State Space Model

Richard Xu

November 29, 2022

1 Topic Summary

1.1 What is time series?

Many definition exist, but let's go with this one: a well-defined collection of observations of data items obtained by repeated measurements over time. Can you give some examples of time series?

2 Dynamic/State space models

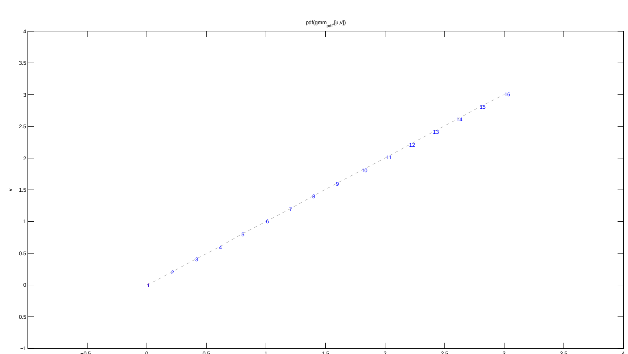
Here we discuss a special kind of continuous dynamic system model/state space model. We'll discuss the Kalman filter, which has been around for 60 years.

But firstly, let's discuss a **high school problem** of describing a dynamic model: a robot that is travelling 0.2 meters every minute in both x and y directions:

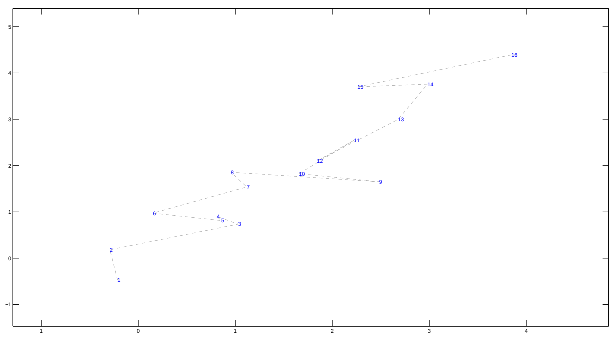
At previous time $t-1$, its position (state) is: x_{t-1}, and at current time t, its position (state) is:

$$x_t = x_{t-1} + \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix} \tag{1}$$

Let's simulate a path:



However, nothing is perfect! The dynamic model always contains a random noise:

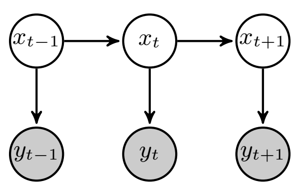


There are other methods for modeling time series, such as autoregressive models, where modeling is applied directly to observations.

However, in a state-space model (SSM), the observation $\mathbf{y}_1, \ldots, \mathbf{y}_t$ are known to us, and they usually come in one at the time. However, the latent state $\mathbf{x}_1, \ldots, \mathbf{x}_t$ is not directly observable, the state at time t only depends on the state at time $t-1$ (markov assumption).

We assume that noise comes from two different types: **(latent) state transitions** and **measurements**.

Let's have a look at the Graphical Model:



Given that a complete and generic state-space model is described as follows:

$$
\begin{aligned}
x_t &= F(x_{t-1}, w_t) \qquad & w_t \sim P_x(\cdot) \\
y_t &= H(x_t, v_t) \qquad & v_t \sim P_y(\cdot)
\end{aligned}
\tag{2}
$$

$p(x_t|x_{t-1})$ is called the (state) transition probability and $p(y_t|x_t)$ is called the measurement/emission probability.

Kalman Filter describe a very specific setting, i.e., very specific transition and measurement probablity.

2.1 Kalman Filter assumptions

For Kalman Filter, it is used to model Linear Dynamic System (LDS) with Gaussian noises. Therefore, we have the following equations:

$$
\begin{aligned}
\mathbf{x}_t &= \mathbf{A}_t \mathbf{x}_{t-1} + \mathbf{B}_t + \mathbf{w}_t & \mathbf{w}_t &\sim \mathcal{N}(0, \mathcal{Q}_t) \\
\mathbf{y}_t &= \mathbf{H}_t \mathbf{x}_t + \mathbf{C}_t + \mathbf{v}_t & \mathbf{v}_t &\sim \mathcal{N}(0, \mathbf{R}_t)
\end{aligned}
\tag{3}
$$

for many machine learning applications we assume the parameters $(\mathbf{A}_t, \mathbf{H}_t, \mathbf{B}_t, \mathbf{C}_t, \mathcal{Q}_t, \mathbf{R}_t)$ are **not** time-varying, so we can simply remove the subscript $_t$

2.2 Motivating examples

This example (or similar) is used in many Engineering textbooks:

A truck on perfectly frictionless, infinitely long straight rails. Initially the truck is stationary at position 0, but it is buffeted this way and that by **random acceleration**, i.e., we assume $a_t \sim \mathcal{N}(0, \sigma^2)$. We measure position of the truck every $\triangle$t seconds, but these measurements are imprecise. We want to maintain a model of where the truck is and what its velocity.

Using simple high school physics (assume you still remember it), where:

1. x: displacement
2. $\dot{x}$: velocity
3. a: acceleration

2.2.1 transition probablity

let's write down the state transtion equation from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$:

$$
\begin{aligned}
x_t &= x_{t-1} + \dot{x}_{t-1} \triangle t + \frac{1}{2} a_t (\triangle t)^2 \\
\dot{x}_t &= \dot{x}_{t-1} + a_t \triangle t
\end{aligned}
\tag{4}
$$

You can write out the complete Linear-Gaussian Dynamic equations:

$$
\begin{aligned}
\begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} &= \underbrace{\begin{bmatrix} 1 & \triangle t \\ 0 & 1 \end{bmatrix}}_{\mathbf{A}_t} \underbrace{\begin{bmatrix} x_{t-1} \\ \dot{x}_{t-1} \end{bmatrix}}_{\mathbf{x}_{t-1}} + \underbrace{\begin{bmatrix} \frac{1}{2} a_t (\triangle t)^2 \\ a_t \triangle t \end{bmatrix}}_{\mathbf{w}_t} \\
\mathbf{x}_t = \mathbf{A} \mathbf{x}_{t-1} + \mathbf{w}_t & \qquad \mathbf{w}_t \sim \mathcal{N}(0, \mathcal{Q}_t)
\end{aligned}
\tag{5}
$$

$$
\mathbf{x}_t = \begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} = \begin{bmatrix} 1 & \triangle t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \dot{x}_{t-1} \end{bmatrix} + \underbrace{\begin{bmatrix} \frac{1}{2} a_t (\triangle t)^2 \\ a_t \triangle t \end{bmatrix}}_{\mathbf{w}_t}
\tag{6}
$$

we know $a_t \sim \mathcal{N}(0, \sigma^2) \;\; \forall t$, then if we let $\mathbf{w}_t \sim \mathcal{N}(0, Q_t)$, what is Q_t? Since $\mathbf{w}_t$ contains both the random variable a_t and constants $\triangle t$

$$
\begin{aligned}
Q_t = \mathbf{Cov}(\mathbf{x}_t) &= \mathbf{Cov}\left(\begin{bmatrix}1 & \triangle t\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_{t-1}\\ \dot{x}_{t-1}\end{bmatrix} + \begin{bmatrix}\frac{1}{2}a_t(\triangle t)^2\\ a_t\triangle t\end{bmatrix}\right)\\
&= \mathbf{Cov}\left(\begin{bmatrix}\frac{1}{2}a_t(\triangle t)^2\\ a_t\triangle t\end{bmatrix}\right) \qquad \text{thanks to additive noise}\\
&= \mathbb{E}\left[(a_t)^2\begin{bmatrix}\frac{1}{2}(\triangle t)^2\\ \triangle t\end{bmatrix}\begin{bmatrix}\frac{1}{2}(\triangle t)^2 & \triangle t\end{bmatrix}\right] \quad \text{separate r.v.} a_t\\
&= \sigma^2\begin{bmatrix}\frac{1}{4}(\triangle t)^4 & \frac{1}{2}(\triangle t)^3\\ \frac{1}{2}(\triangle t)^3 & (\triangle t)^2\end{bmatrix} \qquad \mathbb{E}[a_t] = 0
\end{aligned}
\tag{7}
$$

2.2.2 Measurement Equation

At each time step t, we can make a noisy measurement of the true position of the truck, calling it y_t

Let us suppose the measurement noise, v_t is also normally distributed, with mean 0 and standard deviation σ_y

$$
\begin{aligned}
y_t &= \mathbf{H}\mathbf{x}_t + C + v_t \qquad v_t \sim \mathcal{N}(0, \sigma_y^2)\\
&= \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}x_t\\ \dot{x}_t\end{bmatrix} + v_t \qquad v_t \sim \mathcal{N}(0, \sigma_y^2)
\end{aligned}
\tag{8}
$$

In summary, we have a complete linear-Gaussian dynamic system of the form:

1. transition probability:

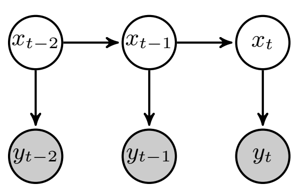
$$
p\left(\begin{bmatrix}x_t\\ \dot{x}_t\end{bmatrix} \middle| \begin{bmatrix}x_{t-1}\\ \dot{x}_{t-1}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}1 & \triangle t\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_{t-1}\\ \dot{x}_{t-1}\end{bmatrix}, \sigma^2\begin{bmatrix}\frac{1}{4}(\triangle t)^4 & \frac{1}{2}(\triangle t)^3\\ \frac{1}{2}(\triangle t)^3 & (\triangle t)^2\end{bmatrix}\right) \tag{9}
$$

2. measurement probability:

$$
p(y_t|\mathbf{x}_t) = \mathcal{N}\Big(\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}x_t\\ \dot{x}_t\end{bmatrix}, \sigma_y^2\Big) \tag{10}
$$

3 Graphical Model and Inference algorithm

let me show the graphical model again:



Markov Assumption, or one can tell from the Markov blanket of x_t and y_t:

$$\begin{aligned} p(x_t|x_1,\ldots,x_{t-1},y_1,\ldots,y_{t-1}) &= p(x_t|x_{t-1}) \\ p(y_t|x_1,\ldots,x_{t-1},x_t,y_1,\ldots,y_{t-1}) &= p(y_t|x_t) \end{aligned} \tag{11}$$

note that in the first equation, y_t has been excluded from the "rest of variables".

3.1 Linear Gaussian Dynamic Model

1. **Transition probability:**

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B} + \mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}(0, \mathcal{Q}_t) \\ p(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}, \mathcal{Q}_t) \end{aligned} \tag{12}$$

2. **Measurement probability:**

$$\begin{aligned} \mathbf{y}_t &= \mathbf{H}\mathbf{x}_t + \mathbf{C} + \mathbf{v}_t \qquad \mathbf{v}_t \sim \mathcal{N}(0, \mathbf{R}_t) \\ p(\mathbf{y}_t|\mathbf{x}_t) &= \mathcal{N}(\mathbf{H}\mathbf{x}_t + \mathbf{C}, \mathbf{R}_t) \end{aligned} \tag{13}$$

Many other dynamic models deal with non-Gaussian noise, non-linear cases such as particle filters, or non-parametric models such as Gaussian processes.

3.2 What is a filtering problem?

The filtering problem can be defined as, given all observations up to the current time t, $\mathbf{y}_{1:t} \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \tag{14}$$

While it's easy to write down $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, you'll see that in the Kalman filter derivation, the problem is further divided into:

1. **Prediction:**

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1} \tag{15}$$

2. **Update:**

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{\int_{\mathbf{s}_t} p(\mathbf{y}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{y}_{1:t-1})\mathrm{d}\mathbf{s}_t} \tag{16}$$

This is because:

$$
\begin{aligned}
\underbrace{p(\mathbf{x}_t|\mathbf{y}_{1:t})} &\propto p(\mathbf{x}_t, \mathbf{y}_{1:t}) \\
&\propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \\
&= \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{\int_{\mathbf{s}_t}(\mathbf{y}_t|\mathbf{s}_t)p(\mathrm{d}\mathbf{s}_t|\mathbf{y}_{1:t-1})} \\
&= p(\mathbf{y}_t|\mathbf{x}_t)\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1} \\
&= p(\mathbf{y}_t|\mathbf{x}_t)\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})\,\underbrace{p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})}\,\mathrm{d}\mathbf{x}_{t-1}
\end{aligned}
\tag{17}
$$

The recurrence relation makes this framework also known as a Recursive Bayesian Filter (RBF).

3.3 Kalman Filter: Prediction

Following the equation of Linear Gaussian (Bishop p.93) [1], given the prior:

$$
p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, \Sigma) \tag{18}
$$

and the likelihood, of which the mean is a linear function of $\mathbf{x}$:

$$
p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mathbf{A}\mathbf{x} + \mathbf{b}, \mathcal{Q}) \tag{19}
$$

then, marginal $p(\mathbf{y})$ is:

$$
\begin{aligned}
p(\mathbf{y}) &= \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{x}} \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathcal{Q})\mathcal{N}(\mathbf{x}|\mu, \Sigma)\mathrm{d}\mathbf{x} \\
&= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mu + \mathbf{b}, \mathcal{Q} + \mathbf{A}\Sigma\mathbf{A}^\top\right)
\end{aligned}
\tag{20}
$$

apply this to predict probabilities and do some pattern matching:

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) &\equiv \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t) \\
&= \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1} \\
&= \int_{\mathbf{x}_{t-1}} \mathcal{N}(\mathbf{x}_t|\mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}, \mathcal{Q}_t)\mathcal{N}(\mathbf{x}_{t-1}|\hat{\mu}_{t-1}, \hat{\Sigma}_{t-1})\mathrm{d}\mathbf{x}_{t-1} \\
&= \mathcal{N}\left(\mathbf{x}_t|\mathbf{A}\hat{\mu}_{t-1} + \mathbf{B}, \mathbf{A}\hat{\Sigma}_{t-1}\mathbf{A}^\top + \mathcal{Q}_t\right)
\end{aligned}
\tag{21}
$$

3.3.1 Direct Computation

Let's use an alternative way to derive prediction equation $p(\mathbf{x}_t \big| \mathbf{y}_{1:t-1})$. Firstly, we write down the expression of the random variable $\mathbf{x}_{t-1} \big| \mathbf{y}_{1:t-1}$ in terms of the constant and random component:

$$\begin{aligned} \mathbf{x}_{t-1} \big| \mathbf{y}_{1:t-1} &= \mathbb{E}[\mathbf{x}_{t-1}] + \epsilon(\mathbf{x}_{t-1}) \quad \big| \mathbf{y}_{1:t-1} \\ &= \hat{\mu}_{t-1} + \epsilon(\mathbf{x}_{t-1}) \quad \big| \mathbf{y}_{1:t-1} \end{aligned} \tag{22}$$

Here we break down $\mathbf{x}_{t-1} \big| \mathbf{y}_{1:t-1}$ into the constant part, $\mathbb{E}[\mathbf{x}_{t-1} \big| \mathbf{y}_{1:t-1}]$ and random part, $\epsilon(\mathbf{x}_{t-1}) \big| \mathbf{y}_{1:t-1}$.

Next, we express $\mathbf{x}_t \big| \mathbf{y}_{1:t-1}$ (note this is different to Eq.22) In the spirit of the recursion, we also express $\epsilon(\mathbf{x}_t) \big| \mathbf{y}_{1:t-1}$ in terms of $\epsilon(\mathbf{x}_{t-1}) \big| \mathbf{y}_{1:t-1}$:

$$\begin{aligned} \mathbf{x}_t \big| \mathbf{y}_{1:t-1} &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t \quad \big| \mathbf{y}_{1:t-1} \\ &= \mathbf{A}(\hat{\mu}_{t-1} + \epsilon(\mathbf{x}_{t-1})) + \mathbf{w}_t \quad \big| \mathbf{y}_{1:t-1} \\ &= \underbrace{\mathbf{A}\hat{\mu}_{t-1}}_{\mathbb{E}[\mathbf{x}_t \mid \mathbf{y}_{1:t-1}]} + \underbrace{\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t}_{\epsilon(\mathbf{x}_t) \mid \mathbf{y}_{1:t-1}} \quad \big| \mathbf{y}_{1:t-1} \end{aligned} \tag{23}$$

Note that in this section, everything is $\big| \mathbf{y}_{1:t-1}$. So we could theoretically remove it for clarity. But we kept it to avoid confusion.

1. **prediction mean**: $\bar{\mu}_t = \mathbb{E}[\mathbf{x}_t \big| \mathbf{y}_{1:t-1}]$:

$$\begin{aligned} \bar{\mu}_t &= \mathbb{E}[\mathbf{A}\hat{\mu}_{t-1} + \mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t \quad \big| \mathbf{y}_{1:t-1}] \\ &= \mathbf{A}\hat{\mu}_{t-1} \end{aligned} \tag{24}$$

2. **prediction covariance**: $\bar{\Sigma}_t = \mathbb{VAR}[\mathbf{x}_t \big| \mathbf{y}_{1:t-1}]$

$$\begin{aligned} \bar{\Sigma}_t &= \mathbb{E}\big[(\mathbf{A}\hat{\mu}_{t-1} + \mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)(\mathbf{A}\hat{\mu}_{t-1} + \mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)^\top \quad \big| \mathbf{y}_{1:t-1}\big] \\ &= \mathbb{E}\big[(\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)(\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)^\top \quad \big| \mathbf{y}_{1:t-1}\big] \quad \text{remove constant terms} \\ &= \mathbb{E}\big[(\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)(\epsilon(\mathbf{x}_{t-1})^\top \mathbf{A}^\top + \mathbf{w}_t^\top) \quad \big| \mathbf{y}_{1:t-1}\big] \quad \text{expand transpose} \\ &= \mathbb{E}\big[\mathbf{A}\epsilon(\mathbf{x}_{t-1})\epsilon(\mathbf{x}_{t-1})^\top \mathbf{A}^\top\big] + \mathbb{E}[\mathbf{w}_t \mathbf{w}_t^\top] \quad \big| \mathbf{y}_{1:t-1} \qquad \because \mathbb{E}[\epsilon(\mathbf{x}_{t-1})\mathbf{w}_t^\top] = \mathbf{0} \\ &= \mathbf{A}\mathbb{E}\big[\epsilon(\mathbf{x}_{t-1})\epsilon(\mathbf{x}_{t-1})^\top\big]\mathbf{A}^\top + \mathbb{E}[\mathbf{w}_t \mathbf{w}_t^\top] \quad \big| \mathbf{y}_{1:t-1} \\ &= \mathbf{A}\hat{\Sigma}_{t-1}\mathbf{A}^\top + \mathcal{Q}_t \end{aligned} \tag{25}$$

since

$$\begin{aligned} \epsilon(\mathbf{x}_{t-1}) \big| \mathbf{y}_{1:t-1} &= \mathbf{x}_{t-1} - \mathbb{E}[\mathbf{x}_{t-1} \big| \mathbf{y}_{1:t-1}] \\ &\sim \mathcal{N}(0, \hat{\Sigma}_{t-1}) \end{aligned} \tag{26}$$

3.4 new expression: systematically!

The big picture here is that we will model jointly:

$$
\begin{aligned}
&p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}_{1:t-1}) \\
&\quad = \mathcal{N}\Big(\begin{bmatrix} \mathbb{E}[\mathbf{x}_t|\mathbf{y}_{1:t-1})] \\ \mathbb{E}[\mathbf{y}_t|\mathbf{y}_{1:t-1})] \end{bmatrix} ,\ \begin{bmatrix} \mathbb{E}[\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})^\top] & \mathbb{E}[\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})^\top] \\ \mathbb{E}[\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})^\top] & \mathbb{E}[\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})^\top] \end{bmatrix} \Big)
\end{aligned}
\tag{27}
$$

Then, we can simply work out from the conditional density $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ using conditional Gaussian formula.

Here we try to express $\mathbf{x}_t\big|\mathbf{y}_{1:t-1}$ and $\mathbf{y}_t\big|\mathbf{y}_{1:t-1}$ in terms of the random part, $\epsilon(\mathbf{x}_t)\big|\mathbf{y}_{1:t-1}$ and $\epsilon(\mathbf{y}_t)\big|\mathbf{y}_{1:t-1}$ respectively. For the purpose of recursion, we need to further express the random part in terms of $\epsilon(\mathbf{x}_{t-1})\big|\mathbf{y}_{1:t-1}$.

$$
\begin{aligned}
\mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t \\
&= \mathbf{A}(\mathbb{E}[\mathbf{x}_{t-1}] + \epsilon(\mathbf{x}_{t-1})) + \mathbf{w}_t \\
&= \underbrace{\mathbf{A}\mathbb{E}[\mathbf{x}_{t-1}]}_{\mathbb{E}[\mathbf{x}_t|\mathbf{y}_{1:t-1}]} + \underbrace{\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t}_{\epsilon(\mathbf{x}_t)|\mathbf{y}_{1:t-1}}
\end{aligned}
\tag{28}
$$

for $\mathbf{y}_t$, it is treated as a random variable in here:

$$
\begin{aligned}
\mathbf{y}_t &= \mathbf{H}\mathbf{x}_t + \mathbf{v}_t \\
&= \mathbf{H}(\mathbf{A}\mathbb{E}[\mathbf{x}_{t-1}] + \mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t) + \mathbf{v}_t \\
&= \underbrace{\mathbf{H}\mathbf{A}\mathbb{E}[\mathbf{x}_{t-1}]}_{\mathbb{E}[\mathbf{y}_t|\mathbf{y}_{1:t-1}]} + \underbrace{\mathbf{H}\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{H}\mathbf{w}_t + \mathbf{v}_t}_{\epsilon(\mathbf{y}_t)|\mathbf{y}_{1:t-1}}
\end{aligned}
\tag{29}
$$

Basically, we break down the $\mathbf{x}_t\big|\mathbf{y}_{1:t-1}$ and $\mathbf{y}_t\big|\mathbf{y}_{1:t-1}$ into two parts:

1. random part:

$$
\begin{aligned}
\epsilon(\mathbf{x}_t)\big|\mathbf{y}_{1:t-1} &= \mathbf{A}\epsilon(\mathbf{x}_{t-1}\big|\mathbf{y}_{1:t-1}) + \mathbf{w}_t \\
\epsilon(\mathbf{y}_t)\big|\mathbf{y}_{1:t-1} &= \mathbf{H}\mathbf{A}\epsilon(\mathbf{x}_{t-1}\big|\mathbf{y}_{1:t-1}) + \mathbf{H}\mathbf{w}_t + \mathbf{v}_t
\end{aligned}
\tag{30}
$$

2. constant parts:

$$
\begin{aligned}
\mathbb{E}[\mathbf{x}_t\big|\mathbf{y}_{1:t-1}] &= \mathbf{A}\mathbb{E}[\mathbf{x}_{t-1}\big|\mathbf{y}_{1:t-1}] \\
\mathbb{E}[\mathbf{y}_t\big|\mathbf{y}_{1:t-1}] &= \mathbf{H}\mathbf{A}\mathbb{E}[\mathbf{x}_{t-1}\big|\mathbf{y}_{1:t-1}]
\end{aligned}
\tag{31}
$$

The Independence assumptions:

$$\begin{aligned}
\mathbf{Cov}(\epsilon(\mathbf{x}_{t-1}), \mathbf{w}_t) &= 0 \\
\mathbf{Cov}(\epsilon(\mathbf{x}_{t-1}), \mathbf{v}_t) &= 0 \\
\mathbf{Cov}(\mathbf{w}_t, \mathbf{v}_t) &= 0
\end{aligned} \tag{32}$$

we should have the following quantities, note that $\epsilon(\mathbf{x}_t)$ and $\epsilon(\mathbf{y}_t)$ are all zero-mean, and in order to compute all the quantities of Eq.(27):

$$\begin{aligned}
\mathbb{E}[\epsilon(\mathbf{x}_t)\ \epsilon(\mathbf{x}_t)^\top \big| \mathbf{y}_{1:t-1}] &= \mathbb{E}[(\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)(\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)^\top \big| \mathbf{y}_{1:t-1}] \\
&= \mathbf{A}\hat{\Sigma}_{t-1}\mathbf{A}^\top + \mathcal{Q}_t \\
&= \bar{\Sigma}_t
\end{aligned} \tag{33}$$

$$\begin{aligned}
\mathbb{E}[\epsilon(\mathbf{y}_t)\ \epsilon(\mathbf{x}_t)^\top \big| \mathbf{y}_{1:t-1}] &= \mathbb{E}\left[(\mathbf{H}\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{H}\mathbf{w}_t + \mathbf{v}_t)(\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{w}_t)^\top \big| \mathbf{y}_{1:t-1}\right] \\
&= \mathbf{H}\big(\mathbf{A}\hat{\Sigma}_{t-1}\mathbf{A}^\top + \mathcal{Q}_t\big) \\
&= \mathbf{H}\bar{\Sigma}_t \quad \text{substitute Eq.(34)} \\
\implies \mathbb{E}[\epsilon(\mathbf{x}_t)\ \epsilon(\mathbf{y}_t)^\top \big| \mathbf{y}_{1:t-1}] &= \bar{\Sigma}_t\mathbf{H}^\top
\end{aligned} \tag{34}$$

$$\begin{aligned}
\mathbb{E}[\epsilon(\mathbf{y}_t)\epsilon(\mathbf{y}_t)^\top \big| \mathbf{y}_{1:t-1}] &= \mathbb{E}\left[(\mathbf{H}\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{H}\mathbf{w}_t + \mathbf{v}_t)(\mathbf{H}\mathbf{A}\epsilon(\mathbf{x}_{t-1}) + \mathbf{H}\mathbf{w}_t + \mathbf{v}_t)^\top \big| \mathbf{y}_{1:t-1}\right] \\
&= \mathbf{H}\big(\mathbf{A}\hat{\Sigma}_{t-1}\mathbf{A}^\top + \mathcal{Q}_t\big)\mathbf{H}^\top + \mathbf{R}_t \\
&= \mathbf{H}\ \bar{\Sigma}_t\ \mathbf{H}^\top + \mathbf{R}_t
\end{aligned} \tag{35}$$

finally the constant part:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}_t \big| \mathbf{y}_{1:t-1}] &= \mathbf{A}\mathbb{E}[\mathbf{x}_{t-1} \big| \mathbf{y}_{1:t-1}] \\
&= \mathbf{A}\hat{\mu}_{t-1}
\end{aligned} \tag{36}$$

$$\begin{aligned}
\mathbb{E}[\mathbf{y}_t \big| \mathbf{y}_{1:t-1}] &= \mathbf{H}\mathbf{A}\mathbb{E}[\mathbf{x}_{t-1} \big| \mathbf{y}_{1:t-1}] \\
&= \mathbf{H}\mathbf{A}\hat{\mu}_{t-1}
\end{aligned} \tag{37}$$

3.4.1 Kalman Filter Update: $p(\mathbf{x}_t | \mathbf{y}_{1:t}) \equiv \mathcal{N}(\hat{\mu}_t, \hat{\Sigma}_t)$

Similar to the Gaussian Process and RKHS lecture notes, when we have joint Gaussian Density:

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\mathbf{u}} \\ \mu_{\mathbf{v}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathbf{uu}} & \Sigma_{\mathbf{uv}} \\ \Sigma_{\mathbf{vu}} & \Sigma_{\mathbf{vv}} \end{bmatrix} \right) \tag{38}$$

the conditional is then:

$$p(\mathbf{u}|\mathbf{v}) = \mathcal{N}\left(\mu_\mathbf{u} + \Sigma_{\mathbf{uv}}\Sigma_{\mathbf{vv}}^{-1}(\mathbf{v} - \mu_\mathbf{v}), \Sigma_{\mathbf{uu}} - \Sigma_{\mathbf{uv}}\Sigma_{\mathbf{vv}}^{-1}\Sigma_{\mathbf{vu}}\right) \tag{39}$$

the corresponding joint density in Kalman Filter case is:

$$\begin{aligned} p(\mathbf{u}, \mathbf{v}) &\equiv p(\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_{1:t-1}) \\ \implies p(\mathbf{u}|\mathbf{v}) &\equiv p(\mathbf{x}_t | \mathbf{y}_t, \mathbf{y}_1, \ldots, \mathbf{y}_{t-1}) \\ &= p(\mathbf{x}_t | \mathbf{y}_{1:t}) \end{aligned} \tag{40}$$

looking at the join Gaussian density in Kalman filter setting, i.e., Eq.(27):

$$\begin{aligned} &p(\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_{1:t-1}) \\ &= \mathcal{N}\left(\begin{bmatrix} \underbrace{\mathbb{E}[\mathbf{x}_t|\mathbf{y}_{1:t-1})]}_{\mu_\mathbf{u}} \\ \underbrace{\mathbb{E}[\mathbf{y}_t|\mathbf{y}_{1:t-1})]}_{\mu_\mathbf{v}} \end{bmatrix}, \begin{bmatrix} \underbrace{\mathbb{E}[\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})^\top]}_{\Sigma_{\mathbf{uu}}} & \underbrace{\mathbb{E}[\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})^\top]}_{\Sigma_{\mathbf{uv}}} \\ \underbrace{\mathbb{E}[\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{x}_t|\mathbf{y}_{1:t-1})^\top]}_{\Sigma_{\mathbf{vu}}} & \underbrace{\mathbb{E}[\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})\epsilon(\mathbf{y}_t|\mathbf{y}_{1:t-1})^\top]}_{\Sigma_{\mathbf{vv}}} \end{bmatrix} \right) \end{aligned} \tag{41}$$

3.4.2 mean: $\hat{\mu}_t = \mathbb{E}[\mathbf{x}_t|\mathbf{y}_{1:t}]$

look at the mean part of Eq.(39):

$$\begin{aligned} \mathbb{E}[\mathbf{u}|\mathbf{v}] &= \mu_\mathbf{u} + \Sigma_{\mathbf{uv}}\Sigma_{\mathbf{vv}}^{-1}(\mathbf{v} - \mu_\mathbf{v}) \\ \implies \mathbb{E}[\mathbf{x}_t|\mathbf{y}_{1:t}] &= \mathbb{E}[\mathbf{x}_t] + \mathbb{E}[\epsilon(\mathbf{x}_t)\epsilon(\mathbf{y}_t)^\top]\ \mathbb{E}[\epsilon(\mathbf{y}_t)\epsilon(\mathbf{y}_t)^\top]^{-1}(\mathbf{y}_t - \mathbb{E}[\mathbf{y}_t]) \quad \Big|\ \mathbf{y}_{1:t-1} \\ &= \mathbf{A}\hat{\mu}_{t-1} + \bar{\Sigma}_t^\top \mathbf{H}(\mathbf{H}\bar{\Sigma}_t\mathbf{H}^\top + \mathbf{R}_t)^{-1}(\mathbf{y}_t - \mathbf{H}\mathbf{A}\hat{\mu}_{t-1}) \end{aligned} \tag{42}$$

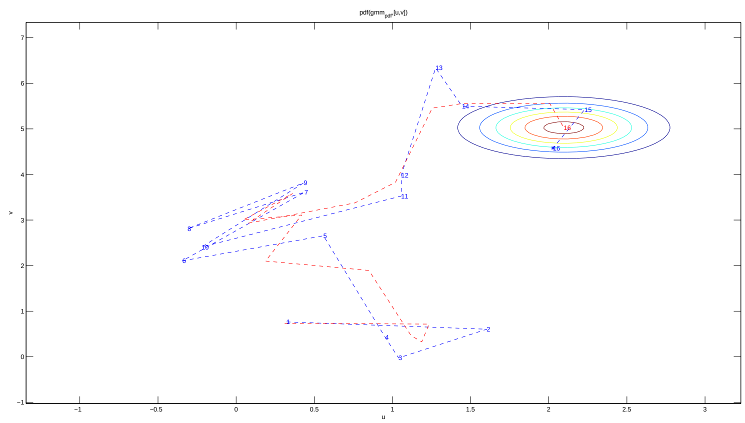
3.4.3 covariance: $\hat{\Sigma}_t = \mathbf{Cov}[\mathbf{x}_t|\mathbf{y}_{1:t}]$

look at the covariance part of Eq.(39):

$$\begin{aligned} \mathbf{Cov}[\mathbf{u}|\mathbf{v}] &= \Sigma_{\mathbf{uu}} - \Sigma_{\mathbf{uv}}\Sigma_{\mathbf{vv}}^{-1}\Sigma_{\mathbf{vu}} \\ \mathbf{Cov}[\mathbf{x}_t|\mathbf{y}_{1:t}] &= \mathbb{E}[\epsilon(\mathbf{x}_t)\epsilon(\mathbf{x}_t)^\top] - \mathbb{E}[\epsilon(\mathbf{x}_t)\epsilon(\mathbf{y}_t)^\top]\mathbb{E}[\epsilon(\mathbf{y}_t)\epsilon(\mathbf{y}_t)^\top]^{-1}\mathbb{E}[\epsilon(\mathbf{y}_t)\epsilon(\mathbf{x}_t)^\top] \quad \Big|\ \mathbf{y}_{1:t-1} \\ &= \bar{\Sigma}_t - \underbrace{\bar{\Sigma}_t\mathbf{H}^\top(\mathbf{H}(\bar{\Sigma}_t)\mathbf{H}^\top + \mathbf{R}_t)^{-1}}_{\mathbf{K}}\mathbf{H}\bar{\Sigma}_t \\ &= (\mathbf{I} - \mathbf{K}\mathbf{H})\bar{\Sigma}_t \end{aligned} \tag{43}$$

3.5 Kalman Filter Demo:

Notice of the **smoothing** effect:



3.6 Kalman Filter 1-d case Optional

3.6.1 mean

$$
\begin{aligned}
\text{k-d:} \quad & \hat{\mu}_t = \mathbf{A}\hat{\mu}_{t-1} + \bar{\Sigma}_t^\top \mathbf{H}(\mathbf{H}\bar{\Sigma}_t\mathbf{H}^\top + \mathbf{R}_t)^{-1}(\mathbf{y}_t - \mathbf{H}\mathbf{A}\hat{\mu}_{t-1}) \\
\text{1-d:} \quad & \hat{\mu}_t = a\hat{\mu}_{t-1} + \frac{\bar{\sigma}_t^2 h(y_t - ha\hat{\mu}_{t-1})}{h^2\bar{\sigma}_t^2 + R_t} \\
& = \frac{a\hat{\mu}_{t-1}(h^2\bar{\sigma}_t^2 + R_t) + \bar{\sigma}_t^2 h(y_t - ha\hat{\mu}_{t-1})}{h^2\bar{\sigma}_t^2 + R_t} \\
& = \frac{a\hat{\mu}_{t-1}R_t + \bar{\sigma}_t^2 h y_t}{h^2\bar{\sigma}_t^2 + R_t}
\end{aligned} \tag{44}
$$

covariance: $\hat{\Sigma}_t = \mathbf{Cov}[\mathbf{x}_t|\mathbf{y}_{1:t}]$:

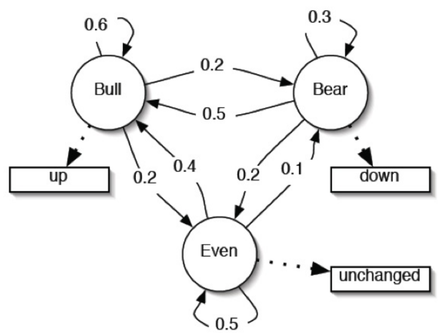
$$
\begin{aligned}
\text{k-d:} \quad & \hat{\Sigma}_t = \bar{\Sigma}_t - \bar{\Sigma}_t\mathbf{H}^\top(\mathbf{H}(\bar{\Sigma}_t)\mathbf{H}^\top + \mathbf{R}_t)^{-1}\mathbf{H}\bar{\Sigma}_t \\
\text{1-d:} \quad & \hat{\sigma}_t = \frac{\bar{\sigma}_t^2(h^2\bar{\sigma}_t^2 + R_t) - (\bar{\sigma}_t^2)^2 h^2}{h^2\bar{\sigma}_t^2 + R_t} \\
& = \frac{\bar{\sigma}_t^2(h^2\bar{\sigma}_t^2 + R_t) - (\bar{\sigma}_t^2)^2 h^2}{h^2\bar{\sigma}_t^2 + R_t} \\
& = \frac{\bar{\sigma}_t^2 R_t}{h^2\bar{\sigma}_t^2 + R_t}
\end{aligned} \tag{45}
$$

4 Discrete States Dynamic Model: Hidden Markov Model

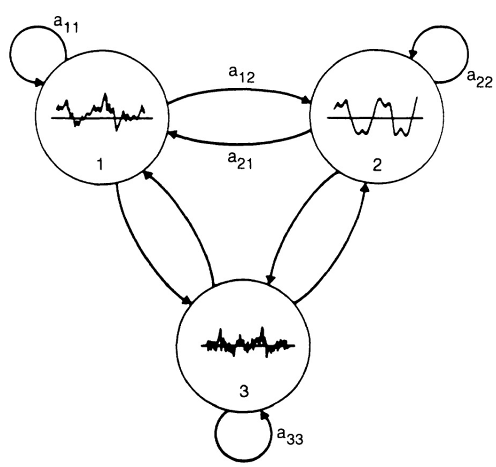
4.1 latent state

Many example of dynamic system may require the latent state to be discrete. Instead of calling it $\mathbf{x}_t$, let's call it q_t, for example:

In simple stock market, the latent states may present $q_t =$ Bull, Bear, Even

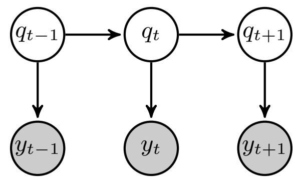


In speech recognition, the latent states may present $q_t =$ each of the "noun" and "consonants" , for example the word "cat" should contain the states $\{null, k, a, t\}$. Can you write down its transition probability?



By the way, speech recognition is now completely replaced by neural network methods nowadays.

4.2 Hidden Markov Model



Discrete Transition Probability:

$$\begin{aligned} p(q_t|q_1, \dots, q_{t-1}, y_1, \dots, y_{t-1}) &= p(q_t|q_{t-1}) \\ &= a_{q_{t-1},q_t} \end{aligned} \tag{46}$$

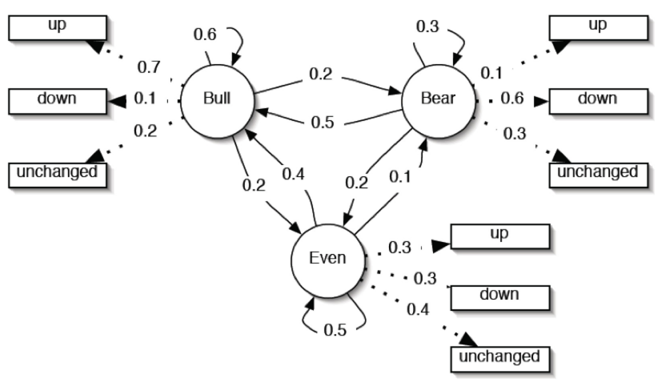
Continuous/Discrete Measurement probability:

$$p(y_t|q_1, \dots, q_{t-1}, q_t, y_1, \dots, y_{t-1}) = p(y_t|q_t) \tag{47}$$

HMM's observation y_t do not need to be discrete. They can be continuous as well. But just in case they are also discrete, $p(y_t|q_t)$ can also describe by a matrix B

4.3 HMM Model

Looking at the following example



let's see what their parameters are:

4.3.1 transition probability:

- Let Bull = 1, Bear = 2, Even = 3:

$$
\begin{aligned}
p(q_t = 1 | q_{t-1} = 1) = 0.6 \\
p(q_t = 2 | q_{t-1} = 1) = 0.2 \\
p(q_t = 3 | q_{t-1} = 1) = 0.2
\end{aligned} \tag{48}
$$

$$
\begin{aligned}
p(q_t = 1 | q_{t-1} = 2) = 0.5 \\
p(q_t = 2 | q_{t-1} = 2) = 0.3 \\
p(q_t = 3 | q_{t-1} = 2) = 0.2
\end{aligned} \tag{49}
$$

$$
\begin{aligned}
p(q_t = 1 | q_{t-1} = 3) = 0.4 \\
p(q_t = 2 | q_{t-1} = 3) = 0.1 \\
p(q_t = 3 | q_{t-1} = 3) = 0.5
\end{aligned} \tag{50}
$$

therefore the parameters can be fully described by matrix A:

$$
A = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \tag{51}
$$

think about the scenario where the rows of A are identical? What does it tell you about the transition?

4.3.2 measurement probability

- Let Bull = 1, Bear = 2, Even = 3:
- Let Up = 1, Down = 2, Uneven = 3:

$$
\begin{aligned}
p(y_t = 1 | q_t = 1) = 0.7 \\
p(y_t = 2 | q_t = 1) = 0.1 \\
p(y_t = 3 | q_t = 1) = 0.2
\end{aligned} \tag{52}
$$

$$
\begin{aligned}
p(y_t = 1 | q_t = 2) = 0.1 \\
p(y_t = 2 | q_t = 2) = 0.6 \\
p(y_t = 3 | q_t = 2) = 0.3
\end{aligned} \tag{53}
$$

$$
\begin{aligned}
p(y_t = 1 | q_t = 3) = 0.3 \\
p(y_t = 2 | q_t = 3) = 0.3 \\
p(y_t = 3 | q_t = 3) = 0.4
\end{aligned} \tag{54}
$$

therefore, we have discrete emission probability to become:

$$
B = \begin{bmatrix} 0.7 & 0.1 & 0.2 \\ 0.1 & 0.6 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix} \tag{55}
$$

4.4 Hidden Markov Model

The HMM Parameter λ (discrete measurement case) contains:

$$\lambda = \{A, B, \pi\} \tag{56}$$

π is the probability of the initial state , i.e., $p(q_1)$. We use $\pi_i \equiv p(q_1 = i)$. This is not captured by A or B:

Let $\mathcal{Q} = q_1, \ldots q_T$ and $\mathbf{Y} = y_1, \ldots y_T$:

Three major operations of HMM:

$$\begin{aligned} &\text{Evaluate } p(\mathbf{Y}|\lambda) \\ &\lambda_{\text{MLE}} = \arg\max_{\lambda} p(\mathbf{Y}|\lambda) \\ &\arg\max_{\mathcal{Q}} p(\mathbf{Y}|\mathcal{Q}, \lambda) \end{aligned} \tag{57}$$

We will discuss Evaluation first.

4.5 Evaluate $p(Y|\lambda)$

The usual way to compute this:

$$\begin{aligned} p(\mathbf{Y}|\lambda) &= \sum_{\mathcal{Q}} [p(\mathbf{Y}, \mathcal{Q}|\lambda)] = \sum_{q_1=1}^{k} \ldots, \sum_{q_T=1}^{k} [p(y_1, \ldots, y_T, q_1, \ldots, q_T|\lambda)] \\ &= \sum_{q_1=1}^{k} \cdots \sum_{q_T=1}^{k} [p(y_1, \ldots, y_T, q_1, \ldots, q_T|\lambda)] \\ &= \sum_{q_1=1}^{k} \cdots \sum_{q_T=1}^{k} p(q_1)p(y_1|q_1)p(q_2|q_1) \ldots p(q_t|q_{t-1})p(y_t|q_t) \\ &= \sum_{q_1=1}^{k} \cdots \sum_{q_T=1}^{k} \pi(q_1) \prod_{t=2}^{T} a_{q_{t-1}, q_t} b_{q_t}(y_t) \end{aligned} \tag{58}$$

we let transition probability:

$$p(q_t = j | q_{t-1} = i) \equiv a_{i,j} \tag{59}$$

and measurement probability

$$p(y_t | q_t = j) \equiv b_j(y_t) \tag{60}$$

There are k^T possible values of $\mathcal{Q}$. We need simpler methods

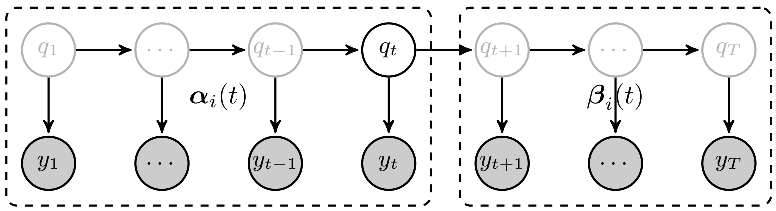
4.6 Forward and Backward Fomula

Forward Algorithm:

$$\boldsymbol{\alpha}_i(t) = p(y_1, y_2, \ldots y_t, q_t = i|\lambda) \tag{61}$$

Backward Algorithm:

$$\boldsymbol{\beta}_i(t) = p(y_{t+1}, \ldots, y_T|q_t = i, \lambda) \tag{62}$$



4.6.1 Forward

Therefore, we define **forward** procedure:

$$\begin{aligned} \boldsymbol{\alpha}_i(t) &= p(y_1, y_2, \ldots y_t, q_t = i|\lambda) \\ \implies p(\mathbf{Y}|\lambda) &= \sum_{i=1}^{k} \boldsymbol{\alpha}_i(T) \end{aligned} \tag{63}$$

This is the probability of partial sequence $y_1, \ldots, y_t$ and ending up in state i at time t. Looking at the following recursion:

$$\begin{aligned} \boldsymbol{\alpha}_j(1) &= p(y_1, q_1 = j|\lambda) \\ &= p(q_1 = j)p(y_1|q_1 = j) \\ &= \pi_j b_j(y_1) \\ \boldsymbol{\alpha}_j(2) &= p(y_1, y_2, q_2 = j|\lambda) \\ &= \sum_{i=1}^{k} p(y_1, y_2, q_1 = i, q_2 = j|\lambda) \quad \text{insert } q_1 = i \\ &= \sum_{i=1}^{k} \underbrace{p(y_1, q_1 = i)}_{\boldsymbol{\alpha}_i(1)} \underbrace{p(q_2 = j|q_1 = i)}_{a_{i,j}} \underbrace{p(y_2|q_2 = j)}_{b_j(y_2)} \\ &= \Big[\sum_{i=1}^{k} \boldsymbol{\alpha}_i(1) a_{i,j}\Big] b_j(y_2) \end{aligned} \tag{64}$$

$$
\begin{aligned}
\boldsymbol{\alpha}_j(3) &= p(y_1, y_2, y_3, q_3 = j|\lambda) \\
&= \sum_{i=1}^{k} p(y_1, y_2, y_3, q_2 = i, q_3 = j|\lambda) \\
&= \sum_{i=1}^{k} \underbrace{p(y_1, y_2, q_2 = i)}_{\boldsymbol{\alpha}_i(2)} \underbrace{p(q_3 = j|q_2 = i)}_{a_{i,j}} \underbrace{p(y_3|q_3 = j)}_{b_j(y_3)} \\
&= \Big[\sum_{i=1}^{k} \boldsymbol{\alpha}_i(2) a_{i,j}\Big] b_j(y_3) \\
&\vdots \\
\boldsymbol{\alpha}_j(t+1) &= \Big[\sum_{i=1}^{k} \boldsymbol{\alpha}_i(t) a_{i,j}\Big] b_j(y_{t+1}) \\
&\vdots \\
\boldsymbol{\alpha}_j(T) &= \Big[\sum_{i=1}^{k} \boldsymbol{\alpha}_i(T-1) a_{i,j}\Big] b_j(y_T) \\
&= p(y_1, y_2, \ldots y_T, q_T = j|\lambda)
\end{aligned}
\tag{65}
$$

which implies that:

$$
\sum_{j=1}^{k} p(y_1, y_2, \ldots y_T, q_T = j|\lambda) = p(\mathbf{Y}|\lambda) \tag{66}
$$

We have $k \times T$ summations to compute all the $\{\boldsymbol{\alpha}_j\}$.

4.6.2 backward

$$
\begin{aligned}
\boldsymbol{\beta}_i(t) = p(y_{t+1}, \ldots, y_T | q_t = i, \lambda) \\
\Longrightarrow \sum_{i=1}^{k} \boldsymbol{\beta}_i(1) \pi_i b_i(y_1) = p(\mathbf{Y}|\lambda)
\end{aligned}
\tag{67}
$$

Probability of partial sequence $y_{1+1}, y_{t+2}, \ldots, y_T$ **given** started at state i at time t:

$$
\begin{aligned}
\boldsymbol{\beta}_i(T) &= 1 \\
\boldsymbol{\beta}_i(T-1) &= p(y_T|q_{T-1}=i) \\
&= \sum_{j=1}^{k} p(q_T=j|q_{T-1}=i)p(y_T|q_T=j) \quad \because \text{insert } q_T=j \\
&= \sum_{j=1}^{k} a_{i,j}b_j(y_T) = \sum_{j=1}^{k} a_{i,j}b_j(y_T)\boldsymbol{\beta}_j(T) \\
\boldsymbol{\beta}_i(T-2) &= p(y_T, y_{T-1}|q_{T-2}=i) \\
&= \sum_{j=1}^{k} p(y_T, y_{T-1}, q_{T-1}=j|q_{T-2}=i) \quad \because \text{insert } q_{T-1}=j \\
&= \sum_{j=1}^{k} \underbrace{p(y_T, y_{T-1}|q_{T-1}=j)}_{\boldsymbol{\beta}_j(T-1)} \underbrace{p(q_{T-1}=j|q_{T-2}=i)}_{a_{i,j}} \underbrace{p(y_{T-1}|q_{T-1}=j)}_{b_j(y_{T-1})} \\
&= \sum_{j=1}^{k} a_{i,j}b_j(y_{T-1})\boldsymbol{\beta}_j(T-1) \\
&\vdots \\
\boldsymbol{\beta}_i(t) &= \sum_{j=1}^{k} a_{i,j}b_j(y_{t+1})\boldsymbol{\beta}_j(t+1) \\
&\vdots \\
\boldsymbol{\beta}_i(1) &= \sum_{j=1}^{k} a_{i,j}b_j(y_2)\boldsymbol{\beta}_j(2)
\end{aligned} \tag{68}
$$

4.7 The probability of being at a particular state

The probability of being in state i at time t for a sequence $\mathbf{Y}$:

$$
\begin{aligned}
p(q_t=i|\mathbf{Y},\lambda) &= \frac{p(\mathbf{Y}, q_t=i|\lambda)}{p(\mathbf{Y}|\lambda)} \\
&= \frac{p(\mathbf{Y}, q_t=i|\lambda)}{\sum_{j=1}^{k} p(\mathbf{Y}, q_t=j|\lambda)} \\
&= \frac{\boldsymbol{\alpha}_i(t)\boldsymbol{\beta}_i(t)}{\sum_{j=1}^{k} \boldsymbol{\alpha}_j(t)\boldsymbol{\beta}_j(t)}
\end{aligned} \tag{69}
$$

$$
\begin{aligned}
p(\mathbf{Y}, q_t = i|\lambda) &= p(\mathbf{Y}|q_t = i)p(q_t = i) \\
&= p(y_1, \dots y_t|q_t = i)p(y_{t+1}, \dots y_T|q_t = i)p(q_t = i) \quad \text{by its graphical model} \\
&= p(y_1, \dots y_t, q_t = i)p(y_{t+1}, \dots y_T|q_t = i) \quad \text{re-arrange} \\
&= \boldsymbol{\alpha}_i(t)\boldsymbol{\beta}_i(t)
\end{aligned} \tag{70}
$$

4.8 Parameter Learning

Looking at the E-M algorithm:

$$
\Theta^{(g+1)} = \arg\max_{\Theta} \left(\int_z \log\left(p(X, Z|\Theta)\right) p(Z|X, \Theta^{(g)}) \right) \mathrm{d}Z \tag{71}
$$

In HMM, we write it as:

$$
\lambda^{(g+1)} = \arg\max_{\lambda} \underbrace{\left(\int_{q \in Q} \ln\left(p(Y, q|\lambda)\right) p(q, Y|\lambda^{(g)}) \right)}_{\mathcal{Q}(\lambda, \lambda^{(g)})} \tag{72}
$$

note that we start with q_0:

$$
\begin{aligned}
\mathcal{Q}(\lambda, \lambda^{(g)}) &= \int_{q \in Q} \ln\left(p(Y, q|\lambda)\right) p(q, Y|\lambda^{(g)}) \\
&= \sum_{q_0=1}^{k} \cdots \sum_{q_T=1}^{k} \Big(\ln \pi_0 + \sum_{t=1}^{T} \ln a_{q_{t-1}, q_t} + \sum_{t=1}^{T} \ln b_{q_t}(y_t) \Big) p(q, Y|\lambda^{(g)})
\end{aligned} \tag{73}
$$

References

[1] Christopher M Bishop and Nasser M Nasrabadi, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.