligned}$$

Each Π_t takes values in the unit interval $[0,1]$, and represents our belief, based on all the evidence available at time t, that the coin is heavy. For this reason, we refer to Π as the *belief signal* (or the *belief process*). We stress that this is a stochastic signal, since each Π_t depends on X (which is unobservable) and on the evidence Y^t. Notice the nice recursive form of (5.11): the belief at time t is a *known deterministic function* of the belief at time $t-1$ and the new observation Y_t. For this reason, we can view Π_t as a state variable.

Let's get some idea about how this belief process evolves. Suppose that we have computed the

belief Π_{t-1}, and toss the coin again. If $Y_t = 1$ (HEADS), then (5.11) gives

$$\begin{aligned}\Pi_t &= \frac{\Pi_{t-1}(1-\theta)\lambda}{\Pi_{t-1}(1-\theta)\lambda + (1-\Pi_{t-1})(1+\theta)\lambda^{-1}} \\ &= \frac{\Pi_{t-1}(1+\theta)}{\Pi_{t-1}(1+\theta) + (1-\Pi_{t-1})(1-\theta)} \\ &= \frac{\Pi_{t-1}(1+\theta)}{2\theta\Pi_{t-1} + 1 - \theta} \\ &= \frac{1+\theta}{1+(2\Pi_{t-1}-1)\theta} \cdot \Pi_{t-1}.\end{aligned}$$

On the other hand, if $Y_t = 0$ (TAILS), then a similar calculation gives

$$\begin{aligned}\Pi_t &= \frac{\Pi_{t-1}(1-\theta)}{\Pi_{t-1}(1-\theta) + (1-\Pi_{t-1})(1+\theta)} \\ &= \frac{1-\theta}{1-(2\Pi_{t-1}-1)\theta} \cdot \Pi_{t-1}.\end{aligned}$$

5.2.2 Hidden Markov Models

Our coin tossing example had to do with a situation when there was only one unobservable quantity (X), and we were updating our beliefs about it based on a stream of observations $Y_0, Y_1, \ldots$. We now consider a more complicated scenario, where we have to track an unobservable stochastic signal $X = (X_t)_{t\in\mathbb{Z}_+}$ on the basis of another stochastic signal $Y = (Y_t)_{t\in\mathbb{N}}$, where, at each time $t \in \mathbb{N}$, we obtain a new observation Y_t, which pertains to X_t. To keep things simple, we assume that the X_t's and the Y_t's take values in some discrete spaces. What do we mean by tracking? At time t, we have observations $Y_1^t = (Y_1, \ldots, Y_t)$, and we wish to compute the posterior probability distribution

$$\Pi_t(\cdot) \triangleq \mathbf{P}[X_t = \cdot | Y_1^t = y_1^t].$$

Conceptually, Π_t quantifies our belief regarding the value of X_t based on the evidence Y_1^t. We take Y_1^0 to be an empty tuple, and in that case $\Pi_0 \equiv p_0$. The process of updating Π_t as new observations arrive is called *Bayesian filtering* or *stochastic filtering*.

As before, in order for Π_t to be well-defined, we first need to specify the joint distribution of X and Y starting from a suitable imperative description. Let p_0 denote the probability distribution of X_0. Let $U = (U_t)_{t\in\mathbb{N}}$ and $V = (V_t)_{t\in\mathbb{N}}$ be two i.i.d. stochastic signals, which are assumed to be independent of each other and of the initial condition X_0. The distribution of U_1 may be different from the distribution of V_1. We also assume that there are two functions, f and g, such that

$$X_{t+1} = f(X_t, U_{t+1}) \tag{5.12a}$$

$$Y_t = g(X_t, V_t) \tag{5.12b}$$

for every $t \in \mathbb{N}$. This type of set-up is called a *Hidden Markov Model* (or HMM), for the following reasons. First of all, from (5.12a) and from the fact that X_0 is independent of U, it follows that X is a Markov chain. Let M denote the matrix of its transition probabilities, i.e.,

$$M(x, x') = \mathbf{P}[f(x, U_1) = x'].$$

We also see that each Y_t is a function only of X_t and of V_t, and so can be thought of as a noisy observation of X_t. This explains the term HMM: the Markov chain X is hidden from direct observation, and we can only see a noisy version of it.

We first need to record a few observations regarding certain conditional probabilities pertaining to (5.12). For any t, we have

$$\begin{aligned} Y_t &= g(X_t, V_t) \\ &= g\left(f(X_{t-1}, U_t), V_t\right) \\ &= g\left(f\left(f(X_{t-2}, U_{t-1}), U_t\right), V_t\right) \\ &= \dots, \end{aligned}$$

which means that Y_1^t is a deterministic function of X_0, U_1^t, and V_1^t. On the other hand, X_t is a deterministic function of X_0 and U_1^t. Therefore, since X_0, (U_1^t, V_1^t), and U_{t+1} are mutually independent, we write

$$\begin{aligned} &\mathbf{P}\Big[\underbrace{X_t = x_t, Y_1^t = y_1^t}_{\text{fctn. of } X_0, U_1^t, V_1^t}, \underbrace{X_{t+1} = x_{t+1}}_{\text{fctn. of } x_t, U_{t+1}}\Big] \\ &\quad = \mathbf{P}[X_t = x_t, Y_1^t = y_1^t]\mathbf{P}[f(x_t, U_{t+1}) = x_{t+1}] \\ &\quad = \mathbf{P}[X_t = x_t, Y_1^t = y_1^t]M(x_t, x_{t+1}). \end{aligned}$$

That is,

$$\mathbf{P}[X_t = x_t, X_{t+1} = x_{t+1}, Y_1^t = y_1^t] = \mathbf{P}[X_t = x_t, Y_1^t = y_1^t]M(x_t, x_{t+1}). \tag{5.13}$$

Reasoning in the same manner, we also have

$$\begin{aligned} &\mathbf{P}\Big[\underbrace{Y_1^t = y_1^t, X_{t+1} = x_{t+1}}_{\text{fctn. of } X_0, U_1^{t+1}, V_1^t}, \underbrace{Y_{t+1} = y_{t+1}}_{\text{fctn. of } x_{t+1}, V_{t+1}}\Big] \\ &\quad = \mathbf{P}[Y_1^t = y_1^t, X_{t+1} = x_{t+1}]\mathbf{P}[g(x_{t+1}, V_{t+1}) = y_{t+1}] \\ &\quad = \mathbf{P}[Y_1^t = y_1^t, X_{t+1} = x_{t+1}]T(x_{t+1}, y_{t+1}). \end{aligned}$$

That is,

$$\mathbf{P}[Y_1^t = y_1^t, X_{t+1} = x_{t+1}, Y_{t+1} = y_{t+1}] = \mathbf{P}[Y_1^t = y_1^t, X^{t+1} = x_{t+1}]T(x_{t+1}, y_{t+1}). \tag{5.14}$$

What we are about to prove is actually quite remarkable: there exists a function h, such that

$$\Pi_{t+1} = h(\Pi_t, Y_{t+1}) \tag{5.15}$$

for all $t \in \mathbb{Z}_+$, with the initial condition $\Pi_0 = p_0$. In fact, we will characterize this function explicitly. Eq. (5.15) is called the *filtering recursion*. To proceed, it will be convenient to define the following notation: for any s, t,

$$\Pi_{t|s} \triangleq \mathbf{P}[X_t = \cdot | Y_1^s = y_1^s].$$

Note that $\Pi_{t|s}$ is a probability distribution on the state space X, and it also depends on the observations $Y_1^s = y_1^s$. We suppress this dependence to minimize notational clutter, but you should always keep it in mind. With this definition, we have $\Pi_t \equiv \Pi_{t|t}$. We will show that the update from $\Pi_t = \Pi_{t|t}$ to $\Pi_{t+1} = \Pi_{t+1|t+1}$ can be realized as a composition of two operations:

$$\Pi_{t|t} \xrightarrow{\text{prediction}} \Pi_{t+1|t} \xrightarrow{\text{correction}} \Pi_{t+1|t+1}.$$

In the prediction step, we compute the belief about the *next* state X_{t+1} before the new observation Y_{t+1} arrives, solely on the basis of what we know at time t; then, once Y_{t+1} is released, we correct our prediction. In order to derive these two steps, we will exploit the imperative model (5.12).

Fix an arbitrary state $a \in \mathsf{X}$. Using the definition of conditional probability, we write

$$\begin{aligned}\Pi_{t+1|t}(a) &= \mathbf{P}[X_{t+1} = a | Y_1^t = y_1^t] \\ &= \frac{\mathbf{P}[X_{t+1} = a, Y_1^t = y_1^t]}{\mathbf{P}[Y_1^t = y_1^t]}.\end{aligned} \tag{5.16}$$

To compute the numerator, we apply the law of total probability to all possible values of X_t and then use (5.13):

$$\begin{aligned}\mathbf{P}[X_{t+1} = a, Y_1^t = y_1^t] &= \sum_{b \in \mathsf{X}} \mathbf{P}[X_{t+1} = a, X_t = b, Y_1^t = y_1^t] \\ &= \sum_{b \in \mathsf{X}} \mathbf{P}[X_t = b, Y_1^t = y_1^t] M(b,a) \\ &= \sum_{b \in \mathsf{X}} \mathbf{P}[Y_1^t = y_1^t] \mathbf{P}[X_t = b | Y_1^t = y_1^t] M(b,a) \\ &= \mathbf{P}[Y_1^t = y_1^t] \sum_{b \in \mathsf{X}} \underbrace{\mathbf{P}[X_t = b | Y_1^t = y_1^t]}_{= \Pi_{t|t}(b)} M(b,a).\end{aligned}$$

Substituting this into (5.16), we see that $\mathbf{P}[Y_1^t = y_1^t]$ cancels, and we end up with

$$\Pi_{t+1|t}(a) = \sum_{b \in \mathsf{X}} \Pi_{t|t}(b) M(b,a). \tag{5.17}$$

We recognize this as the one-step update equation for the distribution of the state of a Markov chain:

$$\Pi_{t+1|t} = \Pi_t M. \tag{5.18}$$

This is the *prediction step* of the filtering recursion. Now we write down the correction step:

$$\begin{aligned}\Pi_{t+1|t+1}(a) &= \mathbf{P}[X_{t+1} = a | Y_1^{t+1} = y_1^{t+1}] \\ &= \frac{\mathbf{P}[Y_1^t = y_1^t, X_{t+1} = a, Y_{t+1} = y_{t+1}]}{\mathbf{P}[Y_1^t = y_1^t, Y_{t+1} = y_{t+1}]}.\end{aligned} \tag{5.19}$$

Using (5.14) and the definition of $\Pi_{t+1|t}$, we can express the numerator as

$$\begin{aligned}\mathbf{P}[Y_1^t = y_1^t, X_{t+1} = a, Y_{t+1} = y_{t+1}] &= \mathbf{P}[Y_1^t = y_1^t, X_{t+1} = a]T(a, y_{t+1}) \\ &= \mathbf{P}[Y_1^t = y_1^t]\mathbf{P}[X_{t+1} = a | Y_1^t = y_1^t]T(a, y_{t+1}) \\ &= \mathbf{P}[Y_1^t = y_1^t]\Pi_{t+1|t}(a)T(a, y_{t+1}).\end{aligned}$$

Using the law of total probability, we can also express the denominator as

$$\begin{aligned}\mathbf{P}[Y_1^t = y_1^t, Y_{t+1} = y_{t+1}] &= \sum_{b \in \mathsf{X}} \mathbf{P}[Y_1^t = y_1^t, X_{t+1} = b, Y_{t+1} = y_{t+1}] \\ &= \sum_{b \in \mathsf{X}} \mathbf{P}[Y_1^t = y_1^t]\Pi_{t+1|t}(b)T(b, y_{t+1}).\end{aligned}$$

Substituting these expressions into (5.19), we obtain

$$\begin{aligned}\Pi_{t+1}(a) &= \Pi_{t+1|t+1}(a) \\ &= \frac{\Pi_{t+1|t}(a)T(a, y_{t+1})}{\sum_{b \in \mathsf{X}} \Pi_{t+1|t}(b)T(b, y_{t+1})}.\end{aligned} \tag{5.20}$$

This is the *correction step* of the filtering recursion. All in all, we have just proved (5.15).