5 Uncertainty

5.1 A glimpse of Bayesian filtering

5.1.1 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 $p = P[X = 1]$. 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 repeatedly, 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

$$X \sim \text{Bern}(p) \text{ and } Y = (Y_t)_{t \in \mathbb{N}} \text{ i.i.d. } \sim \text{ Bern}\left(\frac{1-(1)^X\theta}{2}\right). \quad (5.1)$$

Eq. (5.1) 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

$$Y_t = f(X, U_t), \quad t = 1, 2, \ldots.$$
Based on (5.1), this function takes the following form:

\[
f(x, u) \triangleq \begin{cases} 
0, & \text{if } 0 \leq u < \frac{1 - (-1)^x \theta}{2} \\
1, & \text{if } \frac{1 - (-1)^x \theta}{2} \leq u < 1
\end{cases}
\]  

(5.2)

(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

\[
P[X = x, Y^t = y^t] = P[X = x, f(x, U_1) = y_1, \ldots, f(x, U_t) = y_t] \\
= P[X = x] \prod_{s=1}^t P[f(x, U_t) = y_t].
\]  

(5.3)

It is not hard to show from (5.2) that, for \( U \sim \text{Unif}(0, 1) \) and for any \( x, y \in \{0, 1\} \),

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

Substituting this into (5.3), we get

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

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

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

(5.4)

In particular,

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

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 \( t \),

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

In the language of Bayesian inference, the conditional probability \( P[Y^t = y^t | X = x] \) 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 $Y^t = y^t$:

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

Here, $P[Y^t = y^t]$ is the probability of seeing the evidence $Y^t = y^t$, averaged over all possible values of $X$, and we can use (5.4) to evaluate it explicitly. To that end, let us define the quantity

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

Then, using (5.4), we get

$$P[Y^t = y^t] = P[X = 0, Y^t = y^t] + P[X = 1, Y^t = y^t]$$ 

$$= P[Y^t = y^t | X = 0] P[X = 0] + P[Y^t = y^t | X = 1] P[X = 1]$$ 

$$= (1 - p) \prod_{s=1}^{t} \left( \frac{1 + \theta}{2} \right)^{1-y_s} \left( \frac{1 - \theta}{2} \right)^{y_s} + p \prod_{s=1}^{t} \left( \frac{1 - \theta}{2} \right)^{1-y_s} \left( \frac{1 + \theta}{2} \right)^{y_s}$$ 

$$= (1 - p) \left( \frac{1 + \theta}{2} \right)^t \lambda^{-z_t} + p \left( \frac{1 - \theta}{2} \right)^t \lambda^{z_t}. \quad (5.5)$$

Let $Z_t \triangleq Y_1 + \ldots + Y_t$ be the total number of heads in $t$ tosses. Using this definition in (5.5), we obtain

$$P[Y^t = y^t] = (1 - p) \left( \frac{1 + \theta}{2} \right)^t \lambda^{-z_t} + p \left( \frac{1 - \theta}{2} \right)^t \lambda^{z_t}. \quad (5.6)$$

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

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

$$= \left( \frac{1 - \theta}{2} \right)^t \lambda^{z_t},$$

so, using this and (5.6), we get

$$P[X = 1 | Y^t = y^t] = \frac{p(1 - \theta)^t \lambda^{z_t}}{(1 - p)(1 + \theta)^t \lambda^{-z_t} + p(1 - \theta)^t \lambda^{z_t}}. \quad (5.7)$$

Let us take a moment to analyze the expression (5.7) 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

$$P[X = 1|Y^{t+1} = y^{t+1}] = \frac{p(1-\theta)^{t} \lambda^{z_{t+1}}}{p(1-\theta)^{t+1} \lambda^{z_{t+1}} + (1-p)(1+\theta)^{t+1} \lambda^{-z_{t}}}$$

$$= \frac{(1-\theta)\lambda^{y_{t+1}} \cdot p(1-\theta)^{t} \lambda^{z_{t}}}{(1-\theta)\lambda^{y_{t+1}} \cdot p(1-\theta)^{t} \lambda^{z_{t}} + (1-\theta)(1+\theta)\lambda^{-z_{t}}}.$$  \hspace{1cm} (5.8)

Let $D_t$ denote the denominator of (5.7). Then we can rewrite (5.8) as follows:

$$P[X = 1|Y^{t+1} = y^{t+1}] = \frac{(1-\theta)\lambda^{y_{t+1}} \cdot D_t P[Y^t = y^t|X = x]}{(1-\theta)\lambda^{y_{t+1}} \cdot D_t P[Y^t = y^t|X = x] + (1+\theta)\lambda^{-y_{t+1}} \cdot D_t (1 - P[Y^t = y^t|X = x])}$$

$$= \frac{P[Y^t = y^t|X = x]|(1-\theta)\lambda^{y_{t+1}} + (1-\theta)(1+\theta)\lambda^{-y_{t+1}}}{1}.$$  \hspace{1cm} (5.9)

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 $P[X = 1|Y^{t+1} = y^{t+1}]$ — we see that $P[X = 1|Y^{t+1} = y^{t+1}]$ is a function of $P[X = 1|Y^t = y^t]$ 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 (5.7) with $t = 1$:

$$P[X = 1|Y_1 = y_1] = \frac{p(1-\theta)\lambda^{y_1}}{(1-p)(1+\theta)\lambda^{-y_1} + p(1-\theta)\lambda^{y_1}}$$

$$= g(p, y_1).$$

Looking back at (5.9), we see that

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

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

$$\Pi_t = g(\Pi_{t-1}, Y_t)$$

$$\equiv \frac{\Pi_{t-1}(1-\theta)\lambda^{Y_t}}{\Pi_{t-1}(1-\theta)\lambda^{Y_t} + (1-\Pi_{t-1})(1+\theta)\lambda^{-Y_t}}.$$  \hspace{1cm} (5.10)

Each $\Pi_t$ 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 $\Pi$ as the belief signal (or the belief process). We stress that this is a stochastic signal, since each $\Pi_t$ depends on $X$ (which is unobservable) and on the evidence $Y^t$. Notice the nice recursive form of (5.10): 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 $\Pi_t$ as a state variable.
Let’s get some idea about how this belief process evolves. Suppose that we have computed the belief $\Pi_{t-1}$, and toss the coin again. If $Y_t = 1$ (Heads), then (5.10) gives

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

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

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

5.1.2 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 $Y_t = (Y_1, \ldots, Y_t)$, and we wish to compute the posterior probability distribution

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

Conceptually, $\Pi_t$ 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 $\Pi_t$ as new observations arrive is called Bayesian filtering or stochastic filtering.

As before, in order for $\Pi_t$ 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

$$
X_{t+1} = f(X_t, U_{t+1}) \quad (5.11a) \\
Y_t = g(X_t, V_t) \quad (5.11b)
$$
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 (5.11a) 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.,

\[
M(x, x') = P[f(x, U_1) = x'].
\]

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 (5.11). For any \( t \), we have

\[
Y_t = g(X_t, V_t)
\]

\[
= g(f(X_{t-1}, U_t), V_t)
\]

\[
= g(f(f(X_{t-2}, U_{t-1}), U_t), V_t)
\]

\[
= \ldots,
\]

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

\[
P\left[ X_t = x_t, Y_1^t = y_1^t, X_{t+1} = x_{t+1} \right]
\]

\[
\text{fctn. of } X_0, U_t^t, V_t^t \text{ fctn. of } x_t, U_{t+1}
\]

\[
= P[X_t = x_t, Y_1^t = y_1^t]P[f(x_t, U_{t+1}) = x_{t+1}]
\]

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

That is,

\[
P[X_t = x_t, X_{t+1} = x_{t+1}, Y_1^t = y_1^t] = P[X_t = x_t, Y_1^t = y_1^t]M(x_t, x_{t+1}). \tag{5.12}
\]

Reasoning in the same manner, we also have

\[
P\left[ Y_1^t = y_1^t, X_{t+1} = x_{t+1}, Y_{t+1} = y_{t+1} \right]
\]

\[
\text{fctn. of } X_0, U_t^{t+1}, V_t^t \text{ fctn. of } x_{t+1}, V_{t+1}
\]

\[
= P[Y_1^t = y_1^t, X_{t+1} = x_{t+1}]P[g(x_{t+1}, V_{t+1}) = y_{t+1}]
\]

\[
= P[Y_1^t = y_1^t, X_{t+1} = x_{t+1}]T(x_{t+1}, y_{t+1}).
\]

That is,

\[
P[Y_1^t = y_1^t, X_{t+1} = x_{t+1}, Y_{t+1} = y_{t+1}] = P[Y_1^t = y_1^t, X_{t+1} = x_{t+1}]T(x_{t+1}, y_{t+1}). \tag{5.13}
\]

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

\[
\Pi_{t+1} = h(\Pi_t, Y_{t+1}) \tag{5.14}
\]
for all $t \in \mathbb{Z}_+$, with the initial condition $\Pi_0 = p_0$. In fact, we will characterize this function explicitly. Eq. (5.14) is called the filtering recursion. To proceed, it will be convenient to define the following notation: for any $s, t$,

$$
\Pi_{t|s} \triangleq P[X_t = |Y_s^s = y_s^s].
$$

Note that $\Pi_{t|s}$ is a probability distribution on the state space $X$, and it also depends on the observations $Y_s^s = y_s^s$. We suppress this dependence to minimize notational clutter, but you should always keep it in mind. With this definition, we have $\Pi_t = \Pi_{t|t}$. 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 (5.11).

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

$$
\Pi_{t+1|t}(a) = P[X_{t+1} = a | Y^1_t = y^1_t]
$$

$$
= \frac{P[X_{t+1} = a, Y^1_t = y^1_t]}{P[Y^1_t = y^1_t]}.
$$

(5.15)

To compute the numerator, we apply the law of total probability to all possible values of $X_t$ and then use (5.12):

$$
P[X_{t+1} = a, Y^1_t = y^1_t] = \sum_{b \in X} P[X_{t+1} = a, X_t = b, Y^1_t = y^1_t]
$$

$$
= \sum_{b \in X} P[X_t = b, Y^1_t = y^1_t]M(b, a)
$$

$$
= \sum_{b \in X} P[Y^1_t = y^1_t]P[X_t = b | Y^1_t = y^1_t]M(b, a)
$$

$$
= P[Y^1_t = y^1_t] \sum_{b \in X} P[X_t = b | Y^1_t = y^1_t]M(b, a)_{\Pi_{t|t}(b)}.
$$

Substituting this into (5.15), we see that $P[Y^1_t = y^1_t]$ cancels, and we end up with

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

(5.16)

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

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

(5.17)
This is the prediction step of the filtering recursion. Now we write down the correction step:

\[
\Pi_{t+1|t+1}(a) = \mathbb{P}[X_{t+1} = a|Y_{t+1} = y_{t+1}]
= \frac{\mathbb{P}[Y_{t+1} = y_{t+1}, X_{t+1} = a, Y_{t+1} = y_{t+1}]}{\mathbb{P}[Y_{t+1} = y_{t+1}, Y_{t+1} = y_{t+1}]}.
\] (5.18)

Using (5.13) and the definition of \(\Pi_{t+1|t}\), we can express the numerator as

\[
\mathbb{P}[Y_{t+1} = y_{t+1}, X_{t+1} = a, Y_{t+1} = y_{t+1}] = \mathbb{P}[Y_{t+1} = y_{t+1}, X_{t+1} = a]T(a, y_{t+1})
= \mathbb{P}[Y_{t+1} = y_{t+1}]\mathbb{P}[X_{t+1} = a|Y_{t+1} = y_{t+1}]T(a, y_{t+1})
= \mathbb{P}[Y_{t+1} = y_{t+1}] \Pi_{t+1|t}(a)T(a, y_{t+1}).
\]

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

\[
\mathbb{P}[Y_{t+1} = y_{t+1}, Y_{t+1} = y_{t+1}] = \sum_{b \in \mathcal{X}} \mathbb{P}[Y_{t+1} = y_{t+1}, X_{t+1} = b, Y_{t+1} = y_{t+1}]
= \sum_{b \in \mathcal{X}} \mathbb{P}[Y_{t+1} = y_{t+1}] \Pi_{t+1|t}(b)T(b, y_{t+1}).
\]

Substituting these expressions into (5.18), we obtain

\[
\Pi_{t+1}(a) = \Pi_{t+1|t+1}(a)
= \frac{\Pi_{t+1|t}(a)T(a, y_{t+1})}{\sum_{b \in \mathcal{X}} \mathbb{P}[Y_{t+1} = y_{t+1}] \Pi_{t+1|t}(b)T(b, y_{t+1}).}
\] (5.19)

This is the correction step of the filtering recursion. All in all, we have just proved (5.14).