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, $i = 1, \ldots, u < d$.

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

(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 $d = 6, u = 2, t = 4$.

Exercise 2. *Viterbi algorithm*

For the hidden Markov model

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

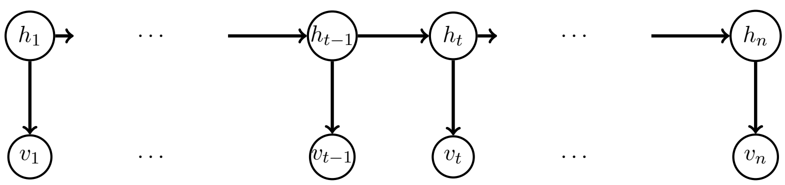
assume you have observations for v_i, $i = 1, \ldots, t$. Use the max-sum algorithm to derive an iterative algorithm to compute

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

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.

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 $(h_1, \ldots, h_n)$

©Michael U. Gutmann, UoE, 2018-22 CC BY 4.0

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.

(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

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

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

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

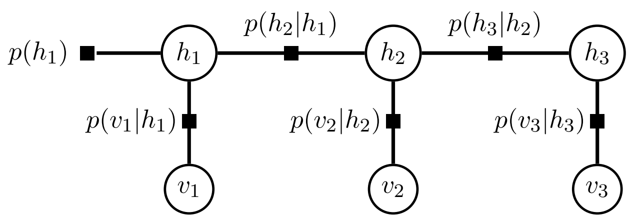
$$p(h_{t-1}|h_t, v_{1:n}) = \frac{\alpha(h_{t-1})}{\alpha(h_t)} p(h_t|h_{t-1})p(v_t|h_t) \tag{3}$$

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).

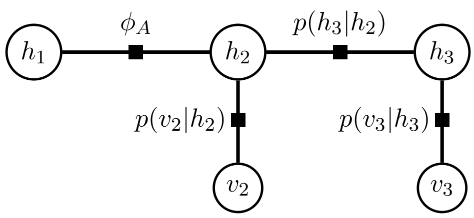
Exercise 4. ***Prediction exercise***

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.



©Michael U. Gutmann, UoE, 2018-22 CC BY 4.0

(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.

(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 $\alpha(h_2)$.

(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

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

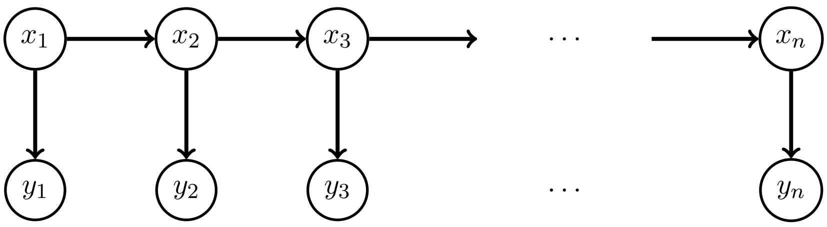
and report the values of x and y.

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

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 $p(x_1, \ldots, x_n)$,

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

©Michael U. Gutmann, UoE, 2018-22 CC BY 4.0

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

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

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 $p(y_1,\dots,y_t)$

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

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

$$p_t(x_1,\dots,x_t) = \frac{1}{Z_t}p_{t-1}(x_1,\dots,x_t)g_t(x_t) \qquad \text{(change of measure)} \tag{10}$$

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

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

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

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

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

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

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

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 $p_t(x_t)$.

Exercise 6. *Kalman filtering*

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 $\mathcal{N}(x|\mu,\sigma^2)$,

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{16}$$

The transition and emission distributions are assumed to be

$$p(h_s|h_{s-1}) = \mathcal{N}(h_s|A_s h_{s-1}, B_s^2) \tag{17}$$

$$p(v_s|h_s) = \mathcal{N}(v_s|C_s h_s, D_s^2). \tag{18}$$

©Michael U. Gutmann, UoE, 2018-22 CC BY 4.0

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

$$h_s = A_s h_{s-1} + B_s \xi_s \tag{19}$$
$$v_s = C_s h_s + D_s \eta_s \tag{20}$$

follow the conditional distributions in (17) and (18). 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.

(b) Show that

$$\int \mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(y|Ax,B^2)\mathrm{d}x \propto \mathcal{N}(y|A\mu, A^2\sigma^2 + B^2) \tag{21}$$

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 (17)-(18) and (19)-(20), and the fact that the weighted sum of two Gaussian random variables is a Gaussian random variable then allows one to obtain the result.

(c) Show that

$$\mathcal{N}(x|m_1,\sigma_1^2)\mathcal{N}(x|m_2,\sigma_2^2) \propto \mathcal{N}(x|m_3,\sigma_3^2) \tag{22}$$

where

$$\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{23}$$
$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(m_2 - m_1) \tag{24}$$

Hint: Work in the negative log domain.

(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"

$$\alpha(h_1) = p(h_1)\cdot p(v_1|h_1) \qquad\qquad \alpha(h_s) = p(v_s|h_s)\sum_{h_{s-1}} p(h_s|h_{s-1})\alpha(h_{s-1}). \tag{25}$$

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

$$\alpha(h_s) = p(v_s|h_s)\int p(h_s|h_{s-1})\alpha(h_{s-1})\mathrm{d}h_{s-1}. \tag{26}$$

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

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

©Michael U. Gutmann, UoE, 2018-22 CC BY 4.0

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

$$\alpha(h_1) \propto \mathcal{N}(h_1|\mu_1, \sigma_1^2). \tag{28}$$

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 (21) to show that

$$I(h_s) \propto \mathcal{N}(h_s|A_s\mu_{s-1}, P_s) \tag{29}$$

where

$$P_s = A_s^2\sigma_{s-1}^2 + B_s^2. \tag{30}$$

(e) Use Equation (22) to show that

$$\alpha(h_s) \propto \mathcal{N}\left(h_s|\mu_s, \sigma_s^2\right) \tag{31}$$

where

$$\mu_s = A_s\mu_{s-1} + \frac{P_sC_s}{C_s^2P_s + D_s^2}(v_s - C_sA_s\mu_{s-1}) \tag{32}$$

$$\sigma_s^2 = \frac{P_sD_s^2}{P_sC_s^2 + D_s^2} \tag{33}$$

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

$$\alpha(h_s) \propto \mathcal{N}\left(h_s|\mu_s, \sigma_s^2\right) \tag{34}$$

where

$$\mu_s = A_s\mu_{s-1} + K_s(v_s - C_sA_s\mu_{s-1}) \tag{35}$$

$$\sigma_s^2 = (1 - K_sC_s)P_s \tag{36}$$

$$K_s = \frac{P_sC_s}{C_s^2P_s + D_s^2} \tag{37}$$

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

(g) Explain Equation (35) in non-technical terms. What happens if the variance D_s^2 of the observation noise goes to zero?

©Michael U. Gutmann, UoE, 2018-22 CC BY 4.0