Exercises for the tutorials: 1, 3.

The other exercises are for self-study and exam preparation. All material is examinable unless otherwise mentioned.

Exercise 1. *Predictive distributions for hidden Markov models*

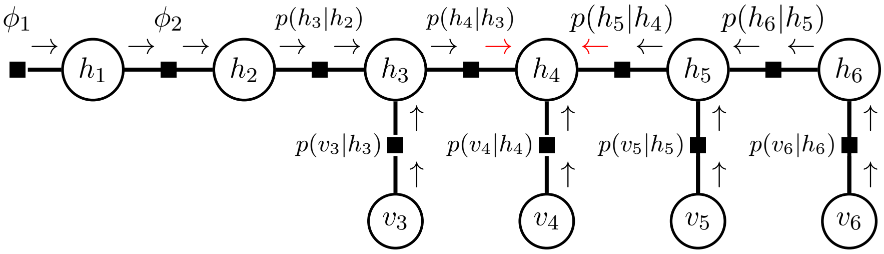
For the hidden Markov model

$$p(h_{1:d}, v_{1:d}) = p(v_1|h_1)p(h_1) \prod_{i=2}^{d} p(v_i|h_i)p(h_i|h_{i-1})$$

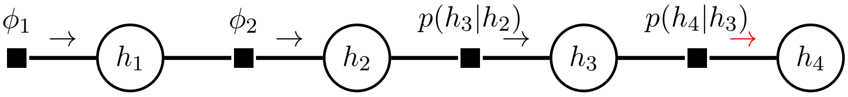
assume you have observations for v_i, $i = 1, \dots, u < d$.

(a) *Use message passing to compute $p(h_t|v_{1:u})$ for $u < t \le d$. For the sake of concreteness, you may consider the case $d = 6, u = 2, t = 4$.*

Solution. The factor graph for $d = 6, u = 2$, with messages that are required for the computation of $p(h_t|v_{1:u})$ for $t = 4$, is as follows.



The messages from the unobserved visibles v_i to their corresponding h_i, e.g. v_3 to h_3, are all one. Moreover, the message from the $p(h_5|h_4)$ node to h_4 equals one as well. This is because all involved factors, $p(v_i|h_i)$ and $p(h_i|h_{i-1})$, sum to one. Hence the factor graph reduces to a chain:



Since the variable nodes copy the messages in case of a chain, we only show the factor-to-variable messages.

The graph shows that we are essentially in the same situation as in filtering, with the difference that we use the factors $p(h_s|h_{s-1})$ for $s \ge u + 1$. Hence, we can use filtering to compute the messages until time $s = u$ and then compute the further messages with the $p(h_s|h_{s-1})$ as factors. This gives the following algorithm:

1. Compute $\alpha(h_u)$ by filtering.
2. For $s = u + 1, \dots, t$, compute

$$\alpha(h_s) = \sum_{h_{s-1}} p(h_s|h_{s-1})\alpha(h_{s-1}) \tag{S.1}$$

3. The required predictive distribution is

$$p(h_t|v_{1:u}) = \frac{1}{Z}\alpha(h_t) \qquad Z = \sum_{h_t} \alpha(h_t) \tag{S.2}$$

For $s \geq u+1$, we have that

$$\sum_{h_s} \alpha(h_s) = \sum_{h_s} \sum_{h_{s-1}} p(h_s|h_{s-1})\alpha(h_{s-1}) \tag{S.3}$$

$$= \sum_{h_{s-1}} \alpha(h_{s-1}) \tag{S.4}$$

since $p(h_s|h_{s-1})$ is normalised. This means that the normalising constant Z above equals

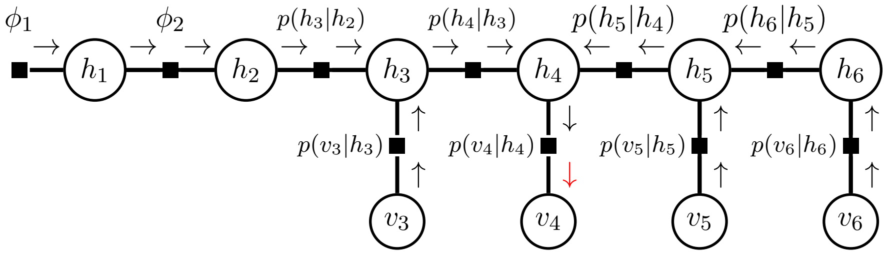
$$Z = \sum_{h_u} \alpha(h_u) = p(v_{1:u}) \tag{S.5}$$

which is the likelihood.

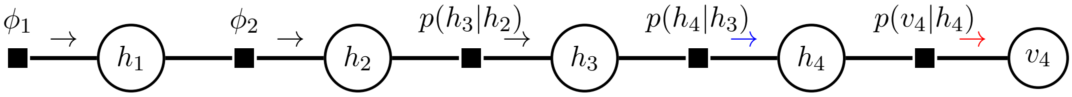
For filtering, we have seen that $\alpha(h_s) \propto p(h_s|v_{1:s})$, $s \leq u$. The $\alpha(h_s)$ for all $s > u$ are proportional to $p(h_s|v_{1:u})$. This may be seen by noting that the above arguments hold for any $t > u$.

(b) *Use message passing to compute $p(v_t|v_{1:u})$ for $u < t \leq d$. For the sake of concreteness, you may consider the case $d = 6, u = 2, t = 4$.*

Solution. The factor graph for $d = 6, u = 2$, with messages that are required for the computation of $p(v_t|v_{1:u})$ for $t = 4$, is as follows.



Due to the normalised factors, as above, the messages to the right of h_t are all one. Moreover the messages that go up from the v_i to the $h_i, i \neq t$, are also all one. Hence the graph simplifies to a chain.



The message in blue is proportional to $p(h_t|v_{1:u})$ computed in question (a). Thus assume that we have computed $p(h_t|v_{1:u})$. The predictive distribution on the level of the visibles thus is

$$p(v_t|v_{1:u}) = \sum_{h_t} p(v_t|h_t)p(h_t|v_{1:u}). \tag{S.6}$$

This follows from message passing since the last node (h_4 in the graph) just copies the (normalised) message and the next factor equals $p(v_t|h_t)$.

An alternative derivation follows from basic definitions and operations, together with the independencies in HMMs:

$$
\begin{aligned}
&\text{(sum rule)} & p(v_t|v_{1:u}) &= \sum_{h_t} p(v_t, h_t|v_{1:u}) && \text{(S.7)}\\
&\text{(product rule)} & &= \sum_{h_t} p(v_t|h_t, v_{1:u})p(h_t|v_{1:u}) && \text{(S.8)}\\
&(v_t \perp\!\!\!\perp v_{1:u} \mid h_t) & &= \sum_{h_t} p(v_t|h_t)p(h_t|v_{1:u}) && \text{(S.9)}
\end{aligned}
$$

Exercise 2. ***Viterbi algorithm***

For the hidden Markov model

$$p(h_{1:t}, v_{1:t}) = p(v_1|h_1)p(h_1)\prod_{i=2}^{t} p(v_i|h_i)p(h_i|h_{i-1})$$

assume you have observations for v_i, $i = 1, \ldots, t$. Use the max-sum algorithm to derive an iterative algorithm to compute

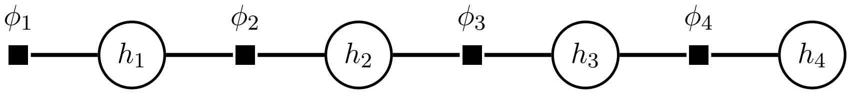
$$\hat{\mathbf{h}} = \underset{h_1,\ldots,h_t}{\operatorname{argmax}}\, p(h_{1:t}|v_{1:t}) \tag{1}$$

Assume that the latent variables h_i can take K different values, e.g. $h_i \in \{0, \ldots, K-1\}$. The resulting algorithm is known as Viterbi algorithm.

Solution. We first form the factors

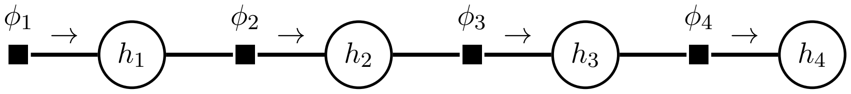
$$
\begin{aligned}
&\phi_1(h_1) = p(v_1|h_1)p(h_1) & \phi_2(h_1, h_2) &= p(v_2|h_2)p(h_2|h_1) && \text{(S.10)}\\
&\ldots & \phi_t(h_{t-1}, h_t) &= p(v_t|h_t)p(h_t|h_{t-1}) && \text{(S.11)}
\end{aligned}
$$

where the v_i are known and fixed. The posterior $p(h_1, \ldots, h_t|v_1, \ldots, v_t)$ is then represented by the following factor graph (assuming $t = 4$).



For the max-sum algorithm, we here choose h_t to be the root. We thus initialise the algorithm with $\gamma_{\phi_1 \to h_1}(h_1) = \log \phi_1(h_1) = \log p(v_1|h_1) + \log p(h_1)$ and then compute the messages from left to right, moving from the leaf ϕ_1 to the root h_t.

Since we are dealing with a chain, the variable nodes, much like in the sum-product algorithm, just copy the incoming messages. It thus suffices to compute the factor to variable messages shown in the graph, and then backtrack to h_1.



With $\gamma_{h_{i-1}\to\phi_i}(h_{i-1}) = \gamma_{\phi_{i-1}\to h_{i-1}}(h_{i-1})$, the factor-to-variable update equation is

$$\gamma_{\phi_i\to h_i}(h_i) = \max_{h_{i-1}} \log\phi_i(h_{i-1}, h_i) + \gamma_{h_{i-1}\to\phi_i}(h_{i-1}) \tag{S.12}$$

$$= \max_{h_{i-1}} \log\phi_i(h_{i-1}, h_i) + \gamma_{\phi_{i-1}\to h_{i-1}}(h_{i-1}) \tag{S.13}$$

To simplify notation, denote $\gamma_{\phi_i\to h_i}(h_i)$ by $V_i(h_i)$. We thus have

$$V_1(h_1) = \log p(v_1|h_1) + \log p(h_1) \tag{S.14}$$

$$V_i(h_i) = \max_{h_{i-1}} \log\phi_i(h_{i-1}, h_i) + V_{i-1}(h_{i-1}) \qquad i = 2, \ldots, t \tag{S.15}$$

In general, $V_1(h_1)$ and $V_i(h_i)$ are functions that depend on h_1 and h_i, respectively. Assuming that the h_i can take on the values $0, \ldots, K-1$, the above equations can be written as

$$v_{1,k} = \log p(v_1|k) + \log p(k) \qquad k = 0, \ldots, K-1 \tag{S.16}$$

$$v_{i,k} = \max_{m\in 0,\ldots,K-1} \log\phi_i(m, k) + v_{i-1,m} \qquad k = 0, \ldots, K-1, \quad i = 2, \ldots, t, \tag{S.17}$$

At the end of the algorithm, we thus have a $t \times K$ matrix $\mathbf{V}$ with elements $v_{i,k}$.

The maximisation can be performed by computing the temporary matrix $\mathbf{A}$ (via broadcasting) where the (m, k)-th element is $\log\phi_i(m, k) + v_{i-1,m}$. Maximisation then corresponds to determining the maximal value in each column.

To support the backtracking, when we compute $V_i(h_i)$ by maximising over h_{i-1}, we compute at the same time the look-up table

$$\gamma_i^*(h_i) = \operatorname*{argmax}_{h_{i-1}} \log\phi_i(h_{i-1}, h_i) + V_{i-1}(h_{i-1}) \tag{S.18}$$

When h_i takes on the values $0, \ldots, K-1$, this can be written as

$$\gamma_{i,k}^* = \operatorname*{argmax}_{m\in 0,\ldots,K-1} \log\phi_i(m, k) + v_{i-1,m} \tag{S.19}$$

This is the (row) index of the maximal element in each column of the temporary matrix $\mathbf{A}$.

After computing $v_{t,k}$ and $\gamma_{t,k}^*$, we then perform backtracking via

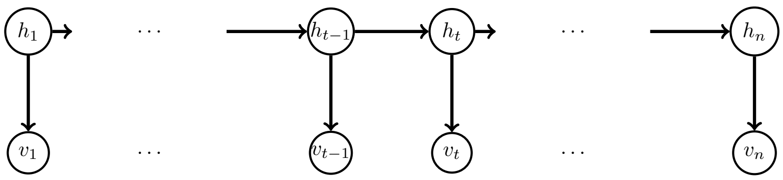
$$\hat{h}_t = \operatorname*{argmax}_{k} v_{t,k} \tag{S.20}$$

$$\hat{h}_i = \gamma_{i+1,\hat{h}_{i+1}}^* \qquad i = t-1, \ldots, 1 \tag{S.21}$$

This gives recursively $\hat{\mathbf{h}} = (\hat{h}_1, \ldots, \hat{h}_t) = \operatorname{argmax}_{h_1,\ldots,h_t} p(h_{1:t}|v_{1:t})$.

Exercise 3. ***Forward filtering backward sampling for hidden Markov models***

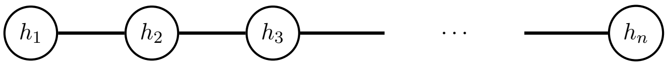
Consider the hidden Markov model specified by the following DAG.



We assume that have already run the alpha-recursion (filtering) and can compute $p(h_t|v_{1:t})$ for all t. The goal is now to generate samples $p(h_1, \dots, h_n|v_{1:n})$, i.e. entire trajectories $(h_1, \dots, h_n)$ from the posterior. Note that this is not the same as sampling from the n filtering distributions $p(h_t|v_{1:t})$. Moreover, compared to the Viterbi algorithm, the sampling approach generates samples from the full posterior rather than just returning the most probable state and its corresponding probability.

(a) *Show that $p(h_1, \dots, h_n|v_{1:n})$ forms a first-order Markov chain.*

Solution. There are several ways to show this. The simplest is to notice that the undirected graph for the hidden Markov model is the same as the DAG but with the arrows removed as there are no colliders in the DAG. Moreover, conditioning corresponds to removing nodes from an undirected graph. This leaves us with a chain that connects the h_i.

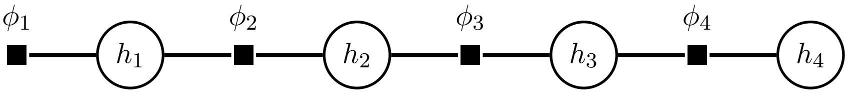


By graph separation, we see that $p(h_1, \dots, h_n|v_{1:n})$ forms a first-order Markov chain so that e.g. $h_{1:t-1} \perp\!\!\!\perp h_{t+1:n}|h_t$ (past independent from the future given the present).

(b) *Since $p(h_1, \dots, h_n|v_{1:n})$ is a first-order Markov chain, it suffices to determine $p(h_{t-1}|h_t, v_{1:n})$, the probability mass function for h_{t-1} given h_t and all the data $v_{1:n}$. Use message passing to show that*

$$p(h_{t-1}, h_t|v_{1:n}) \propto \alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t) \tag{2}$$

Solution. Since all visibles are in the conditioning set, i.e. assumed observed, we can represent the conditional model $p(h_1, \dots, h_n|v_{1:n})$ as a chain factor tree, e.g. as follows in case of $n = 4$



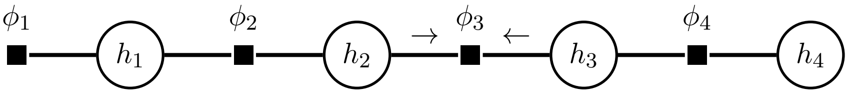
Combining the emission distributions $p(v_s|h_s)$ (and marginal $p(h_1)$) with the transition distributions $p(h_s|h_{s-1})$ we obtain the factors

$$\phi_1(h_1) = p(h_1)p(v_1|h_1) \tag{S.22}$$

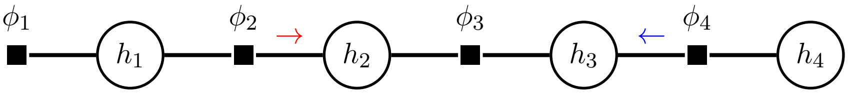
$$\phi_s(h_{s-1}, h_s) = p(h_s|h_{s-1})p(v_s|h_s) \quad \text{for } t = 2, \dots, n \tag{S.23}$$

We see from the factor tree that h_{t-1} and h_t are neighbours, being attached to the same factor node $\phi_t(h_{t-1}, h_t)$, e.g. ϕ_3 in case of $p(h_2, h_3|v_{1:4})$.

By the rules of message passing, the joint $p(h_{t-1}, h_t|v_{1:n})$ is thus proportional to ϕ_t times the messages into ϕ_t. The following graph shows the messages for the case of $p(h_2, h_3|v_{1:4})$.



Since the variable nodes only receive single messages from any direction, they copy the messages so that the messages into ϕ_t are given by $\alpha(h_{t-1})$ and $\beta(h_t)$ shown below in red and blue, respectively.



Hence,

$$p(h_{t-1}, h_t | v_{1:n}) \propto \alpha(h_{t-1})\beta(h_t)\phi_t(h_{t-1}, h_t) \tag{S.24}$$
$$\propto \alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t) \tag{S.25}$$

which is the result that we want to show.

(c) *Show that* $p(h_{t-1}|h_t, v_{1:n}) = \frac{\alpha(h_{t-1})}{\alpha(h_t)} p(h_t|h_{t-1})p(v_t|h_t)$.

Solution. Above, we have shown that

$$p(h_{t-1}, h_t|v_{1:n}) \propto \alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t), \tag{S.26}$$

that is

$$p(h_{t-1}, h_t|v_{1:n}) = \frac{1}{Z(v_{1:n})}\alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t), \tag{S.27}$$

where $Z(v_{1:n})$ is the normalising constant. It follows that

$$p(h_{t-1}|h_t, v_{1:n}) = \frac{p(h_{t-1}, h_t|v_{1:n})}{\sum_{h_{t-1}} p(h_{t-1}, h_t|v_{1:n})} \tag{S.28}$$
$$= \frac{\frac{1}{Z(v_{1:n})}\alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t)}{\sum_{h_{t-1}} \frac{1}{Z(v_{1:n})}\alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t)} \tag{S.29}$$
$$= \frac{\alpha(h_{t-1})p(h_t|h_{t-1})p(v_t|h_t)}{\sum_{h_{t-1}} \alpha(h_{t-1})p(h_t|h_{t-1})p(v_t|h_t)} \tag{S.30}$$

where we have cancelled out $Z(v_{1:n})$ and $\beta(h_t)$ that appear both in the numerator and denominator, and do not depend on h_{t-1}.
We recognise the update rule from the alpha-recursion in the denominator:

$$\alpha(h_t) = \sum_{h_{t-1}} \alpha(h_{t-1})p(h_t|h_{t-1})p(v_t|h_t). \tag{S.31}$$

We thus obtain the desired result:

$$p(h_{t-1}|h_t, v_{1:n}) = \frac{\alpha(h_{t-1})}{\alpha(h_t)}p(h_t|h_{t-1})p(v_t|h_t). \tag{S.32}$$

We thus obtain the following algorithm to generate samples from $p(h_1, \ldots, h_n|v_{1:n})$*:*

1. *Run the alpha-recursion (filtering) to determine all* $\alpha(h_t)$ *forward in time for* $t = 1, \ldots, n$.
2. *Sample* h_n *from* $p(h_n|v_{1:n}) \propto \alpha(h_n)$
3. *Go backwards in time using*

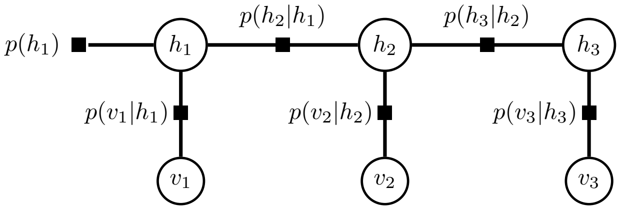
$$p(h_{t-1}|h_t, v_{1:n}) = \frac{\alpha(h_{t-1})}{\alpha(h_t)}p(h_t|h_{t-1})p(v_t|h_t) \tag{3}$$

to generate samples $h_{t-1}|h_t, v_{1:n}$ *for* $t = n, \ldots, 2$.

This algorithm is known as forward filtering backward sampling (FFBS).

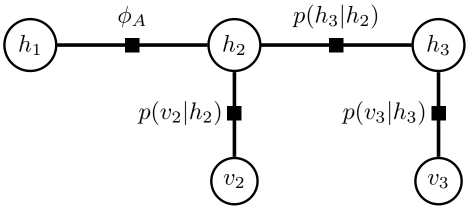
Exercise 4. ***Prediction exercise***

Consider a hidden Markov model with three visibles v_1, v_2, v_3 and three hidden variables h_1, h_2, h_3 which can be represented with the following factor graph:



This question is about computing the predictive probability $p(v_3 = 1|v_1 = 1)$.

(a) *The factor graph below represents $p(h_1, h_2, h_3, v_2, v_3 \mid v_1 = 1)$. Provide an equation that defines ϕ_A in terms of the factors in the factor graph above.*



Solution. $\phi_A(h_1, h_2) \propto p(v_1|h_1)p(h_1)p(h_2|h_1)$ with $v_1 = 1$.

(b) *Assume further that all variables are binary, $h_i \in \{0, 1\}$, $v_i \in \{0, 1\}$; that $p(h_1 = 1) = 0.5$, and that the transition and emission distributions are, for all i, given by:*

$p(h_{i+1}\|h_i)$	h_{i+1}	h_i
0	0	0
1	1	0
1	0	1
0	1	1

$p(v_i\|h_i)$	v_i	h_i
0.6	0	0
0.4	1	0
0.4	0	1
0.6	1	1

Compute the numerical values of the factor ϕ_A.

(c) Given the definition of the transition and emission probabilities, we have $\phi_A(h_1, h_2) = 0$ if $h_1 = h_2$. For $h_1 = 0, h_2 = 1$, we obtain

$$\begin{aligned} \phi_A(h_1 = 0, h_2 = 1) &= p(v_1 = 1|h_1 = 0)p(h_1 = 0)p(h_2 = 1|h_1 = 0) && \text{(S.33)} \\ &= 0.4 \cdot 0.5 \cdot 1 && \text{(S.34)} \\ &= \frac{4}{10} \cdot \frac{1}{2} && \text{(S.35)} \\ &= \frac{2}{10} = 0.2 && \text{(S.36)} \end{aligned}$$

For $h_1 = 1, h_2 = 0$, we obtain

$$\begin{aligned}\phi_A(h_1 = 1, h_2 = 0) &= p(v_1 = 1|h_1 = 1)p(h_1 = 1)p(h_2 = 0|h_1 = 1) && \text{(S.37)}\\ &= 0.6 \cdot 0.5 \cdot 1 && \text{(S.38)}\\ &= \frac{6}{10} \cdot \frac{1}{2} && \text{(S.39)}\\ &= \frac{3}{10} = 0.3 && \text{(S.40)}\end{aligned}$$

Hence

$\phi_A(h_1, h_2)$	h_1	h_2
0	0	0
0.3	1	0
0.2	0	1
0	1	1

(d) *Denote the message from variable node h_2 to factor node $p(h_3|h_2)$ by $\alpha(h_2)$. Use message passing to compute $\alpha(h_2)$ for $h_2 = 0$ and $h_2 = 1$. Report the values of any intermediate messages that need to be computed for the computation of $\alpha(h_2)$.*

Solution. The message from h_1 to ϕ_A is one. The message from ϕ_A to h_2 is

$$\begin{aligned}\mu_{\phi_A \to h_2}(h_2 = 0) &= \sum_{h_1} \phi_A(h_1, h_2 = 0) && \text{(S.41)}\\ &= 0.3 && \text{(S.42)}\\ \mu_{\phi_A \to h_2}(h_2 = 1) &= \sum_{h_1} \phi_A(h_1, h_2 = 1) && \text{(S.43)}\\ &= 0.2 && \text{(S.44)}\end{aligned}$$

Since v_2 is not observed and $p(v_2|h_2)$ normalised, the message from $p(v_2|h_2)$ to h_2 equals one.

This means that the message from h_2 to $p(h_3|h_2)$, which is $\alpha(h_2)$ equals $\mu_{\phi_A \to h_2}(h_2)$, i.e.

$$\begin{aligned}\alpha(h_2 = 0) &= 0.3 && \text{(S.45)}\\ \alpha(h_2 = 1) &= 0.2 && \text{(S.46)}\end{aligned}$$

(e) *With $\alpha(h_2)$ defined as above, use message passing to show that the predictive probability $p(v_3 = 1|v_1 = 1)$ can be expressed in terms of $\alpha(h_2)$ as*

$$p(v_3 = 1|v_1 = 1) = \frac{x\alpha(h_2 = 1) + y\alpha(h_2 = 0)}{\alpha(h_2 = 1) + \alpha(h_2 = 0)} \tag{4}$$

and report the values of x and y.

Solution. Given the definition of $p(h_3|h_2)$, the message $\mu_{p(h_3|h_2)\to h_3}(h_3)$ is

$$\mu_{p(h_3|h_2)\to h_3}(h_3 = 0) = \alpha(h_2 = 1) \tag{S.47}$$
$$\mu_{p(h_3|h_2)\to h_3}(h_3 = 1) = \alpha(h_2 = 0) \tag{S.48}$$

The variable node h_3 copies the message so that we have

$$\begin{aligned}
\mu_{p(v_3|h_3)\to v_3}(v_3 = 0) &= \sum_{h_3} p(v_3 = 0|h_3)\mu_{p(h_3|h_2)\to h_3}(h_3) && \text{(S.49)}\\
&= p(v_3 = 0|h_3 = 0)\alpha(h_2 = 1) + p(v_3 = 0|h_3 = 1)\alpha(h_2 = 0) && \text{(S.50)}\\
&= 0.6\alpha(h_2 = 1) + 0.4\alpha(h_2 = 0) && \text{(S.51)}\\
\mu_{p(v_3|h_h3)\to v_3}(v_3 = 1) &= \sum_{h_3} p(v_3 = 1|h_3))\mu_{p(h_3|h_2)\to h_3}(h_3) && \text{(S.52)}\\
&= p(v_3 = 1|h_3 = 0)\alpha(h_2 = 1) + p(v_3 = 1|h_3 = 1)\alpha(h_2 = 0) && \text{(S.53)}\\
&= 0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0) && \text{(S.54)}
\end{aligned}$$

We thus have

$$\begin{aligned}
p(v_3 = 1|v_1 = 1) &= \frac{0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0)}{0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0) + 0.6\alpha(h_2 = 1) + 0.4\alpha(h_2 = 0)} && \text{(S.55)}\\
&= \frac{0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0)}{\alpha(h_2 = 1) + \alpha(h_2 = 0)} && \text{(S.56)}
\end{aligned}$$

The requested x and y are thus: $x = 0.4$, $y = 0.6$.

(f) *Compute the numerical value of* $p(v_3 = 1|v_1 = 1)$.

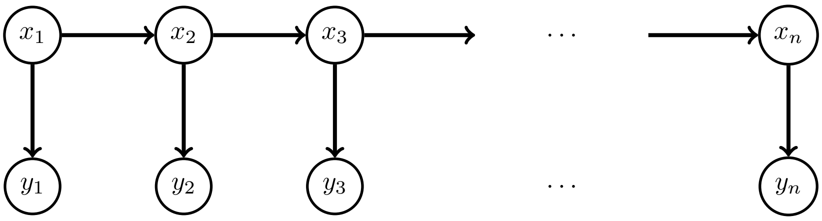
Solution. Inserting the numbers gives $\alpha(h_2 = 0) + \alpha(h_2 = 1) = 5/10 = 1/2$ so that

$$\begin{aligned}
p(v_3 = 1|v_1 = 1) &= \frac{0.4 \cdot 0.2 + 0.6 \cdot 0.3}{\frac{1}{2}} && \text{(S.57)}\\
&= 2 \cdot \left(\frac{4}{10} \cdot \frac{2}{10} + \frac{6}{10}\frac{3}{10}\right) && \text{(S.58)}\\
&= \frac{4}{10} \cdot \frac{4}{10} + \frac{6}{10}\frac{6}{10} && \text{(S.59)}\\
&= \frac{1}{100}(16 + 36) && \text{(S.60)}\\
&= \frac{1}{100}52 && \text{(S.61)}\\
&= \frac{52}{100} = 0.52 && \text{(S.62)}
\end{aligned}$$

Exercise 5. *Hidden Markov models and change of measure*

We take here a change of measure perspective on the alpha-recursion.

Consider the following directed graph for a hidden Markov model where the y_i *correspond to observed (visible) variables and the* x_i *to unobserved (hidden/latent) variables.*



The joint model for $\mathbf{x} = (x_1, \ldots, x_n)$ *and* $\mathbf{y} = (y_1, \ldots, y_n)$ *thus is*

$$p(\mathbf{x}, \mathbf{y}) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) \prod_{i=1}^{n} p(y_i | x_i). \tag{5}$$

(a) *Show that*

$$p(x_1, \ldots, x_n, y_1, \ldots, y_t) = f_1(x_1) \prod_{i=2}^{n} f_i(x_i | x_{i-1}) \prod_{i=1}^{t} p(y_i | x_i) \tag{6}$$

for $t = 0, \ldots, n$. *We take the case* $t = 0$ *to correspond to* $p(x_1, \ldots, x_n)$,

$$p(x_1, \ldots, x_n) = f_1(x_1) \prod_{i=2}^{n} f_i(x_i | x_{i-1}). \tag{7}$$

Solution. The result follows by integrating/summing out $y_{t+1} \ldots n$.

$$p(x_1, \ldots, x_n, y_1, \ldots, y_t) = \int p(x_1, \ldots, x_n, y_1, \ldots, y_n) \mathrm{d}y_{t+1} \ldots \mathrm{d}y_n \tag{S.63}$$

$$= \int p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) \prod_{i=1}^{n} p(y_i | x_i) \mathrm{d}y_{t+1} \ldots \mathrm{d}y_n \tag{S.64}$$

$$= p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) \prod_{i=1}^{t} p(y_i | x_i) \int \prod_{i=t+1}^{n} p(y_i | x_i) \mathrm{d}y_{t+1} \ldots \mathrm{d}y_n \tag{S.65}$$

$$= p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) \prod_{i=1}^{t} p(y_i | x_i) \prod_{i=t+1}^{n} \underbrace{\int p(y_i | x_i) \mathrm{d}y_i}_{=1} \tag{S.66}$$

$$= p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) \prod_{i=1}^{t} p(y_i | x_i) \tag{S.67}$$

The result for $p(x_1, \ldots, x_n)$ is obtained when we integrate out all y's.

(b) *Show that* $p(x_1, \ldots, x_n | y_1, \ldots, y_t)$, $t = 0, \ldots, n$, *factorises as*

$$p(x_1, \ldots, x_n | y_1, \ldots, y_t) \propto p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) \prod_{i=1}^{t} g_i(x_i) \tag{8}$$

where $g_i(x_i) = p(y_i | x_i)$ *for a fixed value of* y_i, *and that its normalising constant* Z_t *equals the likelihood* $p(y_1, \ldots, y_t)$

Solution. The result follows from the basic definition of the conditional

$$p(x_1, \ldots, x_n | y_1, \ldots, y_t) = \frac{p(x_1, \ldots, x_n, y_1, \ldots, y_t)}{p(y_1, \ldots, y_t)} \tag{S.68}$$

together with the expression for $p(x_1, \ldots, x_n, y_1, \ldots, y_t)$ when the y_i are kept fixed.

(c) *Denote $p(x_1, \ldots, x_n | y_1, \ldots, y_t)$ by $p_t(x_1, \ldots, x_n)$. The index $t \leq n$ thus indicates the time of the last y-variable we are conditioning on. Show the following recursion for $1 \leq t \leq n$:*

$$p_{t-1}(x_1, \ldots, x_t) = \begin{cases} p(x_1) & \textit{if } t = 1 \\ p_{t-1}(x_1, \ldots, x_{t-1}) p(x_t | x_{t-1}) & \textit{otherwise} \end{cases} \quad \textit{(extension)} \tag{9}$$

$$p_t(x_1, \ldots, x_t) = \frac{1}{Z_t} p_{t-1}(x_1, \ldots, x_t) g_t(x_t) \quad \textit{(change of measure)} \tag{10}$$

$$Z_t = \int p_{t-1}(x_t) g_t(x_t) \mathrm{d}x_t \tag{11}$$

By iterating from $t = 1$ to $t = n$, we can thus recursively compute $p(x_1, \ldots, x_n | y_1, \ldots, y_n)$, including its normalising constant Z_n, which equals the likelihood $Z_n = p(y_1, \ldots, y_n)$

Solution. We start with (8) which shows that by definition of $p_t(x_1, \ldots, x_n)$ we have

$$p_t(x_1, \ldots, x_n) = p(x_1, \ldots, x_n | y_1, \ldots, y_t) \tag{S.69}$$

$$\propto p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) \prod_{i=1}^{t} g_i(x_i) \tag{S.70}$$

For $t = 1$, we thus have

$$p_1(x_1, \ldots, x_n) \propto p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) g_1(x_1) \tag{S.71}$$

Integrating out $x_2, \ldots, x_n$ gives

$$p_1(x_1) = \int p_1(x_1, \ldots, x_n) \mathrm{d}x_2 \ldots \mathrm{d}x_n \tag{S.72}$$

$$\propto \int p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) g_1(x_1) \mathrm{d}x_2 \ldots \mathrm{d}x_n \tag{S.73}$$

$$\propto p(x_1) g_1(x_1) \int \prod_{i=2}^{n} p(x_i | x_{i-1}) \mathrm{d}x_2 \ldots \mathrm{d}x_n \tag{S.74}$$

$$\propto p(x_1) g_1(x_1) \underbrace{\prod_{i=2}^{n} \int p(x_i | x_{i-1}) \mathrm{d}x_i}_{=1} \tag{S.75}$$

$$\propto p(x_1) g_1(x_1) \tag{S.76}$$

The normalising constant is

$$Z_1 = \int p(x_1) g_1(x_1) \mathrm{d}x_1 \tag{S.77}$$

This establishes the result for $t = 1$.

From (8), we further have

$$
\begin{aligned}
p_{t-1}(x_1,\dots,x_n) &= p(x_1,\dots,x_n|y_1,\dots,y_{t-1}) && \text{(S.78)}\\
&\propto p(x_1)\prod_{i=2}^{n} p(x_i|x_{i-1})\prod_{i=1}^{t-1} g_i(x_i) && \text{(S.79)}
\end{aligned}
$$

Integrating out $x_{t+1},\dots,x_n$ thus gives

$$
\begin{aligned}
p_{t-1}(x_1,\dots,x_t) &= \int p_{t-1}(x_1,\dots,x_n)\mathrm{d}x_{t+1}\dots\mathrm{d}x_n && \text{(S.80)}\\
&\propto \int p(x_1)\prod_{i=2}^{n} p(x_i|x_{i-1})\prod_{i=1}^{t-1} g_i(x_i)\mathrm{d}x_{t+1}\dots\mathrm{d}x_n && \text{(S.81)}\\
&\propto p(x_1)\prod_{i=2}^{t} p(x_i|x_{i-1})\prod_{i=1}^{t-1} g_i(x_i)\int\prod_{i=t+1}^{n} p(x_i|x_{i-1})\mathrm{d}x_{t+1}\dots\mathrm{d}x_n && \text{(S.82)}\\
&\propto p(x_1)\prod_{i=2}^{t} p(x_i|x_{i-1})\prod_{i=1}^{t-1} g_i(x_i)\prod_{i=t+1}^{n}\int p(x_i|x_{i-1})\mathrm{d}x_i && \text{(S.83)}\\
&\propto p(x_1)\prod_{i=2}^{t} p(x_i|x_{i-1})\prod_{i=1}^{t-1} g_i(x_i) && \text{(S.84)}
\end{aligned}
$$

Noting that the product over the g_i does not involve x_t and that $p(x_t|x_{t-1})$ is a pdf, we have further

$$
\begin{aligned}
p_{t-1}(x_1,\dots,x_{t-1}) &= \int p_{t-1}(x_1,\dots,x_t)\mathrm{d}x_t && \text{(S.85)}\\
&\propto p(x_1)\prod_{i=2}^{t-1} p(x_i|x_{i-1})\prod_{i=1}^{t-1} g_i(x_i) && \text{(S.86)}
\end{aligned}
$$

Hence

$$
p_{t-1}(x_1,\dots,x_t) = p_{t-1}(x_1,\dots,x_{t-1})p(x_t|x_{t-1}) \qquad \text{(S.87)}
$$

Note that we can have an equal sign since $p(x_t|x_{t-1})$ is a pdf and hence integrates to one. This is sometimes called the “extension” since the inputs for p_{t-1} are extended from $(x_1,\dots,x_{t-1})$ to $x_1,\dots,x_t$.

From (S.70), we further have

$$
p_t(x_1,\dots,x_n) \propto p_{t-1}(x_1,\dots,x_n)g_t(x_t) \qquad \text{(S.88)}
$$

Integrating out $x_{t+1},\dots,x_n$ thus gives

$$
p_t(x_1,\dots,x_t) \propto p_{t-1}(x_1,\dots,x_t)g_t(x_t) \qquad \text{(S.89)}
$$

This is a change of measure from $p_{t-1}(x_1,\dots,x_t)$ to $p_t(x_1,\dots,x_t)$. Note that $p_{t-1}(x_1,\dots,x_t)$ only involves g_i, and hence observations y_i, up to index (time) $t-1$. The change of measure multiplies-in the additional factor $g_t(x_t) = p(y_t|x_t)$, and thereby incorporates the observation at index (time) t into the model.

The stated recursion is complete by computing the normalising constant Z_t for $p_t(x_1, \ldots, x_t)$, which equals

$$Z_t = \int p_{t-1}(x_1, \ldots, x_t) g_t(x_t) \mathrm{d}x_1, \ldots \mathrm{d}x_t \tag{S.90}$$

$$= \int g_t(x_t) \left[\int p_{t-1}(x_1, \ldots, x_t) \mathrm{d}x_1, \ldots \mathrm{d}x_{t-1} \right] \mathrm{d}x_t \tag{S.91}$$

$$= \int g_t(x_t) p_{t-1}(x_t) \mathrm{d}x_t \tag{S.92}$$

This recursion, and some slight generalisations, forms the basis for what is known as the "forward recursion" in particle filtering and sequential Monte Carlo. These topics are out of scope of the course but an excellent introduction would be the book An Introduction to Sequential Monte Carlo by Chopin and Papaspiliopoulos.

(d) *Use the recursion above to derive the following form of the alpha recursion:*

$$p_{t-1}(x_{t-1}, x_t) = p_{t-1}(x_{t-1}) p(x_t | x_{t-1}) \qquad \textit{(extension)} \tag{12}$$

$$p_{t-1}(x_t) = \int p_{t-1}(x_{t-1}, x_t) \mathrm{d}x_{t-1} \qquad \textit{(marginalisation)} \tag{13}$$

$$p_t(x_t) = \frac{1}{Z_t} p_{t-1}(x_t) g_t(x_t) \qquad \textit{(change of measure)} \tag{14}$$

$$Z_t = \int p_{t-1}(x_t) g_t(x_t) \mathrm{d}x_t \tag{15}$$

with $p_0(x_1) = p(x_1)$.

The term $p_t(x_t)$ *corresponds to* $\alpha(x_t)$ *from the alpha-recursion after normalisation. As in the lecture, we see that* $p_{t-1}(x_t)$ *is a predictive distribution for* x_t *given observations until time* $t-1$. *Multiplying* $p_{t-1}(x_t)$ *with* $g_t(x_t)$ *gives the new* $\alpha(x_t)$. *In the lecture we called* $g_t(x_t) = p(y_t|x_t)$ *the "correction". We see here that the correction has the effect of a change of measure, changing the predictive distribution* $p_{t-1}(x_t)$ *into the filtering distribution* $p_t(x_t)$.

Solution. Let $t > 1$. With (9), we have

$$p_{t-1}(x_{t-1}, x_t) = \int p_{t-1}(x_1, \ldots, x_t) \mathrm{d}x_1 \ldots \mathrm{d}x_{t-2} \tag{S.93}$$

$$= \int p_{t-1}(x_1, \ldots, x_{t-1}) p(x_t | x_{t-1}) \mathrm{d}x_1 \ldots \mathrm{d}x_{t-2} \tag{S.94}$$

$$= p(x_t | x_{t-1}) \int p_{t-1}(x_1, \ldots, x_{t-1}) \mathrm{d}x_1 \ldots \mathrm{d}x_{t-2} \tag{S.95}$$

$$= p(x_t | x_{t-1}) p_{t-1}(x_{t-1}) \tag{S.96}$$

which proves the "extension".

With (10), we have

$$p_t(x_t) = \int p_t(x_1, \ldots, x_t) \mathrm{d}x_1, \ldots \mathrm{d}x_{t-1} \tag{S.97}$$

$$= \frac{1}{Z_t} \int p_{t-1}(x_1, \ldots, x_t) g_t(x_t) \mathrm{d}x_1, \ldots \mathrm{d}x_{t-1} \tag{S.98}$$

$$= \frac{1}{Z_t} g_t(x_t) \int p_{t-1}(x_1, \ldots, x_t) \mathrm{d}x_1, \ldots \mathrm{d}x_{t-1} \tag{S.99}$$

$$= \frac{1}{Z_t} g_t(x_t) p_{t-1}(x_t) \tag{S.100}$$

which proves the "change of measure". Moreover, the normalising constant Z_t is the same as before. Hence completing the iteration until $t = n$ yields the likelihood $p(y_1, \dots, y_n) = Z_n$ as a by-product of the recursion. The initialisation of the recursion with $p_0(x_1) = p(x_1)$ is also the same as above.

Exercise 6. *Reject option*

[Murphy PML1 (2022) Ex 5.1] Consider a K-class discrete variable Y with labels $\mathcal{Y} = \{1, \dots, C\}$. The actions are $\mathcal{A} = \mathcal{Y} \cup \{0\}$, where $a = 0$ denotes the reject option, and choosing action $a = i$ for $i \in \mathcal{Y}$ denotes selecting label i. Define the loss function as follows:

$$\ell(y = j, a = i) = \begin{cases} 0 & \textit{if } i = j \textit{ and } a \in \{1, \dots, C\} \\ \lambda_r & \textit{if } a = 0 \\ \lambda_c & \textit{otherwise,} \end{cases} \tag{16}$$

where λ_r is the cost of a reject, λ_c the cost of an error.

Given information $\mathbf{x}$ we obtain the posterior $p(Y|\mathbf{x})$. Show the the minimum risk is obtained if we decide $Y = j$ if $p(Y = j|\mathbf{x}) \geq p(Y = k|\mathbf{x})$ for all k (i.e. j is the most probable label) and *if $p(Y = j|\mathbf{x}) \geq 1 - \lambda_r/\lambda_c$, otherwise we decide to reject.*

Solution. The risk for the action $a = i$ for $i = 1, \dots, C$ is given by

$$\begin{aligned} R(a = i|\mathbf{x}) &= \sum_k p(Y = k|\mathbf{x}) \ell(Y = k, a = i) \\ &= p(Y = i|\mathbf{x}) \cdot 0 + \lambda_c \sum_{k \neq i} p(Y = k|\mathbf{x}) \\ &= \lambda_c(1 - p(Y = i|\mathbf{x})). \end{aligned}$$

This risk is minimized by choosing the most probable label j. The risk $R(a = 0|\mathbf{x}) = \lambda_r$. We choose the reject option when

$$\begin{aligned} R(a = 0|\mathbf{x}) &\leq R(a = j|\mathbf{x}), \text{i.e.} \\ \lambda_r &\leq \lambda_c(1 - p(Y = j|\mathbf{x})), \end{aligned}$$

which can be rearranged to give $p(Y = j|\mathbf{x}) \geq 1 - \lambda_r/\lambda_c$.

Exercise 7. *Utility of money*

[Based on Koller and Friedman (2009) sec 22.2.2]. Suppose that your utility $U(x)$ for having a bank balance of £x is given by $U(x) = \log_{10}(1 + x)$. Sketch the utility curve.

Suppose you have £50, and are offered a bet π where you double your money with probability 0.5, and lose it all with probability 0.5.

Compute the expected utility of the bet. Show graphically on your sketch graph why the bet has an inferior expected utility compared to having £50 with certainty.

Suppose now that the bet is modified so that you obtain £100 with probability $p > 0.5$, and £0 with probability $1 - p$. Determine p so that your expected utility is indifferent to making the bet or not. Identify p on your sketch graph.

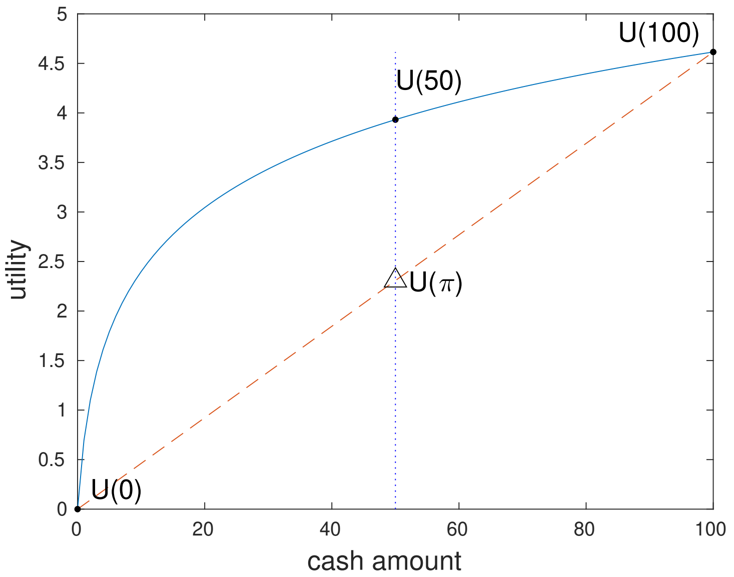


Figure 1: Utility plotted against cash value x for $U(x) = \log_{10}(1 + x)$. The dashed line interpolates between $U(0)$ and $U(100)$. $U(\pi)$ denotes the utility of the 50/50 bet.

Solution. For the bet we have

$$U(\pi) = 0.5 \cdot U(100) + 0.5 \cdot U(0) = 0.5 \cdot 2.0043 + 0 = 1.0022.$$

Otherwise without betting we have $U(50) = 1.7076$. The value of $U(\pi)$ lies halfway along the diagonal line joining $U(0) = 0$ and $U(100)$, and clearly lies below $U(50)$. This follows from the concavity of the utility function.

For the last part we require that $p \cdot U(100) = U(50)$, so $p = U(50)/U(100) = 0.8519$. p can be identified graphically by drawing a horizontal line from $U(50)$ to meet the diagonal line. Let the corresponding cash value of this intercept be c. Then $p = c/100$.

The solution is illustrated in Fig. 1.

Exercise 8. *Kalman filtering (optional, not examinable)*

We here consider filtering for hidden Markov models with Gaussian transition and emission distributions. For simplicity, we assume one-dimensional hidden variables and observables. We denote the probability density function of a Gaussian random variable x with mean μ and variance σ^2 by $\mathcal{N}(x|\mu, \sigma^2)$,

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{17}$$

The transition and emission distributions are assumed to be

$$p(h_s|h_{s-1}) = \mathcal{N}(h_s|A_s h_{s-1}, B_s^2) \tag{18}$$
$$p(v_s|h_s) = \mathcal{N}(v_s|C_s h_s, D_s^2). \tag{19}$$

The distribution $p(h_1)$ is assumed Gaussian with known parameters. The A_s, B_s, C_s, D_s are also assumed known.

(a) *Show that h_s and v_s as defined in the following update and observation equations*

$$h_s = A_s h_{s-1} + B_s \xi_s \tag{20}$$
$$v_s = C_s h_s + D_s \eta_s \tag{21}$$

follow the conditional distributions in (18) and (19). The random variables ξ_s and η_s are independent from the other variables in the model and follow a standard normal Gaussian distribution, e.g. $\xi_s \sim \mathcal{N}(\xi_s|0,1)$.

Hint: For two constants c_1 and c_2, $y = c_1 + c_2 x$ is Gaussian if x is Gaussian. In other words, an affine transformation of a Gaussian is Gaussian.

The equations mean that h_s is obtained by scaling h_{s-1} and by adding noise with variance B_s^2. The observed value v_s is obtained by scaling the hidden h_s and by corrupting it with Gaussian observation noise of variance D_s^2.

Solution. By assumption, ξ_s is Gaussian. Since we condition on h_{s-1}, $A_s h_{s-1}$ in (20) is a constant, and since B_s is a constant too, h_s is Gaussian.

What we have to show next is that (20) defines the same conditional mean and variance as the conditional Gaussian in (18): The conditional expectation of h_s given h_{s-1} is

$$\begin{aligned}
\mathbb{E}(h_s|h_{s-1}) &= A_s h_{s-1} + \mathbb{E}(B_s \xi_s) && \text{(since we condition on } h_{s-1}\text{)} && \text{(S.101)}\\
&= A_s h_{s-1} + B_s \mathbb{E}(\xi_s) && \text{(by linearity of expectation)} && \text{(S.102)}\\
&= A_s h_{s-1} && \text{(since } \xi_s \text{ has zero mean)} && \text{(S.103)}
\end{aligned}$$

The conditional variance of h_s given h_{s-1} is

$$\begin{aligned}
\mathbb{V}(h_s|h_{s-1}) &= \mathbb{V}(B_s \xi_s) && \text{(since we condition on } h_{s-1}\text{)} && \text{(S.104)}\\
&= B_s^2 \mathbb{V}(\xi_s) && \text{(by properties of the variance)} && \text{(S.105)}\\
&= B_s^2 && \text{(since } \xi_s \text{ has variance one)} && \text{(S.106)}
\end{aligned}$$

We see that the conditional mean and variance of h_s given h_{s-1} match those in (18). And since h_s given h_{s-1} is Gaussian as argued above, the result follows.

Exactly the same reasoning also applies to the case of (21). Conditional on h_s, v_s is Gaussian because it is an affine transformation of a Gaussian. The conditional mean of v_s given h_s is:

$$\begin{aligned}
\mathbb{E}(v_s|h_s) &= C_s h_s + \mathbb{E}(D_s \eta_s) && \text{(since we condition on } h_s\text{)} && \text{(S.107)}\\
&= C_s h_s + D_s \mathbb{E}(\eta_s) && \text{(by linearity of expectation)} && \text{(S.108)}\\
&= C_s h_s && \text{(since } \eta_s \text{ has zero mean)} && \text{(S.109)}
\end{aligned}$$

The conditional variance of v_s given h_s is

$$\begin{aligned}
\mathbb{V}(v_s|h_s) &= \mathbb{V}(D_s \eta_s) && \text{(since we condition on } h_s\text{)} && \text{(S.110)}\\
&= D_s^2 \mathbb{V}(\eta_s) && \text{(by properties of the variance)} && \text{(S.111)}\\
&= D_s^2 && \text{(since } \eta_s \text{ has variance one)} && \text{(S.112)}
\end{aligned}$$

Hence, conditional on h_s, v_s is Gaussian with mean and variance as in (19).

(b) *Show that*

$$\int \mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(y|Ax,B^2)\mathrm{d}x \propto \mathcal{N}(y|A\mu, A^2\sigma^2 + B^2) \tag{22}$$

Hint: While this result can be obtained by integration, an approach that avoids this is as follows: First note that $\mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(y|Ax,B^2)$ is proportional to the joint pdf of x and y. We can thus consider the integral to correspond to the computation of the marginal of y from the joint. Using the equivalence of Equations (18)-(19) and (20)-(21), and the fact that the weighted sum of two Gaussian random variables is a Gaussian random variable then allows one to obtain the result.

Solution. We follow the procedure outlined above. The two Gaussian densities correspond to the equations

$$x = \mu + \sigma\xi \tag{S.113}$$

$$y = Ax + B\eta \tag{S.114}$$

where ξ and η are independent standard normal random variables. The mean of y is

$$\mathbb{E}(y) = A\mathbb{E}(x) + B\mathbb{E}(\eta) \tag{S.115}$$

$$= A\mu \tag{S.116}$$

where we have use the linearity of expectation and $\mathbb{E}(\eta) = 0$. The variance of y is

$$\mathbb{V}(y) = \mathbb{V}(Ax) + \mathbb{V}(B\eta) \qquad \text{(since } x \text{ and } \eta \text{ are independent)} \tag{S.117}$$

$$= A^2\mathbb{V}(x) + B^2\mathbb{V}(\eta) \qquad \text{(by properties of the variance)} \tag{S.118}$$

$$= A^2\sigma^2 + B^2 \tag{S.119}$$

Since y is the (weighted) sum of two Gaussians, it is Gaussian itself, and hence its distribution is completely defined by its mean and variance, so that

$$y \sim \mathcal{N}(y|A\mu, A^2\sigma^2 + B^2). \tag{S.120}$$

Now, the product $\mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(y|Ax,B^2)$ is proportional to the joint pdf of x and y, so that the integral can be considered to correspond to the marginalisation of x, and hence its result is proportional to the density of y, which is $\mathcal{N}(y|A\mu, A^2\sigma^2 + B^2)$.

(c) *Show that*

$$\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2) \propto \mathcal{N}(x|m_3,\sigma_3^2) \tag{23}$$

where

$$\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{24}$$

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(m_2 - m_1) \tag{25}$$

Hint: Work in the negative log domain.

Solution. We show the result using a classical technique called "completing the square", see e.g. https://en.wikipedia.org/wiki/Completing_the_square.

We work in the (negative) log-domain and use that

$$-\log\left[\mathcal{N}(x|m,\sigma^2)\right] = \frac{(x-m)^2}{2\sigma^2} + \text{const} \tag{S.121}$$

$$= \frac{x^2}{2\sigma^2} - x\frac{m}{\sigma^2} + \frac{m^2}{2\sigma^2} + \text{const} \tag{S.122}$$

$$= \frac{x^2}{2\sigma^2} - x\frac{m}{\sigma^2} + \text{const} \tag{S.123}$$

where const indicates terms not depending on x. We thus obtain

$$
\begin{aligned}
-\log\left[\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2)\right] &= -\log\left[\mathcal{N}(x|m_1,\sigma_1^2)\right] - \log\left[\mathcal{N}(x|m_2,\sigma_2^2)\right] && \text{(S.124)}\\
&= \frac{(x-m_1)^2}{2\sigma_1^2} + \frac{(x-m_2)^2}{2\sigma_2^2} + \text{const} && \text{(S.125)}\\
&= \frac{x^2}{2\sigma_1^2} - x\frac{m_1}{\sigma_1^2} + \frac{x^2}{2\sigma_2^2} - x\frac{m_2}{\sigma_2^2} + \text{const} && \text{(S.126)}\\
&= \frac{x^2}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) - x\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) + \text{const} && \text{(S.127)}\\
&= \frac{x^2}{2\sigma_3^2} - \frac{x}{\sigma_3^2}\sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) + \text{const}, && \text{(S.128)}
\end{aligned}
$$

where

$$\frac{1}{\sigma_3^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}. \qquad \text{(S.129)}$$

Comparison with (S.123) shows that we can further write

$$\frac{x^2}{2\sigma_3^2} - \frac{x}{\sigma_3^2}\sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = \frac{(x-m_3)^2}{2\sigma_3^2} + \text{const} \qquad \text{(S.130)}$$

where

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) \qquad \text{(S.131)}$$

so that

$$-\log\left[\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2)\right] = \frac{(x-m_3)^2}{2\sigma_3^2} + \text{const} \qquad \text{(S.132)}$$

and hence

$$\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2) \propto \mathcal{N}(x|m_3,\sigma_3^2). \qquad \text{(S.133)}$$

Note that the identity

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}(m_2 - m_1) \qquad \text{(S.134)}$$

is obtained as follows

$$
\begin{aligned}
\sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) &= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) && \text{(S.135)}\\
&= m_1\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} + m_2\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} && \text{(S.136)}\\
&= m_1\left(1 - \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right) + m_2\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} && \text{(S.137)}\\
&= m_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}(m_2 - m_1) && \text{(S.138)}
\end{aligned}
$$

(d) *In the lecture, we have seen that $p(h_t|v_{1:t}) \propto \alpha(h_t)$ where $\alpha(h_t)$ can be computed recursively via the "alpha-recursion"*

$$\alpha(h_1) = p(h_1) \cdot p(v_1|h_1) \qquad\qquad \alpha(h_s) = p(v_s|h_s) \sum_{h_{s-1}} p(h_s|h_{s-1})\alpha(h_{s-1}). \tag{26}$$

For continuous random variables, the sum above becomes an integral so that

$$\alpha(h_s) = p(v_s|h_s) \int p(h_s|h_{s-1})\alpha(h_{s-1})\mathrm{d}h_{s-1}. \tag{27}$$

For reference, let us denote the integral by $I(h_s)$,

$$I(h_s) = \int p(h_s|h_{s-1})\alpha(h_{s-1})\mathrm{d}h_{s-1}. \tag{28}$$

In the lecture, it was pointed out that $I(h_s)$ is proportional to the predictive distribution $p(h_s|v_{1:s-1})$. For a Gaussian prior distribution for h_1 and Gaussian emission probability $p(v_1|h_1)$, $\alpha(h_1) = p(h_1) \cdot p(v_1|h_1) \propto p(h_1|v_1)$ is proportional to a Gaussian. We denote its mean by μ_1 and its variance by σ_1^2 so that

$$\alpha(h_1) \propto \mathcal{N}(h_1|\mu_1, \sigma_1^2). \tag{29}$$

Assuming $\alpha(h_{s-1}) \propto \mathcal{N}(h_{s-1}|\mu_{s-1}, \sigma_{s-1}^2)$ (which holds for $s = 2$), use Equation (22) to show that

$$I(h_s) \propto \mathcal{N}(h_s|A_s\mu_{s-1}, P_s) \tag{30}$$

where

$$P_s = A_s^2\sigma_{s-1}^2 + B_s^2. \tag{31}$$

Solution. We can set $\alpha(h_{s-1}) \propto \mathcal{N}(h_{s-1}|\mu_{s-1}, \sigma_{s-1}^2)$. Since $p(h_s|h_{s-1})$ is Gaussian, see Equation (18), Equation (28) becomes

$$I(h_s) \propto \int \mathcal{N}(h_s|A_s h_{s-1}, B_s^2)\mathcal{N}(h_{s-1}|\mu_{s-1}, \sigma_{s-1}^2)\mathrm{d}h_{s-1}. \tag{S.139}$$

Equation (22) with $x \equiv h_{s-1}$ and $y \equiv h_s$ yields the desired result,

$$I(h_s) \propto \mathcal{N}(h_s|A_s\mu_{s-1}, A_s^2\sigma_{s-1}^2 + B_s^2). \tag{S.140}$$

We can understand the equation as follows: To compute the predictive mean of h_s given $v_{1:s-1}$, we forward propagate the mean of $h_{s-1}|v_{1:s-1}$ using the update equation (20). This gives the mean term $A_s\mu_{s-1}$. Since $h_{s-1}|v_{1:s-1}$ has variance σ_{s-1}^2, the variance of $h_s|v_{1:s-1}$ is given by $A_s^2\sigma_{s-1}^2$ plus an additional term, B_s^2, due to the noise in the forward propagation. This gives the variance term $A_s^2\sigma_{s-1}^2 + B_s^2$.

(e) *Use Equation (23) to show that*

$$\alpha(h_s) \propto \mathcal{N}\left(h_s|\mu_s, \sigma_s^2\right) \tag{32}$$

where

$$\mu_s = A_s\mu_{s-1} + \frac{P_s C_s}{C_s^2 P_s + D_s^2}\left(v_s - C_s A_s \mu_{s-1}\right) \tag{33}$$

$$\sigma_s^2 = \frac{P_s D_s^2}{P_s C_s^2 + D_s^2} \tag{34}$$

Solution. Having computed $I(h_s)$, the final step in the alpha-recursion is

$$\alpha(h_s) = p(v_s|h_s)I(h_s) \tag{S.141}$$

With Equation (19) we obtain

$$\alpha(h_s) \propto \mathcal{N}(v_s|C_s h_s, D_s^2)\mathcal{N}(h_s|A_s\mu_{s-1}, P_s). \tag{S.142}$$

We further note that

$$\mathcal{N}(v_s|C_s h_s, D_s^2) \propto \mathcal{N}\left(h_s|C_s^{-1}v_s, \frac{D_s^2}{C_s^2}\right) \tag{S.143}$$

so that we can apply Equation (23) (with $m_1 = A\mu_{s-1}$, $\sigma_1^2 = P_s$)

$$\begin{aligned}
\alpha(h_s) &\propto \mathcal{N}\left(h_s|C_s^{-1}v_s, \frac{D_s^2}{C_s^2}\right)\mathcal{N}(h_s|A_s\mu_{s-1}, P_s) && \text{(S.144)}\\
&\propto \mathcal{N}\left(h_s, \mu_s, \sigma_s^2\right) && \text{(S.145)}
\end{aligned}$$

with

$$\begin{aligned}
\mu_s &= A_s\mu_{s-1} + \frac{P_s}{P_s + \frac{D_s^2}{C_s^2}}\left(C_s^{-1}v_s - A_s\mu_{s-1}\right) && \text{(S.146)}\\
&= A_s\mu_{s-1} + \frac{P_sC_s^2}{C_s^2P_s + D_s^2}\left(C_s^{-1}v_s - A_s\mu_{s-1}\right) && \text{(S.147)}\\
&= A_s\mu_{s-1} + \frac{P_sC_s}{C_s^2P_s + D_s^2}\left(v_s - C_sA_s\mu_{s-1}\right) && \text{(S.148)}\\
\sigma_s^2 &= \frac{P_s\frac{D_s^2}{C_s^2}}{P_s + \frac{D_s^2}{C_s^2}} && \text{(S.149)}\\
&= \frac{P_sD_s^2}{P_sC_s^2 + D_s^2} && \text{(S.150)}\\
& && \text{(S.151)}
\end{aligned}$$

(f) *Show that $\alpha(h_s)$ can be re-written as*

$$\alpha(h_s) \propto \mathcal{N}\left(h_s|\mu_s, \sigma_s^2\right) \tag{35}$$

where

$$\begin{aligned}
\mu_s &= A_s\mu_{s-1} + K_s\left(v_s - C_sA_s\mu_{s-1}\right) && (36)\\
\sigma_s^2 &= (1 - K_sC_s)P_s && (37)\\
K_s &= \frac{P_sC_s}{C_s^2P_s + D_s^2} && (38)
\end{aligned}$$

These are the Kalman filter equations and K_s is called the Kalman filter gain.

Solution. We start from

$$\mu_s = A_s\mu_{s-1} + \frac{P_sC_s}{C_s^2P_s + D_s^2}\left(v_s - C_sA_s\mu_{s-1}\right), \tag{S.152}$$

and see that

$$\frac{P_s C_s}{C_s^2 P_s + D_s^2} = K_s \tag{S.153}$$

so that

$$\mu_s = A_s \mu_{s-1} + K_s \left(v_s - C_s A_s \mu_{s-1} \right). \tag{S.154}$$

For the variance σ_s^2, we have

$$\begin{align} \sigma_s^2 &= \frac{P_s D_s^2}{P_s C_s^2 + D_s^2} \tag{S.155}\\ &= \frac{D_s^2}{P_s C_s^2 + D_s^2} P_s \tag{S.156}\\ &= \left(1 - \frac{P_s C_s^2}{P_s C_s^2 + D_s^2} \right) P_s \tag{S.157}\\ &= (1 - K_s C_s) P_s, \tag{S.158} \end{align}$$

which is the desired result.

The filtering result generalises to vector valued latents and visibles where the transition and emission distributions in (18) and (19) become

$$\begin{align} p(\mathbf{h}_s | \mathbf{h}_{s-1}) &= \mathcal{N}(\mathbf{h}_s | \mathbf{A}\mathbf{h}_{s-1}, \mathbf{\Sigma}^h), \tag{S.159}\\ p(\mathbf{v}_s | \mathbf{h}_s) &= \mathcal{N}(\mathbf{v}_s | \mathbf{C}_s \mathbf{h}_s, \mathbf{\Sigma}^v), \tag{S.160} \end{align}$$

where $\mathcal{N}()$ denotes multivariate Gaussian pdfs, e.g.

$$\mathcal{N}(\mathbf{v}_s | \mathbf{C}_s \mathbf{h}_s, \mathbf{\Sigma}^v) = \frac{1}{|\det(2\pi \mathbf{\Sigma}^v)|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{v}_s - \mathbf{C}_s \mathbf{h}_s)^\top (\mathbf{\Sigma}^v)^{-1} (\mathbf{v}_s - \mathbf{C}_s \mathbf{h}_s) \right). \tag{S.161}$$

We then have

$$p(\mathbf{h}_t | \mathbf{v}_{1:t}) = \mathcal{N}(\mathbf{h}_t | \boldsymbol{\mu}_t, \mathbf{\Sigma}_t) \tag{S.162}$$

where the posterior mean and variance are recursively computed as

$$\begin{align} \boldsymbol{\mu}_s &= \mathbf{A}_s \boldsymbol{\mu}_{s-1} + \mathbf{K}_s (\mathbf{v}_s - \mathbf{C}_s \mathbf{A}_s \boldsymbol{\mu}_{s-1}) \tag{S.163}\\ \mathbf{\Sigma}_s &= (\mathbf{I} - \mathbf{K}_s \mathbf{C}_s) \mathbf{P}_s \tag{S.164}\\ \mathbf{P}_s &= \mathbf{A}_s \mathbf{\Sigma}_{s-1} \mathbf{A}_s^\top + \mathbf{\Sigma}^h \tag{S.165}\\ \mathbf{K}_s &= \mathbf{P}_s \mathbf{C}_s^\top \left(\mathbf{C}_s \mathbf{P}_s \mathbf{C}_s^\top + \mathbf{\Sigma}^v \right)^{-1} \tag{S.166} \end{align}$$

and initialised with $\boldsymbol{\mu}_1$ and $\mathbf{\Sigma}_1$ equal to the mean and variance of $p(\mathbf{h}_1 | \mathbf{v}_1)$. The matrix $\mathbf{K}_s$ is then called the Kalman gain matrix.

The Kalman filter is widely applicable, see e.g. https://en.wikipedia.org/wiki/Kalman_filter, and has played a role in historic events such as the moon landing, see e.g. http://ieeexplore.ieee.org/document/5466132/

An example of the application of the Kalman filter to tracking is shown in Figure 2.

(g) *Explain Equation (36) in non-technical terms. What happens if the variance D_s^2 of the observation noise goes to zero?*

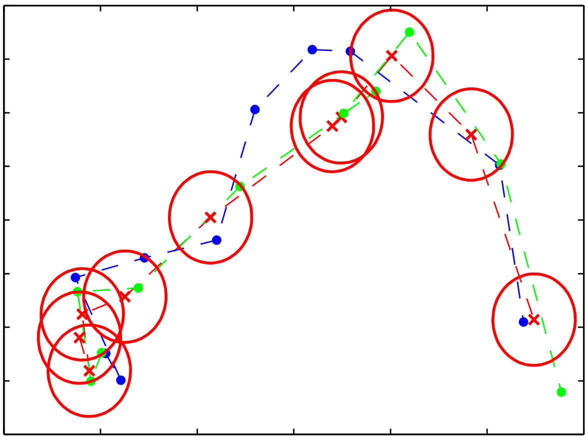


Figure 2: Kalman filtering for tracking of a moving object. The blue points indicate the true positions of the object in a two-dimensional space at successive time steps, the green points denote noisy measurements of the positions, and the red crosses indicate the means of the inferred posterior distributions of the positions obtained by running the Kalman filtering equations. The covariances of the inferred positions are indicated by the red ellipses, which correspond to contours having one standard deviation. (Bishop, Figure 13.22)

Solution. We have already seen that $A_s\mu_{s-1}$ is the predictive mean of h_s given $v_{1:s-1}$. The term $C_sA_s\mu_{s-1}$ is thus the predictive mean of v_s given the observations so far, $v_{1:s-1}$. The difference $v_s - C_sA_s\mu_{s-1}$ is thus the prediction error of the observable. Since $\alpha(h_s)$ is proportional to $p(h_s|v_{1:s})$ and μ_s its mean, we thus see that the posterior mean of $h_s|v_{1:s}$ equals the posterior mean of $h_s|v_{1:s-1}$, $A_s\mu_{s-1}$, updated by the prediction error of the observable weighted by the Kalman gain.

For $D_s^2 \to 0$, $K_s \to C_s^{-1}$ and

$$\begin{aligned}
\mu_s &= A_s\mu_{s-1} + K_s\left(v_s - C_sA_s\mu_{s-1}\right) && \text{(S.167)}\\
&= A_s\mu_{s-1} + C_s^{-1}\left(v_s - C_sA_s\mu_{s-1}\right) && \text{(S.168)}\\
&= A_s\mu_{s-1} + C_s^{-1}v_s - A_s\mu_{s-1} && \text{(S.169)}\\
&= C_s^{-1}v_s, && \text{(S.170)}
\end{aligned}$$

so that the posterior mean of $p(h_s|v_{1:s})$ is obtained by inverting the observation equation. Moreover, the variance σ_s^2 of $h_s|v_{1:s}$ goes to zero so that the value of h_s is known precisely and equals $C_s^{-1}v_s$.