Exercises for Chapter 13 - Sequential Data

abc1306

October 27, 2025

Textbook: ”Pattern Recognition and Machine Learning”, Christopher Bishop (2006), Chapter 13 - Sequential data.

Exercise 13.1

Goal: Use d-separation to verify Markov properties for first and second-order chains.

1. First-Order Markov Chain (Figure 13.3)

We need to show $p(x_n|x_1, \ldots, x_{n-1}) = p(x_n|x_{n-1})$. This requires showing that x_n is conditionally independent of $\{x_1, \ldots, x_{n-2}\}$ given x_{n-1}. Let's consider any node x_i where $i \leq n-2$.

- The only path from x_i to x_n is $x_i \to x_{i+1} \to \cdots \to x_{n-1} \to x_n$.
- We are conditioning on x_{n-1}.
- The node x_{n-1} is head-to-tail on this path ($x_{n-2} \to x_{n-1} \to x_n$).
- Since the path contains an observed, head-to-tail node (x_{n-1}), the path is **blocked**.
- Therefore, x_n is d-separated from $\{x_1, \ldots, x_{n-2}\}$ by x_{n-1}, confirming the first-order Markov property.

2. Second-Order Markov Chain (Figure 13.4)

We need to show $p(x_n|x_1, \ldots, x_{n-1}) = p(x_n|x_{n-1}, x_{n-2})$. This requires showing that x_n is conditionally independent of $\{x_1, \ldots, x_{n-3}\}$ given $\{x_{n-1}, x_{n-2}\}$. Let's consider any node x_i where $i \leq n-3$.

- There are paths from x_i to x_n. Any such path must go through either x_{n-1} or x_{n-2} (or both). Examples:
  - $x_i \to \cdots \to x_{n-2} \to x_n$.
  - $x_i \to \cdots \to x_{n-1} \to x_n$.
  - $x_i \to \cdots \to x_{n-2} \to x_{n-1} \to x_n$.
- We are conditioning on the set $\{x_{n-1}, x_{n-2}\}$.
- In all possible paths from x_i to x_n, at least one of the observed nodes x_{n-1} or x_{n-2} will be present.

- These nodes (x_{n-1}, x_{n-2}) are always head-to-tail with respect to any path segment passing through them towards x_n (e.g., $x_{n-3} \to x_{n-2} \to x_n$, $x_{n-3} \to x_{n-1} \to x_n$, $x_{n-2} \to x_{n-1} \to x_n$).
- Since every path contains an observed, head-to-tail node from the conditioning set, all paths are **blocked**.
- Therefore, x_n is d-separated from $\{x_1, \ldots, x_{n-3}\}$ by $\{x_{n-1}, x_{n-2}\}$, confirming the second-order Markov property.

Exercise 13.2

Goal: Use the sum and product rules to verify Markov properties.

1. First-Order Markov Chain

We start with the definition of conditional probability:

$$p(x_n|x_1, \ldots, x_{n-1}) = \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_{n-1})}$$

Substitute the joint distribution (13.2):

$$p(x_1, \ldots, x_N) = p(x_1) \prod_{m=2}^{N} p(x_m|x_{m-1})$$

So, the numerator is $p(x_1) \prod_{m=2}^{n} p(x_m|x_{m-1})$ and the denominator is $p(x_1) \prod_{m=2}^{n-1} p(x_m|x_{m-1})$.

$$p(x_n|x_1, \ldots, x_{n-1}) = \frac{p(x_1) \left(\prod_{m=2}^{n-1} p(x_m|x_{m-1})\right) p(x_n|x_{n-1})}{p(x_1) \prod_{m=2}^{n-1} p(x_m|x_{m-1})}$$

Cancelling terms leaves:

$$p(x_n|x_1, \ldots, x_{n-1}) = p(x_n|x_{n-1})$$

This verifies the property (13.3).

2. Second-Order Markov Chain

Again, start with the definition of conditional probability:

$$p(x_n|x_1, \ldots, x_{n-1}) = \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_{n-1})}$$

Substitute the joint distribution (13.4):

$$p(x_1, \ldots, x_N) = p(x_1)p(x_2|x_1) \prod_{m=3}^{N} p(x_m|x_{m-1}, x_{m-2})$$

Assuming $n \geq 3$, the numerator is $p(x_1)p(x_2|x_1) \prod_{m=3}^{n} p(x_m|x_{m-1}, x_{m-2})$ and the denominator is $p(x_1)p(x_2|x_1) \prod_{m=3}^{n-1} p(x_m|x_{m-1}, x_{m-2})$.

$$p(x_n|x_1, \ldots, x_{n-1}) = \frac{p(x_1)p(x_2|x_1) \left(\prod_{m=3}^{n-1} p(x_m|x_{m-1}, x_{m-2})\right) p(x_n|x_{n-1}, x_{n-2})}{p(x_1)p(x_2|x_1) \prod_{m=3}^{n-1} p(x_m|x_{m-1}, x_{m-2})}$$

Cancelling terms leaves:

$$p(x_n|x_1,\dots,x_{n-1}) = p(x_n|x_{n-1},x_{n-2})$$

This verifies the property (13.123).

Exercise 13.3

Goal: Show that the observed variables $x_1,\dots,x_N$ in the state space model (Figure 13.5) do not satisfy any finite-order Markov property.

We need to show that for any order M, $p(x_n|x_1,\dots,x_{n-1}) \neq p(x_n|x_{n-1},\dots,x_{n-M})$. This means x_n is generally *not* conditionally independent of x_{n-M-1} (and earlier observations) given $x_{n-1},\dots,x_{n-M}$.

Let's test conditional independence between x_n and x_{n-M-1} given the intermediate observations $C = \{x_{n-1},\dots,x_{n-M}\}$.

- Consider the path $x_{n-M-1} \leftarrow z_{n-M-1} \rightarrow z_{n-M} \rightarrow \cdots \rightarrow z_{n-1} \rightarrow z_n \rightarrow x_n$.
- We need to check if this path is blocked by the conditioning set C.
- The intermediate nodes on this path are the latent variables $z_{n-M-1},\dots,z_n$. None of these are in the conditioning set C.
- All nodes on the segment $z_{n-M-1} \rightarrow \cdots \rightarrow z_n$ are head-to-tail or tail-to-tail.
- The nodes z_{n-M-1} and z_n are head-to-tail with respect to the observed variables x_{n-M-1} and x_n.
- Critically, none of the observed nodes $x_{n-1},\dots,x_{n-M}$ in the conditioning set C lie on this path.
- Therefore, the path is **not blocked**.
- Since there exists an unblocked path, x_n and x_{n-M-1} are *not* d-separated by C.
- This means $p(x_n|x_1,\dots,x_{n-1})$ depends on x_{n-M-1} even when conditioned on $x_{n-1},\dots,x_{n-M}$.

This holds for any finite M. Therefore, the observed sequence does not exhibit the Markov property at any finite order.

Exercise 13.4

Goal: Describe how to learn the parameters w of a parametric emission density $p(x|z,w)$ using maximum likelihood.

This scenario fits directly into the Expectation-Maximization (EM) framework for HMMs. The model parameters are $\theta = \{\pi, A, w\}$.

1. E-Step

- Given the current parameter estimates $\theta^{\text{old}} = \{\pi^{\text{old}}, A^{\text{old}}, w^{\text{old}}\}$, compute the posterior probabilities (responsibilities) of the latent states:

- $\gamma(z_n) = p(z_n|X, \theta^{\text{old}})$
- $\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n|X, \theta^{\text{old}})$

- This is done using the standard forward-backward algorithm (Section 13.2.2). The only modification is that the emission probabilities $p(x_n|z_{nk} = 1)$ needed in the recursions (e.g., in Eq. 13.36 and 13.38) are now calculated using the parametric model $p(x_n|z_{nk} = 1, w^{\text{old}})$.

2. M-Step

- Maximize the expected complete-data log likelihood $Q(\theta, \theta^{\text{old}})$ with respect to the parameters $\theta = \{\pi, A, w\}$.
- The function Q is given by (an extension of Eq. 13.17):

$$Q(\theta, \theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j,k} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(x_n|z_{nk} = 1, w)$$

- The maximization with respect to π and A yields the standard M-step equations (13.18) and (13.19).
- To update w, we need to maximize the third term with respect to w:

$$w^{\text{new}} = \arg\max_{w} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(x_n|z_{nk} = 1, w)$$

- This is equivalent to finding the maximum likelihood parameters w for the emission model, where each data point (x_n) is weighted by the posterior probability $\gamma(z_{nk})$ that it was generated by the component (state k) associated with that part of the emission model.
- The exact method for maximizing this term depends on the specific parametric form of $p(x|z, w)$. For example, if it's a linear regression model, it involves weighted least squares; if it's a neural network, it involves training the network with weighted samples.

Exercise 13.5

Goal: Verify the M-step equations (13.18) for π_k and (13.19) for A_{jk} by maximizing $Q(\theta, \theta^{\text{old}})$ from (13.17).

1. Maximization for π_k

- We need to maximize the first term of Q subject to $\sum_k \pi_k = 1$:

$$L(\pi) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k$$

- Introduce a Lagrange multiplier λ:

$$\tilde{L}(\pi, \lambda) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

- Take the derivative with respect to π_k and set to zero:

$$\frac{\partial \tilde{L}}{\partial \pi_k} = \frac{\gamma(z_{1k})}{\pi_k} + \lambda = 0 \implies \gamma(z_{1k}) = -\lambda \pi_k$$

- Sum both sides over k:

$$\sum_{k=1}^{K} \gamma(z_{1k}) = -\lambda \sum_{k=1}^{K} \pi_k = -\lambda(1) = -\lambda$$

- Substitute $\lambda = -\sum_j \gamma(z_{1j})$ back into the equation for $\gamma(z_{1k})$:

$$\gamma(z_{1k}) = -\left(-\sum_{j=1}^{K} \gamma(z_{1j})\right) \pi_k = \pi_k \sum_{j=1}^{K} \gamma(z_{1j})$$

- Solving for π_k:

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}$$

  This matches Equation (13.18).

2. Maximization for A_{jk}

- We need to maximize the second term of Q subject to the constraints $\sum_k A_{jk} = 1$ for each $j \in \{1, \ldots, K\}$:

$$L(A) = \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk}$$

- Introduce Lagrange multipliers λ_j for each constraint:

$$\tilde{L}(A, \lambda) = \sum_{n=2}^{N} \sum_{j,k} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{j=1}^{K} \lambda_j \left(\sum_{k=1}^{K} A_{jk} - 1\right)$$

- Take the derivative with respect to A_{jk} and set to zero:

$$\frac{\partial \tilde{L}}{\partial A_{jk}} = \sum_{n=2}^{N} \frac{\xi(z_{n-1,j}, z_{nk})}{A_{jk}} + \lambda_j = 0$$

$$\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk}) = -\lambda_j A_{jk}$$

- Sum both sides over k:

$$\sum_{k=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk}) = -\lambda_j \sum_{k=1}^{K} A_{jk} = -\lambda_j(1) = -\lambda_j$$

- Substitute $\lambda_j = -\sum_l \sum_{m=2}^{N} \xi(z_{m-1,j}, z_{ml})$ back into the equation:

$$\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk}) = -\left(-\sum_{l} \sum_{m=2}^{N} \xi(z_{m-1,j}, z_{ml})\right) A_{jk}$$

- Solving for A_{jk}:

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_l \sum_{m=2}^{N} \xi(z_{m-1,j}, z_{ml})}$$

  The denominator sum index m can be replaced by n, and the sum index l by k', giving:

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{k'=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk'})}$$

  This matches Equation (13.19).

Exercise 13.6

Goal: Show that if π_k or A_{jk} are initially zero, they remain zero during EM updates.

- **Initial State Probabilities (π_k):** The M-step update is $\pi_k = \gamma(z_{1k}) / \sum_j \gamma(z_{1j})$. $\gamma(z_{1k})$ is the posterior probability $p(z_{1k} = 1|X, \theta^{\text{old}})$. From the forward recursion initialization (13.37), $\alpha(z_{1k}) \propto \pi_k^{\text{old}} p(x_1|\phi_k^{\text{old}})$. The posterior $\gamma(z_{1k})$ is derived from $\alpha(z_{1k})\beta(z_{1k})$. If $\pi_k^{\text{old}} = 0$, then $\alpha(z_{1k}) = 0$, which implies $\gamma(z_{1k}) = 0$. Therefore, the numerator in the M-step equation (13.18) is zero, and π_k^{new} will also be zero.
- **Transition Probabilities (A_{jk}):** The M-step update is $A_{jk} = (\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk}))/(\sum_l \sum_{n=2}^{N} \xi(z_{n-1,j},$ $\xi(z_{n-1,j}, z_{nk})$ is the posterior $p(z_{n-1,j} = 1, z_{nk} = 1|X, \theta^{\text{old}})$. From (13.43), $\xi(z_{n-1,j}, z_{nk}) \propto \alpha(z_{n-1,j}) p(x_n|z_{nk} = 1) p(z_{nk} = 1|z_{n-1,j} = 1) \beta(z_{nk})$. The term $p(z_{nk} = 1|z_{n-1,j} = 1)$ is exactly A_{jk}^{old}. If $A_{jk}^{\text{old}} = 0$, then $\xi(z_{n-1,j}, z_{nk}) = 0$ for all n. Therefore, the numerator in the M-step equation (13.19) is zero, and A_{jk}^{new} will also be zero.

Thus, zero values in π or A persist through EM updates.

Exercise 13.7

Goal: Derive M-step equations (13.20) and (13.21) for Gaussian emission densities.

- We need to maximize the third term in $Q(\theta, \theta^{\text{old}})$ (Eq. 13.17) with respect to the Gaussian parameters $\phi_k = \{\mu_k, \Sigma_k\}$. The relevant term is:

$$L(\phi) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(x_n|\phi_k)$$

  where $p(x_n|\phi_k) = \mathcal{N}(x_n|\mu_k, \Sigma_k)$.

$$L(\phi) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left[-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma_k| - \frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right]$$

- This expression decouples for each k. For a specific k, we maximize:

$$L_k(\mu_k, \Sigma_k) = \sum_{n=1}^{N} \gamma(z_{nk}) \left[-\frac{1}{2} \ln |\Sigma_k| - \frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right] + \text{const}$$

- This is identical to the maximization required for the parameters of a single Gaussian distribution, but where each data point x_n is weighted by the responsibility $\gamma(z_{nk})$.
- Using the standard maximum likelihood results for a weighted Gaussian, we get:

$$\mu_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\Sigma_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{\text{new}})(x_n - \mu_k^{\text{new}})^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

These match Equations (13.20) and (13.21).

Exercise 13.8

Goal: Derive results for discrete multinomial and Bernoulli observations.

1. Multinomial Observations

- Assume the observation x_n is a 1-of-D binary vector. The emission probability for state k is governed by parameters $\mu_{ik} = p(x_{ni} = 1 | z_{nk} = 1)$, where $\sum_i \mu_{ik} = 1$. The conditional distribution is:

$$p(x_n | z_{nk} = 1, \mu) = \prod_{i=1}^{D} \mu_{ik}^{x_{ni}}$$

Combining over states k using the 1-of-K vector z_n:

$$p(x_n | z_n, \mu) = \prod_{k=1}^{K} \left(\prod_{i=1}^{D} \mu_{ik}^{x_{ni}} \right)^{z_{nk}} = \prod_{i=1}^{D} \prod_{k=1}^{K} \mu_{ik}^{x_{ni} z_{nk}}$$

This matches Equation (13.22).

- **M-Step:** We maximize the emission term in Q w.r.t. μ_{ik}, subject to $\sum_i \mu_{ik} = 1$ for each k.

$$L(\mu) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(x_n | z_{nk} = 1, \mu) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \sum_{i=1}^{D} x_{ni} \ln \mu_{ik}$$

For a fixed k, we maximize $\sum_{n=1}^{N} \gamma(z_{nk}) \sum_{i=1}^{D} x_{ni} \ln \mu_{ik}$ subject to $\sum_i \mu_{ik} = 1$. Using a Lagrange multiplier λ_k:

$$\tilde{L}_k = \sum_n \gamma(z_{nk}) \sum_i x_{ni} \ln \mu_{ik} + \lambda_k \left(\sum_i \mu_{ik} - 1 \right)$$

Taking the derivative w.r.t. μ_{ik} and setting to zero:

$$\frac{\partial \tilde{L}_k}{\partial \mu_{ik}} = \sum_n \gamma(z_{nk}) \frac{x_{ni}}{\mu_{ik}} + \lambda_k = 0 \implies \sum_n \gamma(z_{nk}) x_{ni} = -\lambda_k \mu_{ik}$$

Summing over i:

$$\sum_i \sum_n \gamma(z_{nk}) x_{ni} = -\lambda_k \sum_i \mu_{ik} = -\lambda_k$$

Since $\sum_i x_{ni} = 1$, the left side is $\sum_n \gamma(z_{nk})(1) = N_k$. So, $-\lambda_k = N_k$. Substituting back: $\sum_n \gamma(z_{nk}) x_{ni} = N_k \mu_{ik}$.

$$\mu_{ik}^{\text{new}} = \frac{\sum_{n=1}^N \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^N \gamma(z_{nk})}$$

This matches Equation (13.23).

2. Bernoulli Observations

- Assume x_n is a D-dimensional binary vector, where each element x_{ni} is an independent Bernoulli trial given the state $z_{nk} = 1$. Let $\mu_{ik} = p(x_{ni} = 1 | z_{nk} = 1)$.
- **Conditional Distribution:**

$$p(x_n | z_{nk} = 1, \mu) = \prod_{i=1}^D \mu_{ik}^{x_{ni}} (1 - \mu_{ik})^{1 - x_{ni}}$$

  Combining over states k:

$$p(x_n | z_n, \mu) = \prod_{k=1}^K \left(\prod_{i=1}^D \mu_{ik}^{x_{ni}} (1 - \mu_{ik})^{1 - x_{ni}} \right)^{z_{nk}}$$

- **M-Step:** We maximize the emission term in Q:

$$L(\mu) = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \sum_{i=1}^D [x_{ni} \ln \mu_{ik} + (1 - x_{ni}) \ln(1 - \mu_{ik})]$$

  This decouples for each (i, k) pair. Taking the derivative w.r.t. μ_{ik} and setting to zero:

$$\sum_{n=1}^N \gamma(z_{nk}) \left[\frac{x_{ni}}{\mu_{ik}} - \frac{1 - x_{ni}}{1 - \mu_{ik}} \right] = 0$$

  Solving gives:

$$\mu_{ik}^{\text{new}} = \frac{\sum_{n=1}^N \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^N \gamma(z_{nk})}$$

Exercise 13.9

Goal: Use d-separation to verify properties (13.24)-(13.31).

Refer to the HMM graph (Figure 13.5). Let $X = \{x_1, \ldots, x_N\}$, $X_{1:n} = \{x_1, \ldots, x_n\}$, $X_{n+1:N} = \{x_{n+1}, \ldots, x_N\}$.

- **(13.24)** $p(X|z_n) = p(X_{1:n}|z_n)p(X_{n+1:N}|z_n)$**:** Needs $X_{1:n} \perp\!\!\!\perp X_{n+1:N}|z_n$. Path $x_i \leftarrow z_i \rightarrow \cdots \rightarrow z_n \rightarrow \cdots \rightarrow z_j \rightarrow x_j$ $(i \leq n < j)$. z_n is observed, head-to-tail. Path blocked. Holds.
- **(13.25)** $p(X_{1:n-1}|x_n, z_n) = p(X_{1:n-1}|z_n)$**:** Needs $X_{1:n-1} \perp\!\!\!\perp x_n|z_n$. Path $x_i \leftarrow z_i \rightarrow \cdots \rightarrow z_n \rightarrow x_n$ $(i \leq n-1)$. z_n is observed, head-to-tail. Path blocked. Holds.
- **(13.26)** $p(X_{n+1:N}|z_n, z_{n+1}) = p(X_{n+1:N}|z_{n+1})$**:** Needs $X_{n+1:N} \perp\!\!\!\perp z_n|z_{n+1}$. Path $x_j \leftarrow z_j \leftarrow \cdots \leftarrow z_{n+1} \leftarrow z_n$ $(j \geq n+1)$. z_{n+1} is observed, head-to-tail. Path blocked. Holds.
- **(13.27)** $p(X_{n+2:N}|z_{n+1}, x_{n+1}) = p(X_{n+2:N}|z_{n+1})$**:** Needs $X_{n+2:N} \perp\!\!\!\perp x_{n+1}|z_{n+1}$. Path $x_j \leftarrow z_j \leftarrow \cdots \leftarrow z_{n+1} \rightarrow x_{n+1}$ $(j \geq n+2)$. z_{n+1} is observed, tail-to-tail. Path blocked. Holds.
- **(13.28)** $p(X_{1:n-1}|z_{n-1}, z_n) = p(X_{1:n-1}|z_{n-1})$**:** Needs $X_{1:n-1} \perp\!\!\!\perp z_n|z_{n-1}$. Path $x_i \leftarrow z_i \rightarrow \cdots \rightarrow z_{n-1} \rightarrow z_n$ $(i \leq n-1)$. z_{n-1} is observed, head-to-tail. Path blocked. Holds.
- **(13.29)** $p(X|z_{n-1}, z_n) = p(X_{1:n-1}|z_{n-1})p(x_n|z_n)p(X_{n+1:N}|z_n)$**:** Needs $(X_{1:n-1}, x_n) \perp\!\!\!\perp X_{n+1:N}|\{z_{n-1}, z_n\}$ and $X_{1:n-1} \perp\!\!\!\perp x_n|\{z_{n-1}, z_n\}$. Both hold because paths are blocked by observed z_n or z_{n-1}.
- **(13.30)** $p(x_{N+1}|X, z_{N+1}) = p(x_{N+1}|z_{N+1})$**:** Needs $x_{N+1} \perp\!\!\!\perp X|z_{N+1}$. Path $x_i \leftarrow z_i \rightarrow \cdots \rightarrow z_{N+1} \rightarrow x_{N+1}$ $(i \leq N)$. z_{N+1} is observed, tail-to-tail. Path blocked. Holds.
- **(13.31)** $p(z_{N+1}|z_N, X) = p(z_{N+1}|z_N)$**:** Needs $z_{N+1} \perp\!\!\!\perp X|z_N$. Path $x_i \leftarrow z_i \rightarrow \cdots \rightarrow z_N \rightarrow z_{N+1}$ $(i \leq N)$. z_N is observed, head-to-tail. Path blocked. Holds.

Exercise 13.10

Goal: Verify properties (13.24)-(13.31) using sum/product rules.

Let the joint distribution be $p(X, Z) = p(z_1)\left(\prod_{m=2}^{N} p(z_m|z_{m-1})\right)\left(\prod_{m=1}^{N} p(x_m|z_m)\right)$.

(13.24) $p(X|z_n) = p(X_{1:n}|z_n)p(X_{n+1:N}|z_n)$**:**

$$p(X|z_n) = \frac{p(X, z_n)}{p(z_n)} = \frac{\sum_{Z_{\backslash n}} p(X, Z)}{p(z_n)}$$

The joint $p(X, Z)$ factors into terms involving $Z_{1:n}, X_{1:n}$ and terms involving $Z_{n+1:N}, X_{n+1:N}$, connected only by $p(z_{n+1}|z_n)$. Summing over $Z_{1:n-1}$ relates $X_{1:n-1}$ to z_n. Summing over $Z_{n+1:N}$ relates $X_{n+1:N}$ to z_n. Due to the Markov structure, the dependence factorizes.

$$p(X, z_n) = p(X_{1:n}, z_n)p(X_{n+1:N}|z_n)$$

Dividing by $p(z_n)$ gives $p(X|z_n) = p(X_{1:n}|z_n)p(X_{n+1:N}|z_n)$. Holds.

(13.25) $p(X_{1:n-1}|x_n, z_n) = p(X_{1:n-1}|z_n)$**:**

$$p(X_{1:n-1}|x_n, z_n) = \frac{p(X_{1:n-1}, x_n|z_n)}{p(x_n|z_n)} = \frac{p(X_{1:n-1}|z_n)p(x_n|X_{1:n-1}, z_n)}{p(x_n|z_n)}$$

From the graph structure, x_n only depends directly on z_n, so $p(x_n|X_{1:n-1}, z_n) = p(x_n|z_n)$. Substituting yields $p(X_{1:n-1}|x_n, z_n) = p(X_{1:n-1}|z_n)$. Holds.

(13.26) $p(X_{n+1:N}|z_n, z_{n+1}) = p(X_{n+1:N}|z_{n+1})$**:**

$$p(X_{n+1:N}|z_n, z_{n+1}) = \frac{p(X_{n+1:N}, z_{n+1}|z_n)}{p(z_{n+1}|z_n)}$$

Consider the generation process starting from z_n. The state z_{n+1} is generated via $p(z_{n+1}|z_n)$. All subsequent observations $X_{n+1:N}$ and states $Z_{n+1:N}$ depend only on z_{n+1} (and future states/observations), not directly on z_n. Therefore, $p(X_{n+1:N}, z_{n+1}|z_n) = p(z_{n+1}|z_n)p(X_{n+1:N}|z_{n+1})$. Substituting yields $p(X_{n+1:N}|z_n, z_{n+1}) = p(X_{n+1:N}|z_{n+1})$. Holds.

(Derivations for (13.27)-(13.31) follow similar logic, applying the product rule and identifying terms that cancel or factor out based on the conditional independencies encoded in the joint distribution 13.10).

Exercise 13.11

Goal: Derive the pairwise marginal $\xi(z_{n-1}, z_n)$ (13.43) using the sum-product result (8.72).

- The sum-product algorithm calculates the marginal for the variables within a factor $f_s(\mathbf{x}_s)$ using $p(\mathbf{x}_s|X) \propto f_s(\mathbf{x}_s) \prod_{i \in \text{ne}(f_s)} \mu_{x_i \to f_s}(x_i)$.
- In our simplified HMM factor graph (Figure 13.15), consider the factor $f_n(z_{n-1}, z_n)$. The variables in this factor are $\mathbf{x}_s = \{z_{n-1}, z_n\}$.
- The factor is $f_n(z_{n-1}, z_n) = p(z_n|z_{n-1})p(x_n|z_n)$.
- The neighbours of f_n are z_{n-1} and z_n.
- We need the messages $\mu_{z_{n-1} \to f_n}(z_{n-1})$ and $\mu_{z_n \to f_n}(z_n)$.
- $\mu_{z_{n-1} \to f_n}(z_{n-1})$ is the message coming *into* f_n from z_{n-1}. This is the message coming *out* of factor f_{n-1} towards z_{n-1}, which is $\mu_{f_{n-1} \to z_{n-1}}(z_{n-1}) = \alpha(z_{n-1})$.
- $\mu_{z_n \to f_n}(z_n)$ is the message coming *into* f_n from z_n. This is the message coming *out* of factor f_{n+1} towards z_n, which is $\mu_{f_{n+1} \to z_n}(z_n) = \beta(z_n)$.
- Applying (8.72), the joint marginal $p(z_{n-1}, z_n|X)$ is proportional to the product of the factor and incoming messages:

$$p(z_{n-1}, z_n|X) \propto f_n(z_{n-1}, z_n)\mu_{z_{n-1} \to f_n}(z_{n-1})\mu_{z_n \to f_n}(z_n)$$

$$\xi(z_{n-1}, z_n) \propto [p(z_n|z_{n-1})p(x_n|z_n)]\, \alpha(z_{n-1})\beta(z_n)$$

- The normalization constant is $p(X)$. Therefore:

$$\xi(z_{n-1}, z_n) = \frac{\alpha(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\beta(z_n)}{p(X)}$$

  This matches Equation (13.43).

Exercise 13.12

Goal: Adapt EM for multiple independent sequences $X^{(r)}$, $r = 1, \ldots, R$.

Let $X = \{X^{(1)}, \ldots, X^{(R)}\}$ and $Z = \{Z^{(1)}, \ldots, Z^{(R)}\}$. The total log likelihood is $\ln p(X|\theta) = \sum_r \ln p(X^{(r)}|\theta)$. The complete-data log likelihood is $\ln p(X, Z|\theta) = \sum_r \ln p(X^{(r)}, Z^{(r)}|\theta)$.

1. E-Step

- The expected complete-data log likelihood is:

$$Q(\theta, \theta^{\text{old}}) = \mathbb{E}_{Z|X,\theta^{\text{old}}}[\ln p(X, Z|\theta)] = \mathbb{E}_{Z|X,\theta^{\text{old}}}\left[\sum_r \ln p(X^{(r)}, Z^{(r)}|\theta)\right]$$

- Since the sequences are independent, the posterior factorizes: $p(Z|X, \theta^{\text{old}}) = \prod_r p(Z^{(r)}|X^{(r)}, \theta^{\text{old}})$.
- The expectation becomes:

$$Q(\theta, \theta^{\text{old}}) = \sum_r \mathbb{E}_{Z^{(r)}|X^{(r)},\theta^{\text{old}}}[\ln p(X^{(r)}, Z^{(r)}|\theta)] = \sum_r Q_r(\theta, \theta^{\text{old}})$$

  where Q_r is the Q function for sequence r.
- To evaluate Q, we need the posteriors $\gamma(z_n^{(r)}) = p(z_n^{(r)}|X^{(r)}, \theta^{\text{old}})$ and $\xi(z_{n-1}^{(r)}, z_n^{(r)}) = p(z_{n-1}^{(r)}, z_n^{(r)}|X^{(r)}, \theta^{\text{old}})$
- These are calculated by running the forward-backward algorithm independently for each sequence $X^{(r)}$ using θ^{old}.

2. M-Step

- We maximize $Q = \sum_r Q_r$. Using the form (13.17) for each Q_r:

$$Q = \sum_r \left(\sum_k \gamma(z_{1k}^{(r)}) \ln \pi_k + \sum_{n=2}^{N_r} \sum_{j,k} \xi(z_{n-1,j}^{(r)}, z_{nk}^{(r)}) \ln A_{jk} + \sum_{n=1}^{N_r} \sum_k \gamma(z_{nk}^{(r)}) \ln p(x_n^{(r)}|\phi_k) \right)$$

  (Assuming sequences can have different lengths N_r).
- **For π_k:** Maximize $\sum_r \sum_k \gamma(z_{1k}^{(r)}) \ln \pi_k$ subject to $\sum_k \pi_k = 1$. Using a Lagrange multiplier:

$$\pi_k^{\text{new}} = \frac{\sum_r \gamma(z_{1k}^{(r)})}{\sum_j \sum_r \gamma(z_{1j}^{(r)})}$$

  This matches (13.124).
- **For A_{jk}:** Maximize $\sum_r \sum_{n=2}^{N_r} \sum_{j,k} \xi(z_{n-1,j}^{(r)}, z_{nk}^{(r)}) \ln A_{jk}$ subject to $\sum_k A_{jk} = 1$ for each j. Using Lagrange multipliers:

$$A_{jk}^{\text{new}} = \frac{\sum_r \sum_{n=2}^{N_r} \xi(z_{n-1,j}^{(r)}, z_{nk}^{(r)})}{\sum_l \sum_r \sum_{n=2}^{N_r} \xi(z_{n-1,j}^{(r)}, z_{nl}^{(r)})}$$

  This matches (13.125).

- **For ϕ_k (e.g., Gaussian mean μ_k):** Maximize $\sum_r \sum_{n=1}^{N_r} \sum_k \gamma(z_{nk}^{(r)}) \ln p(x_n^{(r)}|\phi_k)$. For Gaussian emissions:
$$\mu_k^{\text{new}} = \frac{\sum_r \sum_{n=1}^{N_r} \gamma(z_{nk}^{(r)}) x_n^{(r)}}{\sum_r \sum_{n=1}^{N_r} \gamma(z_{nk}^{(r)})}$$
This matches (13.126). Updates for other parameters follow similarly.

Exercise 13.13

Goal: Show sum-product $\alpha(z_n)$ (13.50) matches HMM $\alpha(z_n)$ (13.34).

- The sum-product message from factor f_n to variable z_n is $\mu_{f_n \to z_n}(z_n)$.
- From the general message passing rules (8.64) and (8.69):
$$\mu_{f_n \to z_n}(z_n) = \sum_{z_{n-1}} f_n(z_{n-1}, z_n) \mu_{z_{n-1} \to f_n}(z_{n-1})$$
$$\mu_{z_{n-1} \to f_n}(z_{n-1}) = \mu_{f_{n-1} \to z_{n-1}}(z_{n-1})$$
- Substituting gives the recursion (13.49):
$$\mu_{f_n \to z_n}(z_n) = \sum_{z_{n-1}} f_n(z_{n-1}, z_n) \mu_{f_{n-1} \to z_{n-1}}(z_{n-1})$$
- Using $f_n(z_{n-1}, z_n) = p(z_n|z_{n-1})p(x_n|z_n)$ and defining $\alpha(z_n) = \mu_{f_n \to z_n}(z_n)$, this becomes:
$$\alpha(z_n) = p(x_n|z_n) \sum_{z_{n-1}} \alpha(z_{n-1}) p(z_n|z_{n-1})$$
This is identical to the HMM alpha recursion (13.36).
- For the initial condition, $\alpha(z_1) = \mu_{h \to z_1}(z_1) = h(z_1) = p(z_1)p(x_1|z_1)$. This matches (13.37).
- Since the initial conditions and recursion relations are identical, the sum-product $\alpha(z_n)$ is the same quantity as the HMM $\alpha(z_n) = p(x_1, \ldots, x_n, z_n)$.

Exercise 13.14

Goal: Show sum-product $\beta(z_n)$ (13.52) matches HMM $\beta(z_n)$ (13.35).

- The sum-product message from factor f_{n+1} to variable z_n is $\mu_{f_{n+1} \to z_n}(z_n)$.
- From the general message passing rule (8.67):
$$\mu_{f_{n+1} \to z_n}(z_n) = \sum_{z_{n+1}} f_{n+1}(z_n, z_{n+1}) \mu_{z_{n+1} \to f_{n+1}}(z_{n+1})$$
- The message $\mu_{z_{n+1} \to f_{n+1}}(z_{n+1})$ comes from z_{n+1} into f_{n+1}. Since z_{n+1} has neighbours f_{n+1} and f_{n+2}, this message must be the one coming from f_{n+2}:
$$\mu_{z_{n+1} \to f_{n+1}}(z_{n+1}) = \mu_{f_{n+2} \to z_{n+1}}(z_{n+1})$$

- Substituting this back:

$$\mu_{f_{n+1}\to z_n}(z_n) = \sum_{z_{n+1}} f_{n+1}(z_n, z_{n+1})\mu_{f_{n+2}\to z_{n+1}}(z_{n+1})$$

  This matches the recursion (13.51).

- Using $f_{n+1}(z_n, z_{n+1}) = p(z_{n+1}|z_n)p(x_{n+1}|z_{n+1})$ and defining $\beta(z_n) = \mu_{f_{n+1}\to z_n}(z_n)$, this becomes:

$$\beta(z_n) = \sum_{z_{n+1}} p(z_{n+1}|z_n)p(x_{n+1}|z_{n+1})\beta(z_{n+1})$$

  This is identical to the HMM beta recursion (13.38).

- For the initial condition, the message leaving the root z_N is initialized to 1 (Eq. 8.70). This implies $\beta(z_N) = 1$, matching the HMM initialization.
- Since the initial conditions and recursion relations are identical, the sum-product $\beta(z_n)$ is the same quantity as the HMM $\beta(z_n) = p(x_{n+1}, \ldots, x_N|z_n)$.

Exercise 13.15

Goal: Derive scaled marginals (13.64) and (13.65) using scaled variables $\hat{\alpha}, \hat{\beta}$.

Definitions:

- $\alpha(z_n) = (\prod_{m=1}^{n} c_m)\hat{\alpha}(z_n)$
- $\beta(z_n) = (\prod_{m=n+1}^{N} c_m)\hat{\beta}(z_n)$
- $p(X) = \prod_{m=1}^{N} c_m$

1. Derive $\gamma(z_n)$ (13.64)

- Start with the definition (13.33): $\gamma(z_n) = \frac{\alpha(z_n)\beta(z_n)}{p(X)}$.
- Substitute the scaled versions:

$$\gamma(z_n) = \frac{[(\prod_{m=1}^{n} c_m)\hat{\alpha}(z_n)][(\prod_{m=n+1}^{N} c_m)\hat{\beta}(z_n)]}{\prod_{m=1}^{N} c_m}$$

- The product of scaling factors in the numerator is $(\prod_{m=1}^{n} c_m)(\prod_{m=n+1}^{N} c_m) = \prod_{m=1}^{N} c_m = p(X)$.
- Therefore:

$$\gamma(z_n) = \frac{p(X)\hat{\alpha}(z_n)\hat{\beta}(z_n)}{p(X)} = \hat{\alpha}(z_n)\hat{\beta}(z_n)$$

  This matches (13.64).

2. Derive $\xi(z_{n-1}, z_n)$ (13.65)

- Start with the definition (13.43): $\xi(z_{n-1}, z_n) = \frac{\alpha(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\beta(z_n)}{p(X)}$.

- Substitute the scaled versions:

$$\xi(z_{n-1}, z_n) = \frac{[(\prod_{m=1}^{n-1} c_m)\hat{\alpha}(z_{n-1})]p(x_n|z_n)p(z_n|z_{n-1})[(\prod_{m=n+1}^{N} c_m)\hat{\beta}(z_n)]}{\prod_{m=1}^{N} c_m}$$

- Rearrange the scaling factors:

$$\xi(z_{n-1}, z_n) = \frac{(\prod_{m=1}^{n-1} c_m)(\prod_{m=n+1}^{N} c_m)}{(\prod_{m=1}^{N} c_m)}\hat{\alpha}(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\hat{\beta}(z_n)$$

- The fraction of scaling factors simplifies:

$$\frac{\prod_{m\neq n} c_m}{\prod_{m=1}^{N} c_m} = \frac{1}{c_n}$$

- Therefore:

$$\xi(z_{n-1}, z_n) = \frac{1}{c_n}\hat{\alpha}(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\hat{\beta}(z_n)$$

  Rearranging gives the form in (13.65):

$$\xi(z_{n-1}, z_n) = c_n^{-1}\hat{\alpha}(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\hat{\beta}(z_n)$$

Exercise 13.16

Goal: Derive the Viterbi forward recursion (13.68) directly from the joint distribution (13.6) by taking the log and exchanging max/sum. Verify the initial condition (13.69).

1. **Joint Log Probability:** Take the logarithm of the joint distribution (13.10):

$$\ln p(X, Z|\theta) = \ln p(z_1|\pi) + \sum_{n=2}^{N} \ln p(z_n|z_{n-1}, A) + \sum_{n=1}^{N} \ln p(x_n|z_n, \phi)$$

2. **Most Probable Sequence:** We want to find $\max_Z \ln p(X, Z|\theta)$.

$$\max_{z_1,\ldots,z_N}\left[\ln p(z_1) + \sum_{n=2}^{N} \ln p(z_n|z_{n-1}) + \sum_{n=1}^{N} \ln p(x_n|z_n)\right]$$

   (Omitting parameter dependence for clarity). We can rearrange the terms involving x_n:

$$= \max_{z_1,\ldots,z_N}\left[\ln p(z_1) + \ln p(x_1|z_1) + \sum_{n=2}^{N}(\ln p(z_n|z_{n-1}) + \ln p(x_n|z_n))\right]$$

3. **Dynamic Programming:** Use dynamic programming by pushing the maximizations inwards. Define $\omega(z_n)$ as the maximum log probability of the sequence up to step n, ending in state z_n:

$$\omega(z_n) = \max_{z_1,\ldots,z_{n-1}} \ln p(x_1, \ldots, x_n, z_1, \ldots, z_n)$$

   We can write a recursion for $\omega(z_n)$. Consider the terms up to step n:

$$\ln p(x_1, \ldots, x_n, z_1, \ldots, z_n) = \ln p(x_1, \ldots, x_{n-1}, z_1, \ldots, z_{n-1}) + \ln p(z_n|z_{n-1}) + \ln p(x_n|z_n)$$

Maximizing over $z_1, \ldots, z_{n-1}$:

$$\omega(z_n) = \max_{z_{n-1}} \left[\max_{z_1,\ldots,z_{n-2}} \ln p(\ldots, z_{n-1}) + \ln p(z_n|z_{n-1}) + \ln p(x_n|z_n) \right]$$

$$\omega(z_n) = \max_{z_{n-1}} [\omega(z_{n-1}) + \ln p(z_n|z_{n-1}) + \ln p(x_n|z_n)]$$

$$\omega(z_n) = \ln p(x_n|z_n) + \max_{z_{n-1}} [\ln p(z_n|z_{n-1}) + \omega(z_{n-1})]$$

This matches the form of the max-sum recursion (13.68). The definition here matches (13.70).

4. **Initial Condition:** For $n = 1$, there are no previous states to maximize over.

$$\omega(z_1) = \ln p(x_1, z_1) = \ln p(z_1) + \ln p(x_1|z_1)$$

This matches the initialization (13.69).

Exercise 13.17

Goal: Represent the input-output HMM (Figure 13.18) as the factor graph in Figure 13.15 and define factors.

- The standard HMM factor graph (Figure 13.15) has factors $h(z_1)$ and $f_n(z_{n-1}, z_n)$ representing initial and transition/emission steps.
- In the input-output HMM (Figure 13.18), the transitions $p(z_n|z_{n-1})$ and emissions $p(x_n|z_n)$ now also depend on the input u_n. Let's assume the dependencies are $p(z_n|z_{n-1}, u_n)$ and $p(x_n|z_n, u_n)$.
- The structure of the latent chain is the same, so the factor graph topology remains that of Figure 13.15.
- We just need to redefine the factors to include the conditioning on $U = \{u_1, \ldots, u_N\}$.
- **Initial factor $h(z_1)$:**
$$h(z_1) = p(z_1|\pi)p(x_1|z_1, u_1, \phi)$$
- **General factor $f_n(z_{n-1}, z_n)$ for $n \geq 2$:**
$$f_n(z_{n-1}, z_n) = p(z_n|z_{n-1}, u_n, A)p(x_n|z_n, u_n, \phi)$$

Exercise 13.18

Goal: Derive forward-backward recursions for the input-output HMM.

Using the factor graph representation from Exercise 13.17 and the general sum-product message passing equations.

1. **Forward Recursion (α messages)**
   - Define $\alpha(z_n) = p(x_1, \ldots, x_n, z_n|U, \theta)$.
   - The recursion (analogous to 13.36) uses the factor f_n:

     $$\alpha(z_n) = \sum_{z_{n-1}} f_n(z_{n-1}, z_n)\alpha(z_{n-1})$$

     Substituting f_n from Exercise 13.17:

     $$\alpha(z_n) = \left[\sum_{z_{n-1}} p(z_n|z_{n-1}, u_n, A)\alpha(z_{n-1})\right] p(x_n|z_n, u_n, \phi)$$

   - **Initial Condition:** (Analogous to 13.37) using $h(z_1)$:

     $$\alpha(z_1) = h(z_1) = p(z_1|\pi)p(x_1|z_1, u_1, \phi)$$

     In state-component notation: $\alpha(z_{1k}) = \pi_k p(x_1|z_{1k} = 1, u_1, \phi_k)$.

2. **Backward Recursion (β messages)**
   - Define $\beta(z_n) = p(x_{n+1}, \ldots, x_N|z_n, U, \theta)$.
   - The recursion (analogous to 13.38) uses the factor f_{n+1}:

     $$\beta(z_n) = \sum_{z_{n+1}} f_{n+1}(z_n, z_{n+1})\beta(z_{n+1})$$

     Substituting f_{n+1}:

     $$\beta(z_n) = \sum_{z_{n+1}} p(z_{n+1}|z_n, u_{n+1}, A)p(x_{n+1}|z_{n+1}, u_{n+1}, \phi)\beta(z_{n+1})$$

   - **Initial Condition:** (Analogous to HMM)

     $$\beta(z_N) = 1$$

Exercise 13.19

Goal: Show that for LDS, the sequence of individually most probable states is the same as the globally most probable sequence.

- The Linear Dynamical System (LDS) is a linear-Gaussian model. The joint distribution $p(X, Z)$ is a multivariate Gaussian.
- Conditioning on the observed data X, the posterior distribution $p(Z|X)$ is also a multivariate Gaussian.
- For any multivariate Gaussian distribution, the mode coincides with the mean.

- The globally most probable sequence of latent states, $Z^{\text{MAP}} = \arg\max_Z p(Z|X)$, is the mean of the posterior distribution $\mathbb{E}[Z|X]$.
- The sequence formed by finding the individually most probable state at each time step, $z_n^* = \arg\max_{z_n} p(z_n|X)$, is found by taking the mode of each marginal posterior $p(z_n|X)$.
- Since $p(z_n|X)$ is also Gaussian (marginal of $p(Z|X)$), its mode is equal to its mean $\mathbb{E}[z_n|X]$.
- The mean of the joint posterior $\mathbb{E}[Z|X]$ is composed of the means of the marginal posteriors $\{\mathbb{E}[z_n|X]\}_{n=1}^N$.
- Thus, $Z^{\text{MAP}} = \mathbb{E}[Z|X] = \{\mathbb{E}[z_n|X]\}_{n=1}^N = \{z_n^*\}_{n=1}^N$.
- The globally most probable sequence is identical to the sequence of individually most probable states.

Exercise 13.20

Goal: Prove (13.87) using (2.115).

- Result (2.115) gives the marginal $p(y)$ from $p(y|x) = \mathcal{N}(y|Ax+b, L^{-1})$ and $p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})$. The result is $p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$.
- We want to evaluate $p(z_n) = \int p(z_n|z_{n-1})p(z_{n-1})dz_{n-1}$.
- Here, $p(z_n|z_{n-1}) = \mathcal{N}(z_n|Az_{n-1}, \Gamma)$. Map: $y = z_n$, $x = z_{n-1}$, matrix A, offset $b = 0$, covariance $L^{-1} = \Gamma$.
- And $p(z_{n-1}) = \mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1})$. Map: $\mu = \mu_{n-1}$, covariance $\Lambda^{-1} = V_{n-1}$.
- Applying (2.115), the resulting marginal $p(z_n)$ is Gaussian with:
  - Mean $= A\mu_{n-1} + 0 = A\mu_{n-1}$.
  - Covariance $= L^{-1} + A\Lambda^{-1}A^T = \Gamma + AV_{n-1}A^T$.
- Therefore, $p(z_n) = \mathcal{N}(z_n|A\mu_{n-1}, \Gamma + AV_{n-1}A^T)$.
- Letting $P_{n-1} = AV_{n-1}A^T + \Gamma$ (matching definition 13.88), we get:

  $$\int \mathcal{N}(z_n|Az_{n-1}, \Gamma)\mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1})dz_{n-1} = \mathcal{N}(z_n|A\mu_{n-1}, P_{n-1})$$

  This proves (13.87).

Exercise 13.21

Goal: Derive Kalman filter update equations (13.89)-(13.92) using Gaussian identities.

We start from (13.86) and the result (13.87):

$$c_n\mathcal{N}(z_n|\mu_n, V_n) = \mathcal{N}(x_n|Cz_n, \Sigma)\mathcal{N}(z_n|A\mu_{n-1}, P_{n-1})$$

where $P_{n-1} = AV_{n-1}A^T + \Gamma$. We want the posterior $p(z_n|X_{1:n}) \propto p(x_n|z_n)p(z_n|X_{1:n-1})$.

Consider the joint Gaussian over (z_n, x_n) conditioned on $X_{1:n-1}$.

- $z_n \sim \mathcal{N}(A\mu_{n-1}, P_{n-1})$
- $x_n|z_n \sim \mathcal{N}(Cz_n, \Sigma)$

The joint distribution has:

- $\mathbb{E}[z_n] = A\mu_{n-1}$
- $\mathbb{E}[x_n] = C\mathbb{E}[z_n] = CA\mu_{n-1}$
- $\text{cov}[z_n] = P_{n-1}$
- $\text{cov}[x_n] = C\text{cov}[z_n]C^T + \Sigma = CP_{n-1}C^T + \Sigma$
- $\text{cov}[z_n, x_n] = \text{cov}[z_n]C^T = P_{n-1}C^T$

Using the standard formula for conditional Gaussian distributions $p(z_n|x_n)$:

- Mean $\mu_{z_n|x_n} = \mathbb{E}[z_n] + \text{cov}[z_n, x_n](\text{cov}[x_n])^{-1}(x_n - \mathbb{E}[x_n])$
- Covariance $\Sigma_{z_n|x_n} = \text{cov}[z_n] - \text{cov}[z_n, x_n](\text{cov}[x_n])^{-1}\text{cov}[x_n, z_n]$

Substituting:

- $\mu_n = A\mu_{n-1} + P_{n-1}C^T(CP_{n-1}C^T + \Sigma)^{-1}(x_n - CA\mu_{n-1})$
- $V_n = P_{n-1} - P_{n-1}C^T(CP_{n-1}C^T + \Sigma)^{-1}CP_{n-1}$

Defining the Kalman Gain $K_n = P_{n-1}C^T(CP_{n-1}C^T + \Sigma)^{-1}$ (matches 13.92):

- $\mu_n = A\mu_{n-1} + K_n(x_n - CA\mu_{n-1})$ (Matches 13.89)
- $V_n = P_{n-1} - K_nCP_{n-1} = (I - K_nC)P_{n-1}$ (Matches 13.90).

The normalization constant $c_n = p(x_n|X_{1:n-1})$ is the marginal Gaussian:

$$c_n = \mathcal{N}(x_n|CA\mu_{n-1}, CP_{n-1}C^T + \Sigma)$$

(Matches 13.91).

Exercise 13.22

Goal: Derive c_1 (13.96) using (13.93), (13.76), (13.77), and (2.115).

- Equation (13.93) is $c_1\hat{\alpha}(z_1) = p(z_1)p(x_1|z_1)$.
- Integrating both sides over z_1:

$$c_1 \int \hat{\alpha}(z_1)dz_1 = \int p(z_1)p(x_1|z_1)dz_1$$

- $\hat{\alpha}(z_1) = p(z_1|x_1)$ is normalized, so $\int \hat{\alpha}(z_1)dz_1 = 1$.
- The right hand side is $\int p(x_1, z_1)dz_1 = p(x_1)$.
- Therefore, $c_1 = p(x_1)$.
- We evaluate $p(x_1) = \int p(x_1|z_1)p(z_1)dz_1$.
- Given $p(z_1) = \mathcal{N}(z_1|\mu_0, V_0)$ and $p(x_1|z_1) = \mathcal{N}(x_1|Cz_1, \Sigma)$.

- Use (2.115): $y = x_1$, $x = z_1$, $A = C$, $b = 0$, $L^{-1} = \Sigma$, $\mu = \mu_0$, $\Lambda^{-1} = V_0$.
- Applying (2.115), the marginal $p(x_1)$ is Gaussian with:
  - Mean $= A\mu + b = C\mu_0$.
  - Covariance $= L^{-1} + A\Lambda^{-1}A^T = \Sigma + CV_0C^T$.
- Therefore, $c_1 = p(x_1) = \mathcal{N}(x_1|C\mu_0, CV_0C^T + \Sigma)$.
- This matches (13.96).

Exercise 13.23

Goal: Derive initial Kalman filter state μ_1, V_1 and gain K_1 (13.94, 13.95, 13.97) using (13.93) and Gaussian identities.

- From (13.93), $c_1\hat{\alpha}(z_1) = p(z_1)p(x_1|z_1)$.
- $\hat{\alpha}(z_1) = p(z_1|x_1)$. We want the parameters μ_1, V_1 of $\mathcal{N}(z_1|\mu_1, V_1)$.
- We use the results for conditioning a joint Gaussian $p(x_1, z_1)$, which was derived in Ex 13.22.
- The joint distribution has $\mathbb{E}[z_1] = \mu_0$, $\mathbb{E}[x_1] = C\mu_0$, $\mathrm{cov}[z_1] = V_0$, $\mathrm{cov}[x_1] = CV_0C^T + \Sigma$, $\mathrm{cov}[z_1, x_1] = V_0C^T$.
- Using the conditional Gaussian formulas:
  - Mean $\mu_1 = \mathbb{E}[z_1] + \mathrm{cov}[z_1, x_1](\mathrm{cov}[x_1])^{-1}(x_1 - \mathbb{E}[x_1])$
  - Covariance $V_1 = \mathrm{cov}[z_1] - \mathrm{cov}[z_1, x_1](\mathrm{cov}[x_1])^{-1}\mathrm{cov}[x_1, z_1]$
- Substituting:
$$\mu_1 = \mu_0 + V_0C^T(CV_0C^T + \Sigma)^{-1}(x_1 - C\mu_0)$$
$$V_1 = V_0 - V_0C^T(CV_0C^T + \Sigma)^{-1}CV_0$$
- Define $K_1 = V_0C^T(CV_0C^T + \Sigma)^{-1}$. This matches (13.97).
- Then $\mu_1 = \mu_0 + K_1(x_1 - C\mu_0)$. This matches (13.94).
- And $V_1 = V_0 - K_1CV_0 = (I - K_1C)V_0$. This matches (13.95).

Exercise 13.24

Goal: Show how to handle constant offsets in LDS means by augmenting the state vector.

- Consider the extended model:
$$p(z_n|z_{n-1}) = \mathcal{N}(z_n|Az_{n-1} + a, \Gamma)$$
$$p(x_n|z_n) = \mathcal{N}(x_n|Cz_n + c, \Sigma)$$
- Define an augmented state vector $\tilde{z}_n = \begin{pmatrix} z_n \\ 1 \end{pmatrix}$.

- Define augmented matrices:

$$\tilde{A} = \begin{pmatrix} A & a \\ 0 & 1 \end{pmatrix}, \quad \tilde{C} = \begin{pmatrix} C & c \end{pmatrix}, \quad \tilde{\Gamma} = \begin{pmatrix} \Gamma & 0 \\ 0 & 0 \end{pmatrix}$$

- Check the augmented model equations:
  - $\tilde{A}\tilde{z}_{n-1} = \begin{pmatrix} A & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_{n-1} \\ 1 \end{pmatrix} = \begin{pmatrix} Az_{n-1} + a \\ 1 \end{pmatrix}$. The transition $p(\tilde{z}_n|\tilde{z}_{n-1}) = \mathcal{N}(\tilde{z}_n|\tilde{A}\tilde{z}_{n-1}, \tilde{\Gamma})$ works.
  - $\tilde{C}\tilde{z}_n = \begin{pmatrix} C & c \end{pmatrix} \begin{pmatrix} z_n \\ 1 \end{pmatrix} = Cz_n + c$. The emission $p(x_n|\tilde{z}_n) = \mathcal{N}(x_n|\tilde{C}\tilde{z}_n, \Sigma)$ works.
- Augment initial state: $\tilde{\mu}_0 = (\mu_0^T, 1)^T$, $\tilde{V}_0 = \text{blockdiag}(V_0, 0)$.
- The model with offsets is transformed into the standard LDS form using augmented variables and matrices.

Exercise 13.25

Goal: Show Kalman filter reduces to sequential Bayesian updates for i.i.d. Gaussian data.

- Model i.i.d. $x_1, \ldots, x_N \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ.
- Map to LDS: Let $z_n = \mu$. Constant state $\implies A = 1, \Gamma = 0$. Observations x_n measure $z_n \implies C = 1, \Sigma = \sigma^2$. Initial state $p(z_1) = \mathcal{N}(\mu|\mu_0, V_0)$. Let $\hat{\alpha}(z_n) = p(z_n|x_1, \ldots, x_n) = \mathcal{N}(z_n|\mu_n, V_n)$.
- Apply Kalman filter equations:
  - $P_{n-1} = 1 \cdot V_{n-1} \cdot 1 + 0 = V_{n-1}$.
  - $K_n = V_{n-1} \cdot 1 \cdot (1 \cdot V_{n-1} \cdot 1 + \sigma^2)^{-1} = \frac{V_{n-1}}{V_{n-1}+\sigma^2}$.
  - $\mu_n = \mu_{n-1} + K_n(x_n - \mu_{n-1})$.
  - $V_n = (1 - K_n \cdot 1)P_{n-1} = (1 - \frac{V_{n-1}}{V_{n-1}+\sigma^2})V_{n-1} = \frac{\sigma^2 V_{n-1}}{V_{n-1}+\sigma^2}$.
- Compare with sequential Bayesian updates (Eqs. 2.141, 2.142):
  - Variance/Precision: $V_n^{-1} = \frac{V_{n-1}+\sigma^2}{\sigma^2 V_{n-1}} = \frac{1}{\sigma^2} + \frac{1}{V_{n-1}}$. This matches the precision update rule $1/\sigma_n^2 = 1/\sigma_{n-1}^2 + 1/\sigma^2$.
  - Mean: $\mu_n = \mu_{n-1}(1 - K_n) + K_n x_n = \mu_{n-1}\frac{\sigma^2}{V_{n-1}+\sigma^2} + x_n\frac{V_{n-1}}{V_{n-1}+\sigma^2}$. This can be shown to be equivalent to the sequential mean update $\mu_n = V_n(V_{n-1}^{-1}\mu_{n-1} + (\sigma^2)^{-1}x_n)$.

  The Kalman filter equations reduce to the standard sequential Bayesian updates.

Exercise 13.26

Goal: Show LDS equations reduce to probabilistic PCA posterior when $A = 0, \Gamma = I, \Sigma = \sigma^2 I$.

- PPCA posterior $p(z|x)$ is $\mathcal{N}(z|M^{-1}W^T x/\sigma^2, M^{-1})$ where $M = (I + W^T W/\sigma^2)$ (assuming $\mu = 0$).
- Map to LDS (n=1): $A = 0, \Gamma = I, \Sigma = \sigma^2 I$, $C = W$. Initial state $\mu_0 = 0, V_0 = I$.
- We need $p(z_1|x_1) = \hat{\alpha}(z_1) = \mathcal{N}(z_1|\mu_1, V_1)$. Use initialization equations (13.94), (13.95), (13.97):
  - $K_1 = V_0 C^T (C V_0 C^T + \Sigma)^{-1} = I W^T (W I W^T + \sigma^2 I)^{-1} = W^T (W W^T + \sigma^2 I)^{-1}$.
  - $V_1 = (I - K_1 C) V_0 = (I - W^T (W W^T + \sigma^2 I)^{-1} W) I$.
  - $\mu_1 = \mu_0 + K_1 (x_1 - C\mu_0) = 0 + W^T (W W^T + \sigma^2 I)^{-1} x_1$.
- Check equivalence using matrix inversion identity (C.7): $(P^{-1} + B^T R^{-1} B)^{-1} = P - PB^T (BPB^T + R)^{-1} BP$.
  - Let $P = I, B = W, R = \sigma^2 I$. $(I + W^T (\sigma^2 I)^{-1} W)^{-1} = I - I W^T (W I W^T + \sigma^2 I)^{-1} W I = I - W^T (W W^T + \sigma^2 I)^{-1} W$. So $V_1 = (I + W^T W/\sigma^2)^{-1} = M^{-1}$. Covariances match.
  - Use identity $(I + P^T R^{-1} P)^{-1} P^T R^{-1} = P^T (R + P P^T)^{-1}$. Let $P = W, R = \sigma^2 I$. $(I + W^T W/\sigma^2)^{-1} W^T/\sigma^2 = W^T (W W^T + \sigma^2 I)^{-1}$. The PPCA mean is $M^{-1} W^T x_1/\sigma^2 = (I + W^T W/\sigma^2)^{-1} W^T x_1/\sigma^2$. The Kalman mean is $\mu_1 = W^T (W W^T + \sigma^2 I)^{-1} x_1$. The identity shows the means match.

  The Kalman filter initialization recovers the PPCA posterior.

Exercise 13.27

Goal: Show that if $\Sigma \to 0$, the posterior for z_n has mean $C^{-1} x_n$ and zero variance.

- Consider the Kalman update equations (13.89), (13.90), (13.92). Let $\Sigma \to 0$.
- Kalman Gain: $K_n = P_{n-1} C^T (C P_{n-1} C^T + \Sigma)^{-1}$. Assume C is invertible. As $\Sigma \to 0$: $K_n \approx P_{n-1} C^T (C P_{n-1} C^T)^{-1} = P_{n-1} C^T (C^T)^{-1} P_{n-1}^{-1} C^{-1} = C^{-1}$.
- Covariance Update: $V_n = (I - K_n C) P_{n-1}$. Substituting $K_n \approx C^{-1}$: $V_n \approx (I - C^{-1} C) P_{n-1} = (I - I) P_{n-1} = 0$.
- Mean Update: $\mu_n = A\mu_{n-1} + K_n (x_n - CA\mu_{n-1})$. Substituting $K_n \approx C^{-1}$: $\mu_n \approx A\mu_{n-1} + C^{-1}(x_n - CA\mu_{n-1}) = C^{-1} x_n$.
- The posterior $p(z_n|X_{1:n})$ approaches $\mathcal{N}(z_n|C^{-1} x_n, 0)$.

Exercise 13.28

Goal: Show that if $A = I, \Gamma = 0$ (constant state), the posterior mean μ_n is the average of $x_1, \ldots, x_n$. Assume $V_0 \to \infty$.

- Use induction. Let $p(z_n|x_1, \ldots, x_n) = \mathcal{N}(z_n|\mu_n, V_n)$. Assume $C = I, \Sigma = \sigma^2 I$.
- Base Case $n = 1$: As in Ex 13.25, with $V_0 \to \infty$, we get $\mu_1 = x_1$ and $V_1 = \sigma^2 I$. μ_1 is the average of x_1.

- Inductive Step: Assume $\mu_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i$ and $V_{n-1}^{-1} = V_0^{-1} + (n-1)(\sigma^2)^{-1}I \approx \frac{n-1}{\sigma^2}I$. So $V_{n-1} \approx \frac{\sigma^2}{n-1}I$.
- Apply Kalman equations with $A = I, \Gamma = 0$:
  - $P_{n-1} = V_{n-1}$.
  - $K_n = V_{n-1}(V_{n-1} + \sigma^2 I)^{-1}$.
  - $\mu_n = \mu_{n-1} + K_n(x_n - \mu_{n-1})$.
  - $V_n^{-1} = V_{n-1}^{-1} + (\sigma^2)^{-1}I$.
- Precision update gives $V_n^{-1} \approx \frac{n-1}{\sigma^2}I + \frac{1}{\sigma^2}I = \frac{n}{\sigma^2}I$. So $V_n \approx \frac{\sigma^2}{n}I$.
- Substitute $V_{n-1} \approx \frac{\sigma^2}{n-1}I$ into K_n: $K_n \approx \frac{\sigma^2}{n-1}I(\frac{\sigma^2}{n-1}I + \sigma^2 I)^{-1} = \frac{\sigma^2}{n-1}(\sigma^2(\frac{1}{n-1}+1))^{-1}I = \frac{1}{n-1}(\frac{n}{n-1})^{-1}I = \frac{1}{n}I$.
- Substitute $K_n \approx \frac{1}{n}I$ into mean update: $\mu_n \approx \mu_{n-1} + \frac{1}{n}(x_n - \mu_{n-1}) = \mu_{n-1}(1-\frac{1}{n}) + \frac{1}{n}x_n = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}x_n$.
- Using the inductive hypothesis $\mu_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i$: $\mu_n = \frac{n-1}{n}(\frac{1}{n-1}\sum_{i=1}^{n-1} x_i) + \frac{1}{n}x_n = \frac{1}{n}\sum_{i=1}^{n-1} x_i + \frac{1}{n}x_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- The induction holds. The posterior mean is the average of observations.

Exercise 13.29

Goal: Derive RTS smoother equations (13.100)-(13.102) from backward recursion (13.99).

(Derivation omitted due to algebraic complexity, as noted in the original answer. It involves substituting Gaussian forms into the integral (13.99), relating the result to the desired smoothed marginal $p(z_n|X)$ and the forward-pass marginal $p(z_n|X_{1:n})$, and solving for the parameters of $p(z_n|X)$ recursively.)

The key steps are:

1. Start with $p(z_n|X) = \int p(z_n, z_{n+1}|X)dz_{n+1}$.
2. Use $p(z_n, z_{n+1}|X) = p(z_{n+1}|z_n, X)p(z_n|X) = p(z_{n+1}|z_n)p(z_n|X)$. Wait, this is not right.
3. Use $p(z_n|X) = p(z_n|X_{1:n})\int \frac{p(z_{n+1}|z_n)p(z_{n+1}|X)}{p(z_{n+1}|X_{1:n})}dz_{n+1}$. This involves the prediction $p(z_{n+1}|X_{1:n}) = \mathcal{N}(z_{n+1}|A\mu_n, P_n)$.
4. Substitute the Gaussian forms $\mathcal{N}(z_n|\mu_n, V_n)$ for $p(z_n|X_{1:n})$, $\mathcal{N}(z_{n+1}|\hat{\mu}_{n+1}, \hat{V}_{n+1})$ for $p(z_{n+1}|X)$, and $\mathcal{N}(z_{n+1}|A\mu_n, P_n)$ for the prediction.
5. The integral involves a product/division of Gaussians. Performing the integral and matching the result to $\mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n)$ yields the RTS equations (13.100)-(13.102). Define $J_n = V_n A^T P_n^{-1}$.

Exercise 13.30

Goal: Derive pairwise posterior $\xi(z_{n-1}, z_n)$ (13.103) for LDS from (13.65).

- Equation (13.65) gives the scaled pairwise posterior:

$$\xi(z_{n-1}, z_n) = c_n^{-1}\hat{\alpha}(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\hat{\beta}(z_n)$$

- Substitute the Gaussian forms:

$$\xi(z_{n-1}, z_n) \propto \mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1})\mathcal{N}(x_n|Cz_n, \Sigma)\mathcal{N}(z_n|Az_{n-1}, \Gamma)\hat{\beta}(z_n)$$

- Use $\gamma(z_n) = \hat{\alpha}(z_n)\hat{\beta}(z_n)$ and $\hat{\alpha}(z_n) \propto \mathcal{N}(x_n|Cz_n, \Sigma) \int \mathcal{N}(z_n|Az_{n-1}, \Gamma)\hat{\alpha}(z_{n-1})dz_{n-1}$.
- Also, $p(z_{n-1}, z_n|X) = p(z_n|z_{n-1}, X)p(z_{n-1}|X)$.
- $p(z_{n-1}|X) = \mathcal{N}(z_{n-1}|\hat{\mu}_{n-1}, \hat{V}_{n-1})$.
- $p(z_n|z_{n-1}, X) = p(z_n|z_{n-1}, X_{1:n})$ (using d-separation). This conditional distribution can be found from the Kalman filter steps. It is $\mathcal{N}(z_n|\dots)$.
- Multiplying $p(z_n|z_{n-1}, X_{1:n})$ and $p(z_{n-1}|X)$ gives the joint Gaussian $\xi(z_{n-1}, z_n)$. The form given in (13.103) arises from expressing this joint using the filtered and smoothed marginals.

Exercise 13.31

Goal: Verify the cross-covariance $\text{cov}[z_n, z_{n-1}] = J_{n-1}\hat{V}_n$ (13.104).

- The joint posterior $\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n|X)$ is Gaussian.
- We know the marginals $p(z_{n-1}|X) = \mathcal{N}(z_{n-1}|\hat{\mu}_{n-1}, \hat{V}_{n-1})$ and $p(z_n|X) = \mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n)$.
- From standard RTS smoother derivations (e.g., Rauch et al., 1965, or Bishop App B), the cross covariance between consecutive smoothed states is given by:

$$\text{cov}[z_n, z_{n-1}|X] = \mathbb{E}[(z_n - \hat{\mu}_n)(z_{n-1} - \hat{\mu}_{n-1})^T|X] = J_{n-1}\hat{V}_n$$

  where $J_{n-1} = V_{n-1}A^T P_{n-1}^{-1}$ is the smoother gain used in the backward pass relating z_n to z_{n-1}.
- Alternatively, $\text{cov}[z_{n-1}, z_n|X] = \hat{V}_{n-1}J_{n-1}^T$. Since $\hat{V}_n$ is symmetric, $J_{n-1}\hat{V}_n = (\hat{V}_n J_{n-1}^T)^T$.
- This result arises naturally from the structure of the RTS equations.

Exercise 13.32

Goal: Verify M-step equations (13.110), (13.111) for μ_0, V_0.

- The relevant part of $Q(\theta, \theta^{\text{old}})$ is $L(\mu_0, V_0) = \mathbb{E}_{Z|\theta^{\text{old}}}[\ln p(z_1|\mu_0, V_0)]$.

$$L(\mu_0, V_0) = -\frac{1}{2}\ln|V_0| - \frac{1}{2}\mathbb{E}[(z_1 - \mu_0)^T V_0^{-1}(z_1 - \mu_0)] + \text{const}$$

- This is the expected log likelihood for a Gaussian $p(z_1|\mu_0, V_0)$, where expectation is over the posterior $p(z_1|X, \theta^{\text{old}})$.
- Maximizing w.r.t μ_0, V_0 gives the ML estimates based on the expected sufficient statistics:
  - $\mu_0^{\text{new}} = \mathbb{E}[z_1]$
  - $V_0^{\text{new}} = \mathbb{E}[z_1 z_1^T] - \mathbb{E}[z_1]\mathbb{E}[z_1]^T = \text{cov}[z_1|X]$
- Using the smoother outputs $\mathbb{E}[z_1] = \hat{\mu}_1$ and $\text{cov}[z_1|X] = \hat{V}_1$:
  - $\mu_0^{\text{new}} = \hat{\mu}_1$ (Matches 13.110 if $\mathbb{E}[z_1]$ means $\hat{\mu}_1$)
  - $V_0^{\text{new}} = \hat{V}_1$ (Matches 13.111 if the expression means $\hat{V}_1$)

Exercise 13.33

Goal: Verify M-step equations (13.113), (13.114) for A, Gamma.

- The relevant part of Q is $L(A, \Gamma) = \mathbb{E}[\sum_{n=2}^{N} \ln p(z_n|z_{n-1}, A, \Gamma)]$.

$$L(A, \Gamma) = -\frac{N-1}{2} \ln |\Gamma| - \frac{1}{2} \sum_{n=2}^{N} \mathbb{E}[(z_n - Az_{n-1})^T \Gamma^{-1} (z_n - Az_{n-1})] + \text{const}$$

- Maximize w.r.t A: Treat as weighted linear regression from z_{n-1} to z_n. The standard solution is:

$$A^{\text{new}} = \left(\sum_{n=2}^{N} \mathbb{E}[z_n z_{n-1}^T] \right) \left(\sum_{n=2}^{N} \mathbb{E}[z_{n-1} z_{n-1}^T] \right)^{-1}$$

  This matches (13.113).
- Maximize w.r.t Γ: The ML estimate for covariance is the expected outer product of the residuals:

$$\Gamma^{\text{new}} = \frac{1}{N-1} \sum_{n=2}^{N} \mathbb{E}[(z_n - A^{\text{new}} z_{n-1})(z_n - A^{\text{new}} z_{n-1})^T]$$

  Expanding the expectation:

$$\mathbb{E}[zz^T - Az_{n-1}z_n^T - z_n z_{n-1}^T A^T + Az_{n-1}z_{n-1}^T A^T]$$
$$= \mathbb{E}[z_n z_n^T] - A^{\text{new}}\mathbb{E}[z_{n-1}z_n^T] - \mathbb{E}[z_n z_{n-1}^T](A^{\text{new}})^T + A^{\text{new}}\mathbb{E}[z_{n-1}z_{n-1}^T](A^{\text{new}})^T$$

  Substituting into the expression for Γ^{new} matches (13.114).

Exercise 13.34

Goal: Verify M-step equations (13.115), (13.116) for C, Sigma.

- The relevant part of Q is $L(C, \Sigma) = \mathbb{E}[\sum_{n=1}^{N} \ln p(x_n|z_n, C, \Sigma)]$.

$$L(C, \Sigma) = -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} \mathbb{E}[(x_n - Cz_n)^T \Sigma^{-1} (x_n - Cz_n)] + \text{const}$$

- Maximize w.r.t C: Treat as weighted linear regression from z_n to x_n.

$$C^{\text{new}} = \left(\sum_{n=1}^{N} x_n \mathbb{E}[z_n^T]\right)\left(\sum_{n=1}^{N} \mathbb{E}[z_n z_n^T]\right)^{-1}$$

  This matches (13.115).

- Maximize w.r.t Σ: ML estimate is expected outer product of residuals:

$$\Sigma^{\text{new}} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[(x_n - C^{\text{new}} z_n)(x_n - C^{\text{new}} z_n)^T]$$

  Expanding the expectation (note x_n is fixed):

$$\begin{aligned} &\mathbb{E}[x_n x_n^T - C^{\text{new}} z_n x_n^T - x_n z_n^T (C^{\text{new}})^T + C^{\text{new}} z_n z_n^T (C^{\text{new}})^T] \\ &= x_n x_n^T - C^{\text{new}} \mathbb{E}[z_n] x_n^T - x_n \mathbb{E}[z_n^T](C^{\text{new}})^T + C^{\text{new}} \mathbb{E}[z_n z_n^T](C^{\text{new}})^T \end{aligned}$$

  Substituting into the expression for Σ^{new} matches (13.116).