

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

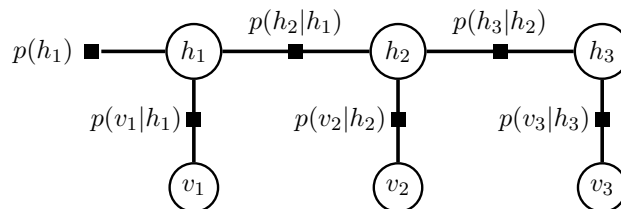
$$p(h_{1:d}, v_{1:d}) = p(v_1|h_1)p(h_1) \prod_{i=2}^d p(v_i|h_i)p(h_i|h_{i-1})$$

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

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

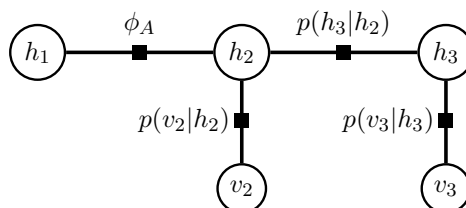
Exercise 2. *Prediction exercise*

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 $p(v_3 = 1|v_1 = 1)$.

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



- 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	$p(v_i h_i)$	v_i	h_i
0	0	0	0.6	0	0
1	1	0	0.4	1	0
1	0	1	0.4	0	1
0	1	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

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

and report the values of x and y .

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

Exercise 3. 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 $X \rightarrow Y$. (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 $\mathbb{E}[Y] = \mu_y$.
- (b) Show that the covariance between X and Y is $\text{covar}(X, Y) = w$ and that the variance of Y equals $\mathbf{V}(y) = w^2 + \sigma^2$.
- (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

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

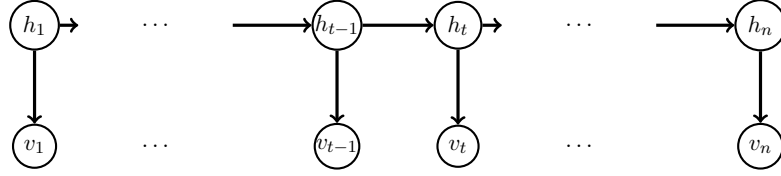
then the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is multivariate Gaussian with mean and variance equal to

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

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

Exercise 4. Forward filtering backward sampling for hidden Markov models (optional, not examinable)

Consider the hidden Markov model specified by the following DAG.



We assume that we 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.
- (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

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

- (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 $p(h_1, \dots, h_n|v_{1:n})$:

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

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

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

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

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

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

The transition and emission distributions are assumed to be

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

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

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

$$h_s = A_s h_{s-1} + B_s \xi_s \quad (7)$$

$$v_s = C_s h_s + D_s \eta_s \quad (8)$$

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 .

- (b) Show that

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

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.

- (c) Show that

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

where

$$\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1} = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (11)$$

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

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”

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

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

$$\alpha(h_s) = p(v_s|h_s) \int p(h_s|h_{s-1})\alpha(h_{s-1})dh_{s-1}. \quad (14)$$

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

$$I(h_s) = \int p(h_s|h_{s-1})\alpha(h_{s-1})dh_{s-1}. \quad (15)$$

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

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

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

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

where

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

(e) Use Equation (10) to show that

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

where

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

$$\sigma_s^2 = \frac{P_s D_s^2}{P_s C_s^2 + D_s^2} \quad (21)$$

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

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

where

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

$$\sigma_s^2 = (1 - K_s C_s) P_s \quad (24)$$

$$K_s = \frac{P_s C_s}{C_s^2 P_s + D_s^2} \quad (25)$$

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

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