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

Exercise 1. *Predictive distributions for hidden Markov models*

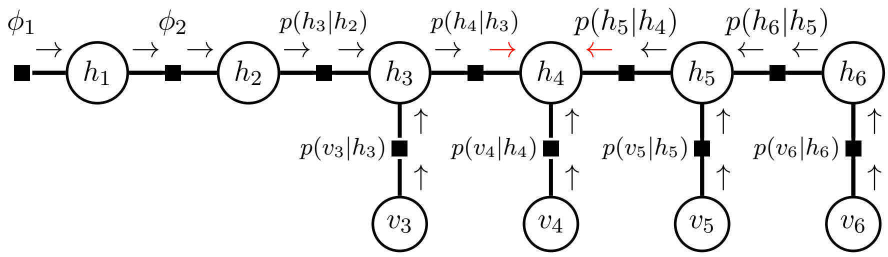
For the hidden Markov model

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

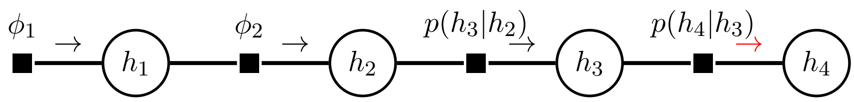
assume you have observations for v_i, $i = 1, \ldots, u < d$.

(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 $s = u+1, \ldots, t$, 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

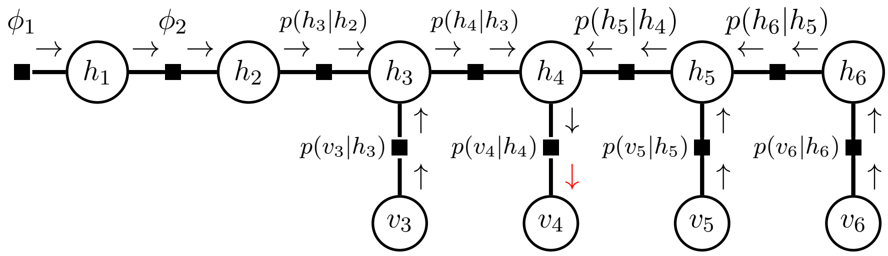
$$Z = \sum_{h_u} \alpha(h_u) = p(v_{1:u}) \tag{S.5}$$

which is the likelihood.

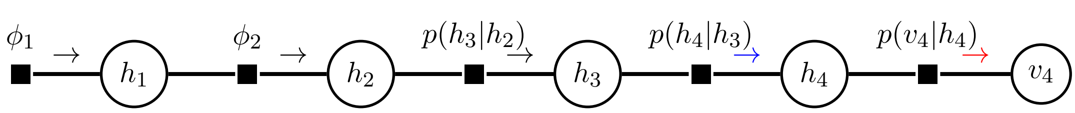
For filtering, we have seen that $\alpha(h_s) \propto p(h_s|v_{1:s})$, $s \leq u$. The $\alpha(h_s)$ for all $s > u$ are proportional to $p(h_s|v_{1:u})$. This may be seen by noting that the above arguments hold for any $t > u$.

(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:

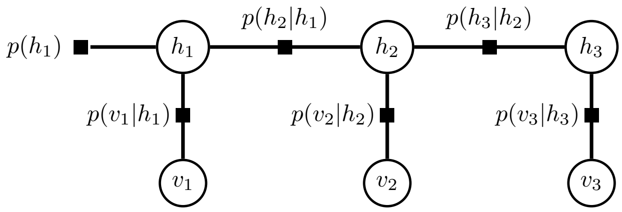
$$p(v_t|v_{1:u}) = \sum_{h_t} p(v_t, h_t|v_{1:u}) \quad \text{(sum rule)} \tag{S.7}$$

$$= \sum_{h_t} p(v_t|h_t, v_{1:u})p(h_t|v_{1:u}) \quad \text{(product rule)} \tag{S.8}$$

$$= \sum_{h_t} p(v_t|h_t)p(h_t|v_{1:u}) \quad (v_t \perp\!\!\!\perp v_{1:u} \mid h_t) \tag{S.9}$$

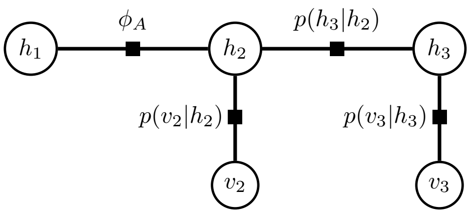
Exercise 2. ***Prediction exercise***

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

$$\phi_A(h_1 = 0, h_2 = 1) = p(v_1 = 1|h_1 = 0)p(h_1 = 0)p(h_2 = 1|h_1 = 0) \tag{S.10}$$
$$= 0.4 \cdot 0.5 \cdot 1 \tag{S.11}$$
$$= \frac{4}{10} \cdot \frac{1}{2} \tag{S.12}$$
$$= \frac{2}{10} = 0.2 \tag{S.13}$$

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

$$\phi_A(h_1 = 1, h_2 = 0) = p(v_1 = 1|h_1 = 1)p(h_1 = 1)p(h_2 = 0|h_1 = 1) \tag{S.14}$$
$$= 0.6 \cdot 0.5 \cdot 1 \tag{S.15}$$
$$= \frac{6}{10} \cdot \frac{1}{2} \tag{S.16}$$
$$= \frac{3}{10} = 0.3 \tag{S.17}$$

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

$$\mu_{\phi_A \to h_2}(h_2 = 0) = \sum_{h_1} \phi_A(h_1, h_2 = 0) \tag{S.18}$$
$$= 0.3 \tag{S.19}$$
$$\mu_{\phi_A \to h_2}(h_2 = 1) = \sum_{h_1} \phi_A(h_1, h_2 = 1) \tag{S.20}$$
$$= 0.2 \tag{S.21}$$

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.

$$\alpha(h_2 = 0) = 0.3 \tag{S.22}$$
$$\alpha(h_2 = 1) = 0.2 \tag{S.23}$$

(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*

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

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

$$\mu_{p(h_3|h_2)\to h_3}(h_3 = 0) = \alpha(h_2 = 1) \tag{S.24}$$
$$\mu_{p(h_3|h_2)\to h_3}(h_3 = 1) = \alpha(h_2 = 0) \tag{S.25}$$

The variable node h_3 copies the message so that we have

$$\mu_{p(v_3|h_3)\to v_3}(v_3 = 0) = \sum_{h_3} p(v_3 = 0|h_3)\mu_{p(h_3|h_2)\to h_3}(h_3) \tag{S.26}$$
$$= p(v_3 = 0|h_3 = 0)\alpha(h_2 = 1) + p(v_3 = 0|h_3 = 1)\alpha(h_2 = 0) \tag{S.27}$$
$$= 0.6\alpha(h_2 = 1) + 0.4\alpha(h_2 = 0) \tag{S.28}$$
$$\mu_{p(v_3|h_h3)\to v_3}(v_3 = 1) = \sum_{h_3} p(v_3 = 1|h_3))\mu_{p(h_3|h_2)\to h_3}(h_3) \tag{S.29}$$
$$= p(v_3 = 1|h_3 = 0)\alpha(h_2 = 1) + p(v_3 = 1|h_3 = 1)\alpha(h_2 = 0) \tag{S.30}$$
$$= 0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0) \tag{S.31}$$

We thus have

$$p(v_3 = 1|v_1 = 1) = \frac{0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0)}{0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0) + 0.6\alpha(h_2 = 1) + 0.4\alpha(h_2 = 0)} \tag{S.32}$$
$$= \frac{0.4\alpha(h_2 = 1) + 0.6\alpha(h_2 = 0)}{\alpha(h_2 = 1) + \alpha(h_2 = 0)} \tag{S.33}$$

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

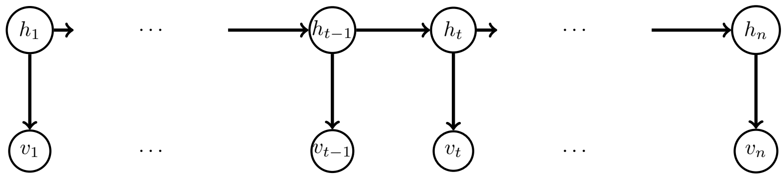
(f) *Compute the numerical value of $p(v_3 = 1|v_1 = 1)$.*

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

$$p(v_3 = 1|v_1 = 1) = \frac{0.4 \cdot 0.2 + 0.6 \cdot 0.3}{\frac{1}{2}} \tag{S.34}$$
$$= 2 \cdot \left(\frac{4}{10} \cdot \frac{2}{10} + \frac{6}{10}\frac{3}{10}\right) \tag{S.35}$$
$$= \frac{4}{10} \cdot \frac{4}{10} + \frac{6}{10}\frac{6}{10} \tag{S.36}$$
$$= \frac{1}{100}(16 + 36) \tag{S.37}$$
$$= \frac{1}{100}52 \tag{S.38}$$
$$= \frac{52}{100} = 0.52 \tag{S.39}$$

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

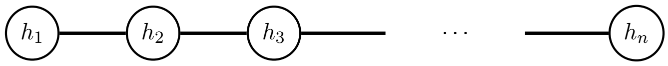
Consider the hidden Markov model specified by the following DAG.



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 $p(h_1, \ldots, h_n|v_{1:n})$, i.e. entire trajectories $(h_1, \ldots, h_n)$ 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 $p(h_1, \ldots, h_n|v_{1:n})$ 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.

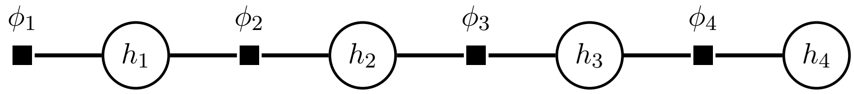


By graph separation, we see that $p(h_1, \ldots, h_n|v_{1:n})$ forms a first-order Markov chain so that e.g. $h_{1:t-1} \perp\!\!\!\perp h_{t+1:n}|h_t$ (past independent from the future given the present).

(b) *Since $p(h_1, \ldots, h_n|v_{1:n})$ 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 $p(h_1, \ldots, h_n|v_{1:n})$ as a chain factor tree, e.g. as follows in case of $n = 4$



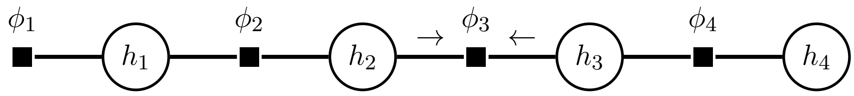
Combining the emission distributions $p(v_s|h_s)$ (and marginal $p(h_1)$) with the transition distributions $p(h_s|h_{s-1})$ we obtain the factors

$$\phi_1(h_1) = p(h_1)p(v_1|h_1) \tag{S.40}$$

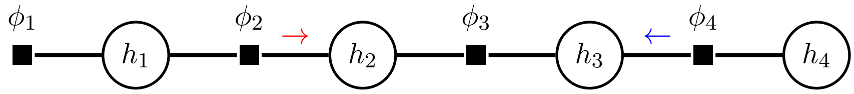
$$\phi_s(h_{s-1}, h_s) = p(h_s|h_{s-1})p(v_s|h_s) \quad \text{for } t = 2, \ldots, n \tag{S.41}$$

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,

$$p(h_{t-1}, h_t|v_{1:n}) \propto \alpha(h_{t-1})\beta(h_t)\phi_t(h_{t-1}, h_t) \tag{S.42}$$
$$\propto \alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t) \tag{S.43}$$

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. The conditional $p(h_{t-1}|h_t, v_{1:n})$ can be written as the ratio

$$p(h_{t-1}|h_t, v_{1:n}) = \frac{p(h_{t-1}, h_t|v_{1:n})}{p(h_t|v_{1:n})}. \tag{S.44}$$

Above, we have shown that the numerator satisfies

$$p(h_{t-1}, h_t|v_{1:n}) \propto \alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t). \tag{S.45}$$

The denominator $p(h_t|v_{1:n})$ is proportional to $\alpha(h_t)\beta(h_t)$ since it is the smoothing distribution than can be determined via the alpha-beta recursion.

Normally, we needed to sum the messages over all values of (h_{t-1}, h_t) to find the normalising constant of the numerator. For the denominator, we had to sum over all values of h_t. Next, I will argue qualitatively that this summation is not needed; the normalising constants are both equal to $p(v_{1:t})$. A more mathematical argument is given below.

We started with a factor graph and factors that represent the joint $p(h_{1:n}, v_{1:n})$. The conditional $p(h_{1:n}, v_{1:n})$ equals

$$p(h_{1:n}|v_{1:n}) = \frac{p(h_{1:n}, v_{1:n})}{p(v_{1:n})} \tag{S.46}$$

Message passing is variable elimination. Hence, when computing $p(h_t|v_{1:n})$ as $\alpha(h_t)\beta(h_t)$ from a factor graph for $p(h_{1:n}, v_{1:n})$, we only need to divide by $p(v_{1:n})$ for normalisation; explicitly summing out h_t is not needed. In other words,

$$p(h_t|v_{1:n}) = \frac{\alpha(h_t)\beta(h_t)}{p(v_{1:n})}. \tag{S.47}$$

Similarly, $p(h_{t-1}, h_t|v_{1:n})$ is also obtained from (S.46) by marginalisation/variable elimination. Again, when computing $p(h_{t-1}, h_t|v_{1:n})$ as $\alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t)$ from a factor graph for $p(h_{1:n}, v_{1:n})$, we do not need to explicitly sum over all values of h_t and h_{t-1}

for normalisation. The definition of the factors in the factor graph together with (S.46) shows that we can simply divide by $p(v_{1:n})$. This gives

$$p(h_{t-1}, h_t|v_{1:n}) = \frac{1}{p(v_{1:n})}\alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t). \tag{S.48}$$

The desired conditional thus is

$$\begin{aligned} p(h_{t-1}|h_t, v_{1:n}) &= \frac{p(h_{t-1}, h_t|v_{1:n})}{p(h_t|v_{1:n})} && \text{(S.49)} \\ &= \frac{\alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t)}{\alpha(h_t)\beta(h_t)} && \text{(S.50)} \\ &= \frac{\alpha(h_{t-1})p(h_t|h_{t-1})p(v_t|h_t)}{\alpha(h_t)} && \text{(S.51)} \end{aligned}$$

which is the result that we wanted to show. Note that $\beta(h_t)$ cancels out and that $p(h_{t-1}|h_t, v_{1:n})$ only involves the α's, the (forward) transition distribution $p(h_t|h_{t-1})$ and the emission distribution at time t.

Alternative solution: An alternative, mathematically rigorous solution is as follows. The conditional $p(h_{t-1}|h_t, v_{1:n})$ can be written as the ratio

$$p(h_{t-1}|h_t, v_{1:n}) = \frac{p(h_{t-1}, h_t|v_{1:n})}{p(h_t|v_{1:n})}. \tag{S.52}$$

We first determine the denominator. From the properties of the alpha and beta recursion, we know that

$$\alpha(h_t) = p(h_t, v_{1:t}) \qquad \beta(h_t) = p(v_{t+1:n}|h_t) \tag{S.53}$$

Using that $v_{t+1:n} \perp\!\!\!\perp v_{1:t}|h_t$, we can thus express the denominator $p(h_t|v_{1:n})$ as

$$\begin{aligned} p(h_t|v_{1:n}) &= \frac{p(h_t, v_{1:n})}{p(v_{1:n})} && \text{(S.54)} \\ &= \frac{p(h_t, v_{1:t})p(v_{t+1:n}|h_t)}{p(v_{1:n})} && \text{(S.55)} \\ &= \frac{\alpha(h_t)\beta(h_t)}{p(v_{1:n})} && \text{(S.56)} \end{aligned}$$

For the numerator, we have

$$
\begin{aligned}
p(h_{t-1}, h_t|v_{1:n}) &= \frac{p(h_{t-1}, h_t, v_{1:n})}{p(v_{1:n})} && \text{(S.57)}\\
&= \frac{p(h_{t-1}, v_{1:t-1}, h_t, v_{t:n})}{p(v_{1:n})} && \text{(S.58)}\\
&= \frac{p(h_{t-1}, v_{1:t-1})p(h_t, v_{t:n}|h_{t-1}, v_{1:t-1})}{p(v_{1:n})} && \text{(S.59)}\\
&= \frac{p(h_{t-1}, v_{1:t-1})p(h_t, v_{t:n}|h_{t-1})}{p(v_{1:n})} \qquad \text{(using } h_t, v_{1:t} \perp\!\!\!\perp v_{1:t-1}|h_{t-1}) && \text{(S.60)}\\
&= \frac{\alpha(h_{t-1})p(h_t, v_{t:n}|h_{t-1})}{p(v_{1:n})} \qquad \text{(using } \alpha(h_{t-1}) = p(h_{t-1}, v_{1:t-1})) && \text{(S.61)}\\
&= \frac{\alpha(h_{t-1})p(v_t|h_t, h_{t-1}, v_{t+1:n})p(h_t, v_{t+1:n}|h_{t-1})}{p(v_{1:n})} && \text{(S.62)}\\
&= \frac{\alpha(h_{t-1})p(v_t|h_t)p(h_t, v_{t+1:n}|h_{t-1})}{p(v_{1:n})} \qquad \text{(using } v_t \perp\!\!\!\perp h_{t-1}, v_{t+1:n}|h_t) && \text{(S.63)}\\
&= \frac{\alpha(h_{t-1})p(v_t|h_t)p(h_t|h_{t-1})p(v_{t+1:n}|h_{t-1}, h_t)}{p(v_{1:n})} && \text{(S.64)}\\
&= \frac{\alpha(h_{t-1})p(v_t|h_t)p(h_t|h_{t-1})p(v_{t+1:n}|h_t)}{p(v_{1:n})} \qquad \text{(using } v_{t+1:n} \perp\!\!\!\perp h_{t-1}|h_t) && \text{(S.65)}\\
&= \frac{\alpha(h_{t-1})p(v_t|h_t)p(h_t|h_{t-1})\beta(h_t)}{p(v_{1:n})} \qquad \text{(using } \beta(h_t) = p(v_{t+1:n}|h_t)) && \text{(S.66)}
\end{aligned}
$$

The desired conditional thus is

$$
\begin{aligned}
p(h_{t-1}|h_t, v_{1:n}) &= \frac{p(h_{t-1}, h_t|v_{1:n})}{p(h_t|v_{1:n})} && \text{(S.67)}\\
&= \frac{\alpha(h_{t-1})\beta(h_t)p(h_t|h_{t-1})p(v_t|h_t)}{\alpha(h_t)\beta(h_t)} && \text{(S.68)}\\
&= \frac{\alpha(h_{t-1})p(h_t|h_{t-1})p(v_t|h_t)}{\alpha(h_t)} && \text{(S.69)}
\end{aligned}
$$

which is the result that we wanted to show.

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. ***Inference for two linearly dependent Gaussian RVs (adapted from an exercise by Chris Williams)***

You have prior knowledge that an unknown variable $X \sim \mathcal{N}(X; 0, 1)$. You can make an observation of the variable Y which is related to X by $Y = wX + \mu_y + \epsilon$, with w and μ constants, and $\epsilon \sim \mathcal{N}(\epsilon; 0, \sigma^2)$ independent of X. The graphical model is $X \to Y$. (The factor w might arise e.g. because you want to measure X in centimeters, but your ruler is in inches. μ_y may arise due to an offset between the origins of the coordinates in the X and Y spaces.)

(a) *Show that $\mathbb{E}[Y] = \mu_y$.*

Solution.

$$\mathbb{E}[Y] = w\mathbb{E}[X] + \mathbb{E}[\mu_y] + \mathbb{E}[\epsilon] = 0 + \mu_y + 0 = \mu_y. \tag{S.70}$$

(b) *Show that the covariance between X and Y is $\operatorname{covar}(X, Y) = w$ and that the variance of Y equals $\mathbf{V}(y) = w^2 + \sigma^2$.*

Solution. We use that $Y = wX + \mu_y + \epsilon$, with $\epsilon \sim \mathcal{N}(\epsilon; 0, 1)$, and that $\mathbb{E}(X\epsilon) = \mathbb{E}(X)\mathbb{E}(\epsilon) = 0$ since X and ϵ are independent and $\mathbb{E}(\epsilon) = 0$.

$$\operatorname{covar}(X, Y) = \mathbb{E}[(X - 0)(Y - \mu_y)] = \mathbb{E}[X(wX + \epsilon)] = w\mathbb{V}(X) + 0 = w. \tag{S.71}$$

$$\mathbb{V}(Y) = \mathbb{E}[(Y - \mu_y)^2] = \mathbb{E}[(wX + \epsilon)^2] = w^2\mathbb{V}(X) + \mathbb{V}(\epsilon) = w^2 + \sigma^2. \tag{S.72}$$

(c) *You now want to make inferences for X given the observation $Y = y$. The conditional distribution $p(X = x|Y = y)$ is Gaussian. Compute its posterior mean and variance. HINT: You may use that if $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ is multivariate normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ partitioning into*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

then the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is multivariate Gaussian with mean and variance equal to

$$\boldsymbol{\mu}^c_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \qquad \boldsymbol{\Sigma}^c_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Alternatively, you can manipulate the expressions for the Gaussians $p(x)$ and $p(y|x)$ directly.

Solution. We compute the elements of the 2×2 covariance matrix $\boldsymbol{\Sigma}$. We have that $\Sigma_{xx} = \mathbb{V}(X) = 1$. The other entries have been computed in the question above:

$$\Sigma_{xy} = \operatorname{covar}(X, Y) = w.$$
$$\Sigma_{yy} = \mathbb{V}(Y) = w^2 + \sigma^2.$$

The conditioning formula given as hint reads in the context of this question as:

$$\mu^c_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y),$$
$$\Sigma^c_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx},$$

where $\Sigma^c_{x|y}$ denotes the conditional variance of x given y. We rename $(\mu^c_{x|y}, \Sigma^c_{x|y})$ as (μ_c, σ_c^2) for simplicity below. Substituting in and simplifying, we obtain

$$\mu_c = 0 + \frac{w}{w^2 + \sigma^2}(y - \mu_y) = \frac{w}{w^2 + \sigma^2}(y - \mu_y) \tag{S.73}$$

$$\sigma_c^2 = 1 - \frac{w^2}{w^2 + \sigma^2} = \frac{w^2 + \sigma^2}{w^2 + \sigma^2} - \frac{w^2}{(w^2 + \sigma^2)} = \frac{\sigma^2}{w^2 + \sigma^2}. \tag{S.74}$$

If there was no noise we would write $x = (y - \mu_y)/w$, and this is indeed what would be obtained for μ_c in the limit $\sigma^2 \to 0$ in (S.73). Notice that the posterior variance σ_c^2 is less than the prior variance, 1, so the observation has reduced uncertainty about X.

An alternative approach is to manipulate the expressions for the Gaussians $p(x)$ and $p(y|x)$ directly. We have $p(x|y) \propto p(x, y)$ when y is fixed. Hence

$$p(x|y) \propto p(x, y) = p(x)p(y|x) \tag{S.75}$$

$$\propto \exp\left(-\frac{1}{2}x^2\right) \cdot \exp\left(-\frac{1}{2}\frac{(y - (wx + \mu_y))^2}{\sigma^2}\right). \tag{S.76}$$

Note that the normalization factors can be omitted. The above equation is (for fixed y) a Gaussian distribution in x. We first extract the quadratic form from within the exponents (omitting the $-1/2$ factor) to give

$$Q(x) = x^2 + \frac{(y - wx - \mu_y)^2}{\sigma^2} \tag{S.77}$$

$$= x^2 + \frac{(wx + \mu_y - y)^2}{\sigma^2} \tag{S.78}$$

$$= x^2(1 + \frac{w^2}{\sigma^2}) + 2x\frac{w(\mu_y - y)}{\sigma^2} + \text{const} \tag{S.79}$$

where *const* includes any terms that do not involve powers of x.
The conditional Gaussian in x has mean μ_c and variance σ_c^2. Its quadratic form is

$$Q_c(x) = \frac{(x - \mu_c)^2}{\sigma_c^2} = \frac{x^2}{\sigma_c^2} - 2x\frac{\mu_c}{\sigma_c^2} + \text{const.} \tag{S.80}$$

By comparing the coefficients of x^2 and x we obtain

$$\frac{1}{\sigma_c^2} = \frac{\sigma^2 + w^2}{\sigma^2} \Rightarrow \sigma_c^2 = \frac{\sigma^2}{\sigma^2 + w^2}, \tag{S.81}$$

$$\frac{\mu_c}{\sigma_c^2} = \frac{w}{\sigma^2}(y - \mu_y) \Rightarrow \mu_c = \frac{w}{\sigma^2 + w^2}(y - \mu_y), \tag{S.82}$$

in agreement with (S.73) and (S.74).

Exercise 5. ***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 $\mathcal{N}(x|\mu, \sigma^2)$,

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]. \tag{4}$$

The transition and emission distributions are assumed to be

$$p(h_s|h_{s-1}) = \mathcal{N}(h_s|A_s h_{s-1}, B_s^2) \tag{5}$$

$$p(v_s|h_s) = \mathcal{N}(v_s|C_s h_s, D_s^2). \tag{6}$$

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*

$$h_s = A_s h_{s-1} + B_s \xi_s \tag{7}$$
$$v_s = C_s h_s + D_s \eta_s \tag{8}$$

follow the conditional distributions in (5) and (6). 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 (7) is a constant, and since B_s is a constant too, h_s is Gaussian.

What we have to show next is that (7) defines the same conditional mean and variance as the conditional Gaussian in (5): The conditional expectation of h_s given h_{s-1} is

$$\begin{aligned}
\mathbb{E}(h_s|h_{s-1}) &= A_s h_{s-1} + \mathbb{E}(B_s \xi_s) && \text{(since we condition on } h_{s-1}) && \text{(S.83)}\\
&= A_s h_{s-1} + B_s \mathbb{E}(\xi_s) && \text{(by linearity of expectation)} && \text{(S.84)}\\
&= A_s h_{s-1} && \text{(since } \xi_s \text{ has zero mean)} && \text{(S.85)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{V}(h_s|h_{s-1}) &= \mathbb{V}(B_s \xi_s) && \text{(since we condition on } h_{s-1}) && \text{(S.86)}\\
&= B_s^2 \mathbb{V}(\xi_s) && \text{(by properties of the variance)} && \text{(S.87)}\\
&= B_s^2 && \text{(since } \xi_s \text{ has variance one)} && \text{(S.88)}
\end{aligned}$$

We see that the conditional mean and variance of h_s given h_{s-1} match those in (5). 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 (8). 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:

$$\begin{aligned}
\mathbb{E}(v_s|h_s) &= C_s h_s + \mathbb{E}(D_s \eta_s) && \text{(since we condition on } h_s) && \text{(S.89)}\\
&= C_s h_s + D_s \mathbb{E}(\eta_s) && \text{(by linearity of expectation)} && \text{(S.90)}\\
&= C_s h_s && \text{(since } \eta_s \text{ has zero mean)} && \text{(S.91)}
\end{aligned}$$

The conditional variance of v_s given h_s is

$$\begin{aligned}
\mathbb{V}(v_s|h_s) &= \mathbb{V}(D_s \eta_s) && \text{(since we condition on } h_s) && \text{(S.92)}\\
&= D_s^2 \mathbb{V}(\eta_s) && \text{(by properties of the variance)} && \text{(S.93)}\\
&= D_s^2 && \text{(since } \eta_s \text{ has variance one)} && \text{(S.94)}
\end{aligned}$$

Hence, conditional on h_s, v_s is Gaussian with mean and variance as in (6).

(b) *Show that*

$$\int \mathcal{N}(x|\mu, \sigma^2) \mathcal{N}(y|Ax, B^2) \mathrm{d}x \propto \mathcal{N}(y|A\mu, A^2\sigma^2 + B^2) \tag{9}$$

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 (5)-(6) and (7)-(8), 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

$$x = \mu + \sigma\xi \tag{S.95}$$
$$y = Ax + B\eta \tag{S.96}$$

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

$$\mathbb{E}(y) = A\mathbb{E}(x) + B\mathbb{E}(\eta) \tag{S.97}$$
$$= A\mu \tag{S.98}$$

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

$$\mathbb{V}(y) = \mathbb{V}(Ax) + \mathbb{V}(B\eta) \qquad \text{(since } x \text{ and } \eta \text{ are independent)} \tag{S.99}$$
$$= A^2\mathbb{V}(x) + B^2\mathbb{V}(\eta) \qquad \text{(by properties of the variance)} \tag{S.100}$$
$$= A^2\sigma^2 + B^2 \tag{S.101}$$

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

$$y \sim \mathcal{N}(y|A\mu, A^2\sigma^2 + B^2). \tag{S.102}$$

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*

$$\mathcal{N}(x|m_1, \sigma_1^2)\mathcal{N}(x|m_2, \sigma_2^2) \propto \mathcal{N}(x|m_3, \sigma_3^2) \tag{10}$$

where

$$\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{11}$$
$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(m_2 - m_1) \tag{12}$$

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

$$-\log\left[\mathcal{N}(x|m, \sigma^2)\right] = \frac{(x-m)^2}{2\sigma^2} + \text{const} \tag{S.103}$$
$$= \frac{x^2}{2\sigma^2} - x\frac{m}{\sigma^2} + \frac{m^2}{2\sigma^2} + \text{const} \tag{S.104}$$
$$= \frac{x^2}{2\sigma^2} - x\frac{m}{\sigma^2} + \text{const} \tag{S.105}$$

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

$$-\log\left[\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2)\right] = -\log\left[\mathcal{N}(x|m_1,\sigma_1^2)\right] - \log\left[\mathcal{N}(x|m_2,\sigma_2^2)\right] \tag{S.106}$$

$$= \frac{(x-m_1)^2}{2\sigma_1^2} + \frac{(x-m_2)^2}{2\sigma_2^2} + \text{const} \tag{S.107}$$

$$= \frac{x^2}{2\sigma_1^2} - x\frac{m_1}{\sigma_1^2} + \frac{x^2}{2\sigma_2^2} - x\frac{m_2}{\sigma_2^2} + \text{const} \tag{S.108}$$

$$= \frac{x^2}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) - x\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) + \text{const} \tag{S.109}$$

$$= \frac{x^2}{2\sigma_3^2} - \frac{x}{\sigma_3^2}\sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) + \text{const}, \tag{S.110}$$

where

$$\frac{1}{\sigma_3^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}. \tag{S.111}$$

Comparison with (S.105) shows that we can further write

$$\frac{x^2}{2\sigma_3^2} - \frac{x}{\sigma_3^2}\sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = \frac{(x-m_3)^2}{2\sigma_3^2} + \text{const} \tag{S.112}$$

where

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) \tag{S.113}$$

so that

$$-\log\left[\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2)\right] = \frac{(x-m_3)^2}{2\sigma_3^2} + \text{const} \tag{S.114}$$

and hence

$$\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2) \propto \mathcal{N}(x|m_3,\sigma_3^2). \tag{S.115}$$

Note that the identity

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}(m_2 - m_1) \tag{S.116}$$

is obtained as follows

$$\sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) \tag{S.117}$$

$$= m_1\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} + m_2\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} \tag{S.118}$$

$$= m_1\left(1 - \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right) + m_2\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} \tag{S.119}$$

$$= m_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}(m_2 - m_1) \tag{S.120}$$

(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"*

$$\alpha(h_1) = p(h_1) \cdot p(v_1|h_1) \qquad \alpha(h_s) = p(v_s|h_s) \sum_{h_{s-1}} p(h_s|h_{s-1})\alpha(h_{s-1}). \tag{13}$$

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

$$\alpha(h_s) = p(v_s|h_s) \int p(h_s|h_{s-1})\alpha(h_{s-1})\mathrm{d}h_{s-1}. \tag{14}$$

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

$$I(h_s) = \int p(h_s|h_{s-1})\alpha(h_{s-1})\mathrm{d}h_{s-1}. \tag{15}$$

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*

$$\alpha(h_1) \propto \mathcal{N}(h_1|\mu_1, \sigma_1^2). \tag{16}$$

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 (9) to show that*

$$I(h_s) \propto \mathcal{N}(h_s|A_s\mu_{s-1}, P_s) \tag{17}$$

where

$$P_s = A_s^2\sigma_{s-1}^2 + B_s^2. \tag{18}$$

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 (5), Equation (15) becomes

$$I(h_s) \propto \int \mathcal{N}(h_s|A_s h_{s-1}, B_s^2)\mathcal{N}(h_{s-1}|\mu_{s-1}, \sigma_{s-1}^2)\mathrm{d}h_{s-1}. \tag{S.121}$$

Equation (9) with $x \equiv h_{s-1}$ and $y \equiv h_s$ yields the desired result,

$$I(h_s) \propto \mathcal{N}(h_s|A_s\mu_{s-1}, A_s^2\sigma_{s-1}^2 + B_s^2). \tag{S.122}$$

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 (7). 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 (10) to show that*

$$\alpha(h_s) \propto \mathcal{N}\left(h_s|\mu_s, \sigma_s^2\right) \tag{19}$$

where

$$\mu_s = A_s\mu_{s-1} + \frac{P_s C_s}{C_s^2 P_s + D_s^2}\left(v_s - C_s A_s \mu_{s-1}\right) \tag{20}$$

$$\sigma_s^2 = \frac{P_s D_s^2}{P_s C_s^2 + D_s^2} \tag{21}$$

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

$$\alpha(h_s) = p(v_s|h_s)I(h_s) \tag{S.123}$$

With Equation (6) we obtain

$$\alpha(h_s) \propto \mathcal{N}(v_s|C_s h_s, D_s^2)\mathcal{N}(h_s|A_s\mu_{s-1}, P_s). \tag{S.124}$$

We further note that

$$\mathcal{N}(v_s|C_s h_s, D_s^2) \propto \mathcal{N}\left(h_s|C_s^{-1}v_s, \frac{D_s^2}{C_s^2}\right) \tag{S.125}$$

so that we can apply Equation (10) (with $m_1 = A\mu_{s-1}$, $\sigma_1^2 = P_s$)

$$\alpha(h_s) \propto \mathcal{N}\left(h_s|C_s^{-1}v_s, \frac{D_s^2}{C_s^2}\right)\mathcal{N}(h_s|A_s\mu_{s-1}, P_s) \tag{S.126}$$

$$\propto \mathcal{N}\left(h_s, \mu_s, \sigma_s^2\right) \tag{S.127}$$

with

$$\mu_s = A_s\mu_{s-1} + \frac{P_s}{P_s + \frac{D_s^2}{C_s^2}}\left(C_s^{-1}v_s - A_s\mu_{s-1}\right) \tag{S.128}$$

$$= A_s\mu_{s-1} + \frac{P_s C_s^2}{C_s^2 P_s + D_s^2}\left(C_s^{-1}v_s - A_s\mu_{s-1}\right) \tag{S.129}$$

$$= A_s\mu_{s-1} + \frac{P_s C_s}{C_s^2 P_s + D_s^2}\left(v_s - C_s A_s\mu_{s-1}\right) \tag{S.130}$$

$$\sigma_s^2 = \frac{P_s \frac{D_s^2}{C_s^2}}{P_s + \frac{D_s^2}{C_s^2}} \tag{S.131}$$

$$= \frac{P_s D_s^2}{P_s C_s^2 + D_s^2} \tag{S.132}$$

$$\tag{S.133}$$

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

$$\alpha(h_s) \propto \mathcal{N}\left(h_s|\mu_s, \sigma_s^2\right) \tag{22}$$

where

$$\mu_s = A_s\mu_{s-1} + K_s\left(v_s - C_s A_s\mu_{s-1}\right) \tag{23}$$

$$\sigma_s^2 = (1 - K_s C_s)P_s \tag{24}$$

$$K_s = \frac{P_s C_s}{C_s^2 P_s + D_s^2} \tag{25}$$

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

Solution. We start from

$$\mu_s = A_s\mu_{s-1} + \frac{P_s C_s}{C_s^2 P_s + D_s^2}\left(v_s - C_s A_s\mu_{s-1}\right), \tag{S.134}$$

and see that

$$\frac{P_s C_s}{C_s^2 P_s + D_s^2} = K_s \tag{S.135}$$

so that

$$\mu_s = A_s \mu_{s-1} + K_s \left(v_s - C_s A_s \mu_{s-1} \right). \tag{S.136}$$

For the variance σ_s^2, we have

$$\begin{aligned} \sigma_s^2 &= \frac{P_s D_s^2}{P_s C_s^2 + D_s^2} && \text{(S.137)} \\ &= \frac{D_s^2}{P_s C_s^2 + D_s^2} P_s && \text{(S.138)} \\ &= \left(1 - \frac{P_s C_s^2}{P_s C_s^2 + D_s^2} \right) P_s && \text{(S.139)} \\ &= (1 - K_s C_s) P_s, && \text{(S.140)} \end{aligned}$$

which is the desired result.

The filtering result generalises to vector valued latents and visibles where the transition and emission distributions in (5) and (6) become

$$\begin{aligned} p(\mathbf{h}_s | \mathbf{h}_{s-1}) &= \mathcal{N}(\mathbf{h}_s | \mathbf{A}\mathbf{h}_{s-1}, \boldsymbol{\Sigma}^h), && \text{(S.141)} \\ p(\mathbf{v}_s | \mathbf{h}_s) &= \mathcal{N}(\mathbf{v}_s | \mathbf{C}_s \mathbf{h}_s, \boldsymbol{\Sigma}^v), && \text{(S.142)} \end{aligned}$$

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

$$\mathcal{N}(\mathbf{v}_s | \mathbf{C}_s \mathbf{h}_s, \boldsymbol{\Sigma}^v) = \frac{1}{|\det(2\pi \boldsymbol{\Sigma}^v)|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{v}_s - \mathbf{C}_s \mathbf{h}_s)^\top (\boldsymbol{\Sigma}^v)^{-1} (\mathbf{v}_s - \mathbf{C}_s \mathbf{h}_s) \right). \tag{S.143}$$

We then have

$$p(\mathbf{h}_t | \mathbf{v}_{1:t}) = \mathcal{N}(\mathbf{h}_t | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \tag{S.144}$$

where the posterior mean and variance are recursively computed as

$$\begin{aligned} \boldsymbol{\mu}_s &= \mathbf{A}_s \boldsymbol{\mu}_{s-1} + \mathbf{K}_s (\mathbf{v}_s - \mathbf{C}_s \mathbf{A}_s \boldsymbol{\mu}_{s-1}) && \text{(S.145)} \\ \boldsymbol{\Sigma}_s &= (\mathbf{I} - \mathbf{K}_s \mathbf{C}_s) \mathbf{P}_s && \text{(S.146)} \\ \mathbf{P}_s &= \mathbf{A}_s \boldsymbol{\Sigma}_{s-1} \mathbf{A}_s^\top + \boldsymbol{\Sigma}^h && \text{(S.147)} \\ \mathbf{K}_s &= \mathbf{P}_s \mathbf{C}_s^\top \left(\mathbf{C}_s \mathbf{P}_s \mathbf{C}_s^\top + \boldsymbol{\Sigma}^v \right)^{-1} && \text{(S.148)} \end{aligned}$$

and initialised with $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ 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 1.

(g) *Explain Equation (23) in non-technical terms. What happens if the variance D_s^2 of the observation noise goes to zero?*

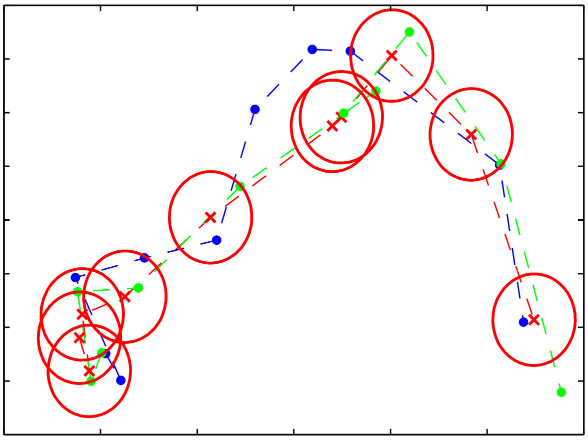


Figure 1: Kalman filtering for tracking of a moving object. The blue points indicate the true positions of the object in a two-dimensional space at successive time steps, the green points denote noisy measurements of the positions, and the red crosses indicate the means of the inferred posterior distributions 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 $C_s A_s \mu_{s-1}$ is thus the predictive mean of v_s given the observations so far, $v_{1:s-1}$. The difference $v_s - C_s A_s \mu_{s-1}$ 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

$$
\begin{aligned}
\mu_s &= A_s\mu_{s-1} + K_s\left(v_s - C_s A_s \mu_{s-1}\right) && \text{(S.149)}\\
&= A_s\mu_{s-1} + C_s^{-1}\left(v_s - C_s A_s \mu_{s-1}\right) && \text{(S.150)}\\
&= A_s\mu_{s-1} + C_s^{-1} v_s - A_s\mu_{s-1} && \text{(S.151)}\\
&= C_s^{-1} v_s, && \text{(S.152)}
\end{aligned}
$$

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 $C_s^{-1} v_s$.