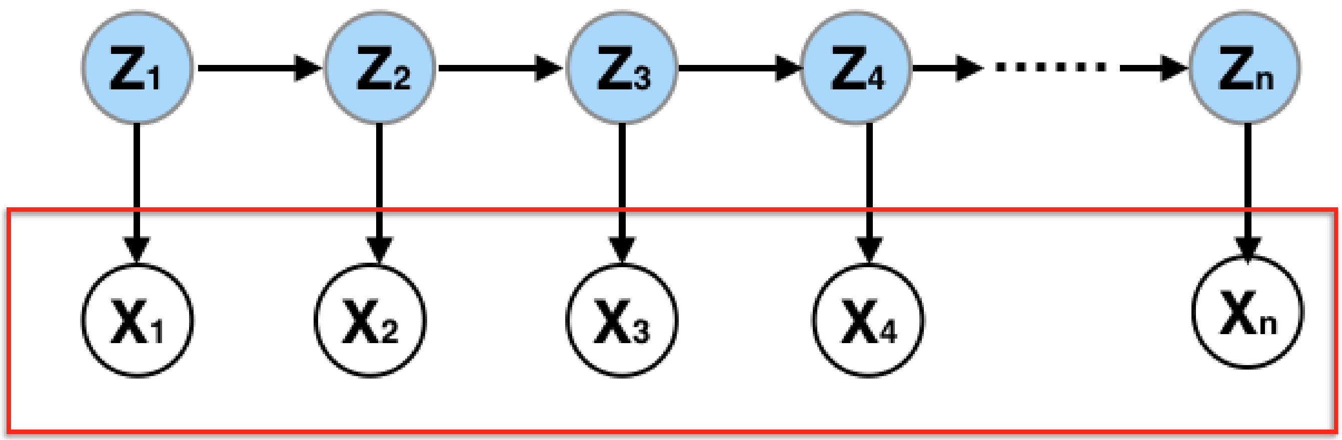
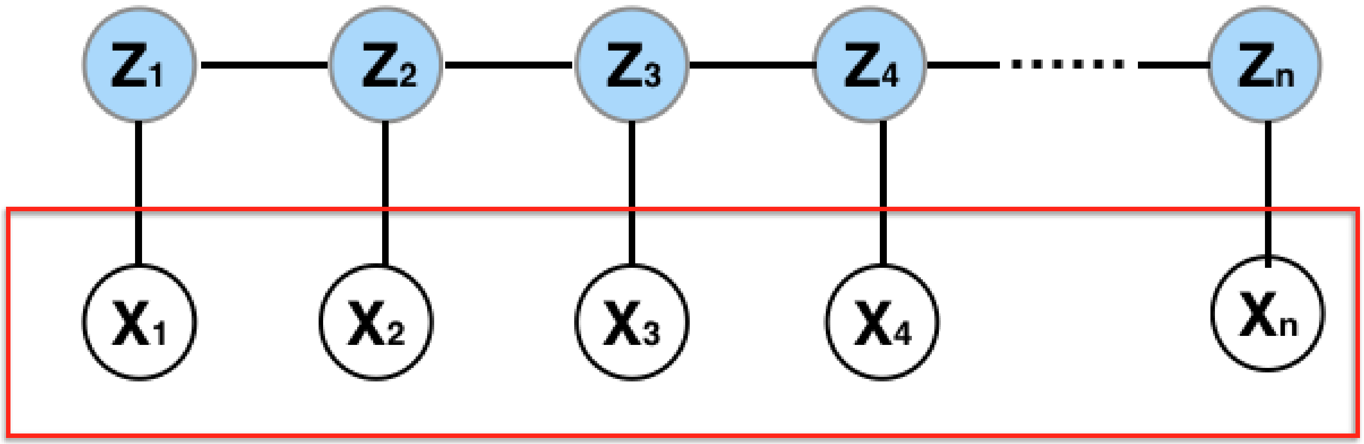
Hidden Markov Model (HMM)



Hidden Markov Model (HMM)



Consider a HMM for $(\mathbf{Z}, \mathbf{X}) = (Z_1, \ldots, Z_n, X_1, \ldots, X_n)$ where X_i's are observed and Z_i's are hidden. Let's assume that both Z and X are discrete random variables, taking m_z and m_x possible values, respectively. So the HMM is parameterized by $\theta = (w, A, B)$ where

- $w_{m_z \times 1}$: distribution for Z_1;
- $A_{m_z \times m_z}$: the transition probability matrix from Z_t to Z_{t+1};
- $B_{m_z \times m_x}$: the probability matrix (the emission distribution) from Z_t to X_t.

Symbols in red are the five elements of a discrete HMM.

Issues

- Forward probabilities: $\alpha_t(i) = p_\theta(x_1, \ldots, x_t, Z_t = i)$
  - How uncertainty is the latent state at time t given all the data till time t?
  - What's the likelihood of the data sequence?

$$p(\mathbf{x}|\theta) = \sum_i p_\theta(x_1, \ldots, x_n, Z_n = i) = \sum_i \alpha_n(i).$$

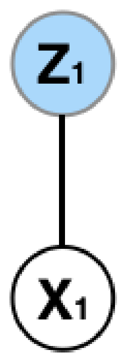
- Backward probabilities: $\beta_t(i) = p_\theta(x_{t+1}, \ldots, x_n | Z_t = i)$
  - Given the latent state at time t, what's our prediction on future data?
- Given $\mathbf{x} = (x_1 \ldots, x_n)$, how to compute the MLE of $\theta = (w, A, B)$?
- Given $\mathbf{x} = (x_1 \ldots, x_n)$ and θ, what can we say about the latent states, i.e. $p_\theta(Z_1, \ldots, Z_n \mid \mathbf{x})$?

Forward Probabilities

The forward probabilities $\alpha_t(i) = p_\theta(x_1, \ldots, x_t, Z_t = i)$ can be calculated recursively.

When $t = 1$,

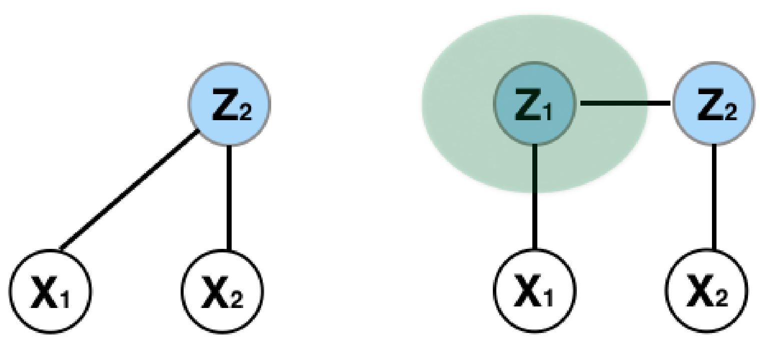
$$\begin{aligned} \alpha_1(i) &= p_\theta(x_1, Z_1 = i) \\ &= p_\theta(Z_1 = i) \times p_\theta(x_1 \mid Z_1 = i) \\ &= w(i)B(i, x_1). \end{aligned}$$



When $t = 1$, $\alpha_1(j) = p_\theta(x_1, Z_1 = j)$ is known.

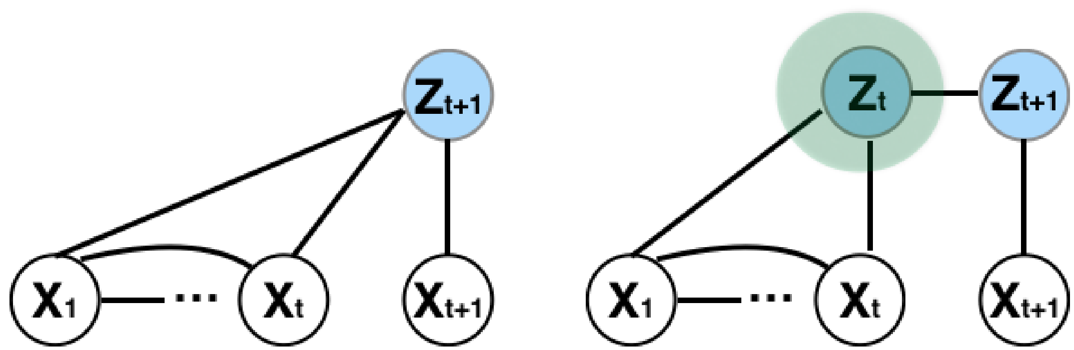
When $t = 2$,

$$
\begin{aligned}
\alpha_2(i) &= p_\theta(x_1, x_2, Z_2 = i) \\
&= \sum_j p_\theta(x_1, x_2, Z_1 = j, Z_2 = i) \\
&= \sum_j p_\theta(x_1, Z_1 = j) \times p_\theta(Z_2 | x_1, Z_1 = j) \times p_\theta(x_2 | x_1, Z_1 = j, Z_2 = i) \\
&= \sum_j \alpha_1(j) A(j, i) B(i, x_2)
\end{aligned}
$$



The forward probabilities $\alpha_t(i) = p_\theta(x_1, \ldots, x_t, Z_t = i)$ can be calculated recursively. Suppose $\alpha_t(j) = p_\theta(x_1, \ldots, x_t, Z_t = j)$ is known.

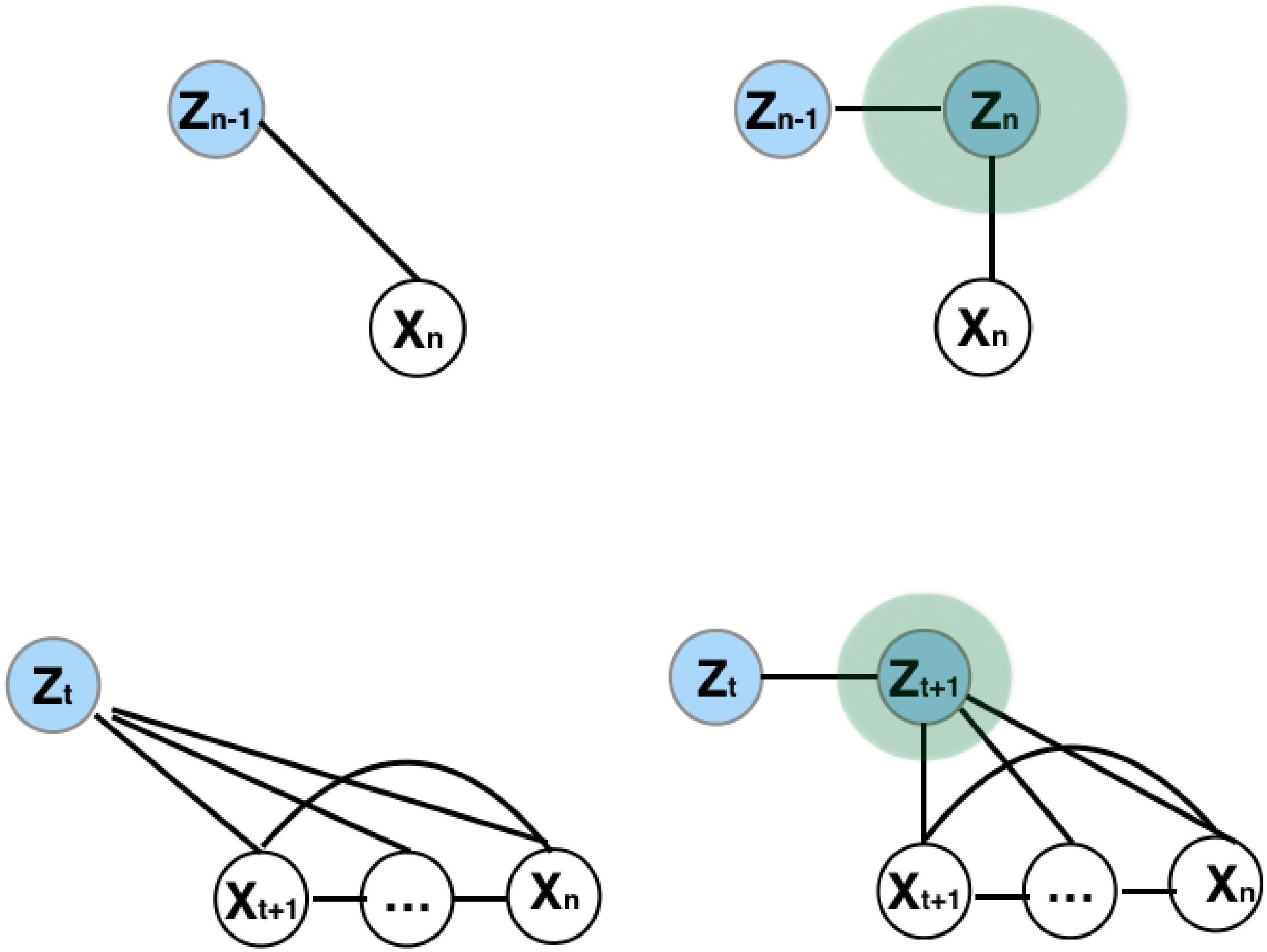
$$\begin{aligned}
\alpha_{t+1}(i) &= p_\theta(x_1, \ldots, x_{t+1}, Z_{t+1} = i) \\
&= \sum_j p_\theta(x_1, \ldots, x_{t+1}, Z_t = j, Z_{t+1} = i) \\
&= \sum_j p_\theta(x_1, \ldots, x_t, Z_t = j) \times p_\theta(Z_{t+1} = i | x_1, \ldots, x_t, Z_t = j) \\
&\quad \times p_\theta(x_{t+1} | x_1, \ldots, x_t, Z_t = j, Z_{t+1} = i) \\
&= \sum_j \alpha_t(j) A(j, i) B(i, x_{t+1})
\end{aligned}$$



Backward Probabilities

The backward probabilities $\beta_t(i) = p_\theta(x_{t+1}, \ldots, x_n | Z_t = i)$ can be calculated recursively.

$$
\begin{aligned}
\beta_{n-1}(i) &= p_\theta(x_n | Z_{n-1} = i) \\
&= \sum_j p_\theta(x_n, Z_n = j | Z_{n-1} = i) \\
&= \sum_j p_\theta(Z_n = j | Z_{n-1} = i) \times p_\theta(x_n | Z_n = j, Z_{n-1} = i) \\
&= \sum_j A(i,j) B(j, x_n) \\
&= \sum_j A(i,j) B(j, x_n) \beta_n(j), \qquad \text{Define } \beta_n(j) = 1 \text{ for all } j
\end{aligned}
$$



Zn-1
Xn
Zn-1
Zn
Xn
Zt
Xt+1
...
Xn
Zt
Zt+1
Xt+1
...
Xn

The backward probabilities $\beta_t(i) = p_\theta(x_{t+1}, \ldots, x_n | Z_t = i)$ can be calculated recursively.

$$\begin{aligned}
\beta_t(i) &= p_\theta(x_{t+1}, \ldots, x_n | Z_t = i) \\
&= \sum_j p_\theta(x_{t+1}, \ldots, x_n, Z_{t+1} = j | Z_t = i) \\
&= \sum_j p_\theta(Z_{t+1} = j | Z_t = i) \times p_\theta(x_{t+1} | Z_{t+1} = j, Z_t = i) \\
&\qquad \times p_\theta(x_{t+2}, \ldots, x_n | x_{t+1}, Z_{t+1} = j, Z_t = i) \\
&= \sum_j A(i,j) B(j, x_{t+1}) \beta_{t+1}(j)
\end{aligned}$$

with

$$\beta_n(i) = 1.$$

The Baum-Welch Algorithm

- The log likelihood on the observed data is given by

$$\log\Big[p(\mathbf{x}|\theta)\Big] = \log\Big[\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}|\theta)\Big],$$

  which is difficult to optimize due to the summation inside the log.

- The log likelihood for the complete data $(\mathbf{Z}, \mathbf{x})$ is given by

$$\begin{aligned} & \log\Big[w(Z_1)\prod_{t=1}^{n-1} A(Z_t, Z_{t+1}) \prod_{t=1}^{n} B(Z_t, x_t)\Big] \\ = \quad & \log w(Z_1) + \sum_{t=1}^{n-1} \log A(Z_t, Z_{t+1}) + \sum_{t=1}^{n} \log B(Z_t, x_t). \end{aligned}$$

To describe the EM (a.k.a. the Baum-Welch) algorithm, define

$$\gamma_t(i,j) = p_\theta(Z_t = i, Z_{t+1} = j|\mathbf{x}), \quad \gamma_t(i) = p_\theta(Z_t = i|\mathbf{x}) = \sum_j \gamma_t(i,j).$$

At the E-step, we have

$$\begin{aligned}
& \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log p(\mathbf{Z},\mathbf{x}|\theta) \\
= \;& \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0}\Big[\log w(Z_1) + \sum_{t=1}^{n-1} \log A(Z_t, Z_{t+1}) + \sum_{t=1}^{n} \log B(Z_t, x_t)\Big] \\
= \;& \sum_{i=1}^{m_z} \gamma_1(i) \log w(i) + \sum_{t=1}^{n-1} \sum_{i,j=1}^{m_z} \gamma_t(i,j) \log A(i,j) + \sum_{t=1}^{n} \sum_{i=1}^{m_z} \gamma_t(i) \log B(i, x_t) \\
= \;& \sum_{i=1}^{m_z} \gamma_1(i) \log w(i) + \sum_{i,j=1}^{m_z} \sum_{t=1}^{n-1} \gamma_t(i,j) \log A(i,j) + \sum_{i=1}^{m_z} \sum_{t=1}^{n} \gamma_t(i) \log B(i, x_t).
\end{aligned}$$

$$
\begin{aligned}
& \sum_{i=1}^{m_z} \gamma_1(i) \log w(i) + \sum_{i,j=1}^{m_z} \sum_{t=1}^{n-1} \gamma_t(i,j) \log A(i,j) + \sum_{i=1}^{m_z} \sum_{t=1}^{n} \gamma_t(i) \log B(i, x_t) \\
= \ & \sum_{i=1}^{m_z} \gamma_1(i) \log w(i) + \sum_{i}^{m_z} \Big[\sum_{j=1}^{m_z} \gamma_+(i,j) \log A(i,j) \Big] + \\
& \qquad \sum_{i=1}^{m_z} \sum_{l=1}^{m_x} \Big(\sum_{t:x_t=l} \gamma_t(i) \Big) \log B(i,l) \qquad (*)
\end{aligned}
$$

At the M-step, when updating the parameters (w, A, B), we will repeatedly use the following result (which we have proved when discussing two component Gaussian mixturers): consider the following function of $(w_1, \ldots, w_m)$:

$$
J(\mathbf{w}) = a_1 \log w_1 + a_2 \log w_2 + \cdots + a_m \log w_m
$$

where $a_j \geq 0$, and $(w_1, \ldots, w_m)$ is a probability vector (i.e., $0 \leq w_j \leq 1$ and $\sum_j w_j = 1$). The maximum of $J(\mathbf{w})$ is achieved by $w_j = a_j / \sum_{j'} a_{j'}$.

- Update w: The maximum of $\sum_{i=1}^{m_z} \gamma_1(i) \log w(i)$ is achieved by

$$w^*(i) = \gamma_1(i), \quad i = 1, \ldots, m_z.$$

Note that $\sum_i \gamma_1(i) = 1$.

- Update $A_{m_z \times m_z}$: Note that each row of A, $\{A(i,j)\}_{j=1}^{m_z}$, is a probability vector, so

$$\sum_{j=1}^{m_z} \gamma_+(i,j) \log A(i,j)$$

is maximized by

$$A^*(i,j) = \frac{\gamma_+(i,j)}{\sum_{j'} \gamma_+(i,j')} = \frac{\sum_{t=1}^{n-1} \gamma_t(i,j)}{\sum_{j'} \sum_{t=1}^{n-1} \gamma_t(i,j')}, \quad i,j = 1, \ldots, m_z.$$

- Update $B_{m_z \times m_x}$: Note that each row of B, $\{B(i,l)\}_{l=1}^{m_x}$, is a probability vector, so

$$\sum_{l=1}^{m_x} \Big(\sum_{t: x_t = l} \gamma_t(i) \Big) \log B(i,l)$$

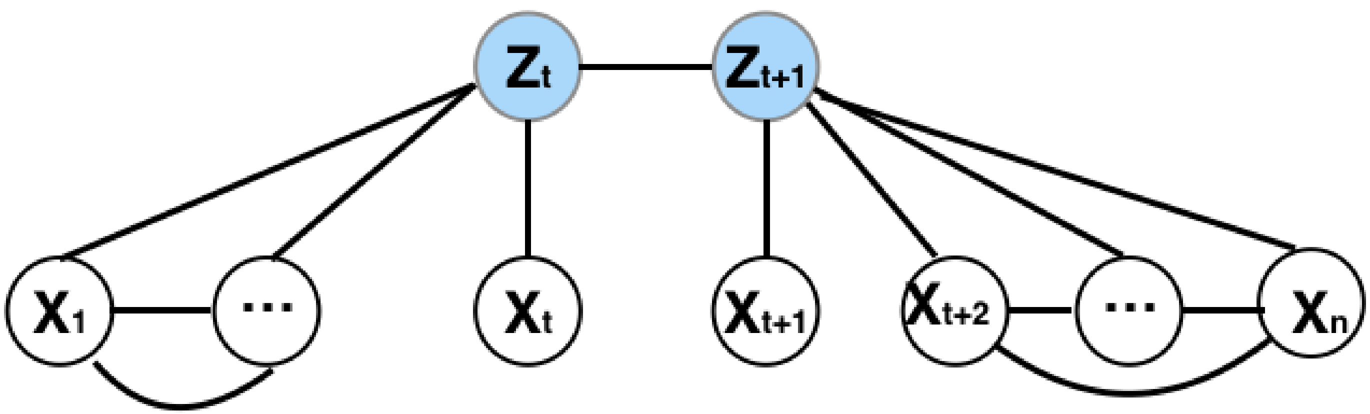
is maximized by

$$B^*(i,l) = \frac{\sum_{t: x_t = l} \gamma_t(i)}{\sum_t \gamma_t(i)}, \quad i = 1, \ldots, m_z,\ l = 1, \ldots, m_x.$$

- In the coding assignment, you will be asked not to update $w^*(i)$.

How to calculate $\gamma_t(i,j)$?

$$
\begin{aligned}
& p_\theta(Z_t = i, Z_{t+1} = j|\mathbf{x}) \\
\propto \quad & p_\theta(\mathbf{x}_{1:t}, Z_t = i, Z_{t+1} = j, x_{t+1}, \mathbf{x}_{t+2:n}) \\
= \quad & p_\theta(\mathbf{x}_{1:t}, Z_t = i) \times p_\theta(Z_{t+1} = j|Z_t = i) \\
& \quad \times p_\theta(x_{t+1}|Z_{t+1} = j) \times p_\theta(\mathbf{x}_{t+2:n}|Z_{t+1} = j) \\
= \quad & \alpha_t(i) A(i,j) B(j, x_{t+1}) \beta_{t+1}(j),
\end{aligned}
$$



Inference on the Hidden States $\mathbf{Z}$

There are several ways to find the "optimal" hidden state sequence $\mathbf{Z}$ given an observation sequence $\mathbf{x}$, depending on the definition of "optimality".

- One optimality criterion is to choose the hidden states Z_t's that are *individually* most likely, that is,

$$\hat{Z}_t^* = \arg\max_i \, p_\theta(Z_t = i|\mathbf{x}) = \arg\max_i \gamma_t(i).$$

  Here we can plug in $\hat{\theta}$ from the aforementioned EM algorithm. Such a solution is optimal in the sense that it maximizes the expected number of correct states (by choosing the most likely state for each t). However, the resulting sequence may not the most likely one and it may not even be a valid sequence, for example, $\hat{Z}_t^* = 1$ and $\hat{Z}_{t+1}^* = 2$, but $A_{12} = 0$.

- An alternative approach is to find the most likely *single* sequence (or path), that is,

$$\hat{\mathbf{Z}}^* = \arg\max_{i_1,\ldots,i_n} p_\theta(Z_1 = i_1, \ldots, Z_n = i_n|\mathbf{x}).$$

The solution is obtained via a dynamic programming method, known as the Viterbi algorithm. Define

$$\delta_t(i) = \max_{j_1,\ldots,j_{t-1}} p_\theta(Z_1 = j_1, \ldots, Z_{t-1} = j_{t-1}, Z_t = i, \mathbf{x}_{1:t}),$$

which is the highest probability along a single path from time 1 to t, which accounts for the first t observations $\mathbf{x}_{1:t}$ and ends in a hidden state $Z_t = i$. In particular

$$\delta_1(i) = p_\theta(Z_1 = i, x_1) = w(i)B(i, x_1).$$

By induction we have

$$
\begin{aligned}
\delta_{t+1}(i) &= \max_{j_{1:(t-1)},j} p_\theta(Z_1 = j_1, \ldots, Z_{t-1} = j_{t-1}, Z_t = j, Z_{t+1} = i, \mathbf{x}_{1:(t+1)}) \\
&= \max_{j_{1:(t-1)},j} \Big[p_\theta(Z_1 = j_1, \ldots, Z_{t-1} = j_{t-1}, Z_t = j, \mathbf{x}_{1:t}) \times \\
&\qquad p_\theta(Z_{t+1} = i | Z_t = j) \times p_\theta(x_{t+1} | Z_{t+1} = i) \Big] \\
&= \Big[\max_j \delta_t(j) A(j,i) \Big] B(i, x_{t+1}).
\end{aligned}
$$

Note that the recursive formula above is similar to the one for $\alpha_t(i)$. The major difference is that the maximization over previous states is used for δ_t and the integration/summation is used for α_t.

- Next we solve for the most likely *single* sequence $\hat{\mathbf{Z}}^*$ backward

$$\hat{\mathbf{Z}}^* = \arg\max_{i_1,\ldots,i_n} p_\theta(Z_1 = i_1, \ldots, Z_n = i_n | \mathbf{x})$$

  - The best value for $\hat{Z}_n^*$. $\delta_n(i)$ stores the highest probability of a $\mathbf{Z}$ sequence that ends with $Z_n = i$. So

$$\hat{Z}_n^* = \arg\max_i \delta_n(i).$$

  - The best value for $\hat{Z}_{n-1}^*$. Note that we have already known $\hat{Z}_n^* = j_n^*$:

$$\hat{Z}_{n-1}^* = \arg\max_i \left[\delta_{n-1}(i) A(i, j_n^*)\right]$$

  - The best value for $\hat{Z}_{t-1}^*$, given we have known the optimal values for $\hat{Z}_{t:n}^*$.

$$\hat{Z}_{t-1}^* = \arg\max_i \left[\delta_{t-1}(i) A(i, j_t^*)\right]$$