TTIC 31170:
Robot Learning and Estimation

Spring 2021

Matthew Walter
TTI-Chicago

Lecture 3: Recursive Bayesian Estimation

Office hours

- **Matt**: Wednesdays, 4pm–5pm CT
- **Shengjie**: Mondays, 10am–11am CT

Held over Zoom using the same meeting as lectures

Topics

- Stochastic Processes
- Markov Assumption
- Hidden Markov Models
  - Forward-backward algorithm
  - Viterbi Algorithm

Stochastic (random) process

Definition: A ***stochastic process*** is a sequence (discrete-time) or waveform (continuous-time) of random variables

	Discrete RV	**Continuous RV**
Discrete Time	*discrete random sequence* $X[n, w) \quad n \in \{0, \pm 1, \pm 2, \ldots\}$	*discrete random process* $x[n, w) \quad n \in \{0, \pm 1, \pm 2, \ldots\}$
Continuous Time	*continuous random sequence* $X(t, w) \quad t \in \mathbb{R}^+$	*continuous random process* $x(t, w) \quad t \in \mathbb{R}^+$

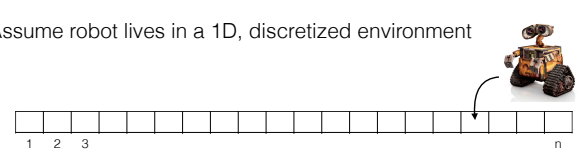
Stochastic (random) process

Definition: A ***stochastic process*** is a sequence (discrete-time) or waveform (continuous-time) of random variables

- With $w \in \Omega$ fixed, $X[n, w) = X[n]$ and $X(t, w) = X(t)$ represent *sample paths* of the process
- Stochastic processes are often viewed as functions of a single variable viewed as "time", but they need not be, e.g.,
  - Image intensity as a function of two spatial variables
  - Video as a function of two spatial and one temporal variable

Example: Robot localization

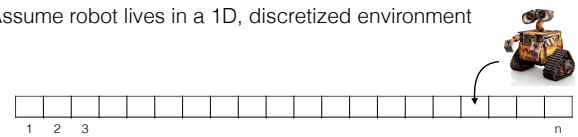
- Assume robot lives in a 1D, discretized environment



$X_t \in \{1, 2, 3, \ldots, n\}$ Robot's position at time t

Example: Robot localization

- Assume robot lives in a 1D, discretized environment



$X_t \in \{1, 2, 3, \ldots, n\}$ Robot's position at time t

- The robot's motion is uncertain (e.g., wheel slip, model error)
- The robot's trajectory is a stochastic process

$$X^T = \{X_0, X_1, X_2, \ldots, X_T\}$$

we use the superscript T to denote the history of up to time T

Note: Slight abuse of notation $X_t = X[t]$

Example: Language modeling

- Let X_i be the word at position i in a spoken utterance

$$X_0 = \text{“}wreck\text{”} \quad X_1 = \text{“}a\text{”} \quad X_2 = \text{“}nice\text{”} \quad X_3 = \text{“}beach\text{”}$$

$$X^T = \{X_0, X_1, X_2, \ldots, X_T\}$$

- Applications:
  - Speech recognition
  - Machine translation

Example: Binary communication

- Consider a sequence of binary numbers

$$X_t \in \{0, 1\}$$

transmitted over a (noisy) channel

$$0010011001000101000001000011101101010100$$

$$X^T = \{X_0, X_1, X_2, \ldots, X_T\}$$

Stochastic processes

- Since the RV is a function of time, so are its distribution (density)

$$F_x(x, t) = P(x(t) \le x) = P(x_1(t) \le x_1, x_2(t) \le x_2, \ldots, x_n(t) \le x_n)$$

$$f_x(x, t) = \frac{dF_x(x, t)}{dx} = \frac{d^n F_x(x, t)}{dx_1, \cdots, dx_n}$$

- Since $x(t_1)$ and $x(t_2)$ are random variables, we can consider the joint distribution (density)

$$F_x(x_1, x_2, t_1, t_2) = P(x(t_1) \le x_1, x(t_2) \le x_2)$$

$$f_x(x, t) = \frac{dF(x_1, x_2, t_1, t_2)}{dx_1 dx_2}$$

First- and second-order statistics

- First and second moments of continuous-time stochastic process $x(t)$ consists of
  - *mean function*

$$\mu_x(t) = \mathbb{E}[x(t)]$$

  - *autocovariance function*

$$K_{xx}(t, s) = \text{cov}(x(t), x(s)) = \mathbb{E}[(x(t) - \mu_x(t))(x(s) - \mu_x(s))^\top]$$

  - alternatively, the *autocorrelation function*

$$\begin{aligned} R_{xx}(t, s) &= \mathbb{E}[x(t)x^\top(s)] \\ &= K_{xx}(t, s) + \mu_x(t)\mu_x^\top(s) \end{aligned}$$

First- and second-order statistics

- First and second moments of discrete-time stochastic process consists of $x[n]$
  - *mean function*

$$\mu_x[n] = \mathbb{E}[x[n]]$$

  - *autocovariance function*

$$K_{xx}[n, m] = \text{cov}(x[n], x[m]) = \mathbb{E}[(x[n] - \mu_x[n])(x[m] - \mu_x[m])^\top]$$

  - alternatively, the *correlation function*

$$\begin{aligned} R_{xx}[n, m] &= \mathbb{E}[x[n]x^\top[m]] \\ &= K_{xx}[n, m] + \mu_x[n]\mu_x^\top[m] \end{aligned}$$

Example: Random walk

- Consider a *discrete-time random walk*

$$x[n] = x[n-1] + w[n]$$

where $x[0]$ is deterministic and $w[n]$ is an i.i.d. stochastic sequence

$$\mathbb{E}\left[w[n]\right] = \mu_w$$
$$\mathbb{E}\left[(w[n]-\mu_w)(w[m]-\mu_w)\right] = \sigma^2\delta[n-m]$$

- We can express $x[n] = x[0] + \sum_{k=1}^{n} w[k]$, in which case

$$\mathbb{E}\left[x[n]\right] = n\mu_w \quad \text{cov}\left(x[n], x[n]\right) = n\sigma^2$$

Example: Random walk

- Consider the interval $x[n] - x[m]$ where $n > m$

$$x[n] - x[m] = \sum_{k=m+1}^{n} w[k]$$

- The mean and covariance follow as

$$\mathbb{E}\left[x[n] - x[m]\right] = (n-m)\mu_w$$
$$\text{cov}\left(x[n]-x[m], x[n]-x[m]\right) = (n-m)\sigma^2$$

- $(x[n_2] - x[n_1])$ and $(x[n_4] - x[n_3])$ for $n_4 > n_3 \geq n_2 > n_1$ are *independent (uncorrelated) increments*

$$\text{cov}\left(x[n]-x[m], x[k]\right) = 0 \quad \forall n > m \geq k$$

Coviariance functions

- A function $F(t,s)$ is positive semi-definite if for all $a(t)$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(t)F(t,s)a(s)\,dt\,ds \geq 0$$

- A bivariate function $F(t,s)$ is a valid covariance function if and only if $F(t,s)$ is positive semidefinite

Coviariance functions

- A function $F(t,s)$ is positive semi-definite if for all $a(t)$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(t)F(t,s)a(s)\,dt\,ds \geq 0$$

- A bivariate function $F(t,s)$ is a valid covariance function if and only if $F(t,s)$ is positive semidefinite
- If we choose $a(t) = \sum_{i=1}^{N} a_i\delta(t-t_i)$

then

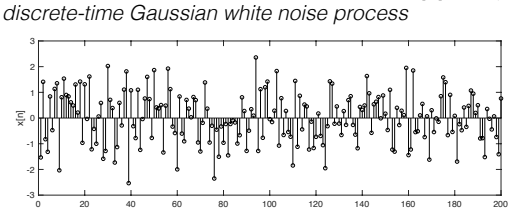
$$K_{xx}(t,s) = \begin{bmatrix} K_{xx}(t_1,t_1) & K_{xx}(t_1,t_2) & \cdots & K_{xx}(t_1,t_N) \\ K_{xx}(t_2,t_1) & K_{xx}(t_2,t_2) & \cdots & K_{xx}(t_2,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ K_{xx}(t_N,t_1) & K_{xx}(t_N,t_2) & \cdots & K_{xx}(t_N,t_N) \end{bmatrix} \geq 0$$

Example: White noise

- A *discrete-time white noise processes* has samples that are uncorrelated but can have arbitrary means, e.g.,

$$\mathbb{E}[x[n]] = 0$$
$$\mathbb{E}[x[n]x[m]] = \sigma^2\delta[n-m]$$

- When $x[n]$ are independent and Gaussian, i.e., $x[n] \sim \mathcal{N}(0,\sigma^2)$, this is a *discrete-time Gaussian white noise process*



Least squares estimation

- Let $\mathbf{x} \in \mathbb{R}^n$ be a constant unknown vector
- Suppose that we have m noisy observations that are linear in $\mathbf{x}$

$$z_1 = H_{11}x_1 + H_{12}x_2 + \ldots + H_{1n}x_n + v_1$$
$$z_2 = H_{21}x_1 + H_{22}x_2 + \ldots + H_{2n}x_n + v_2$$
$$\vdots$$
$$z_m = H_{m1}x_1 + H_{m2}x_2 + \ldots + H_{mn}x_n + v_m$$

where v_i denotes measurement noise

- We can represent this in matrix form as

$$\mathbf{z} = H\mathbf{x} + \mathbf{v}$$

where $\mathbf{z} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^m$

Least squares estimation

- Find the best estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$. What is "best"?

> "The most probable value of the unknown quantities will be that in which the sum of the squares of the differences between the actually observed and the computed values multiplied by numbers that measure the degree of precision is a minimum."
>
> – Carl Friedrich Gauss, 1809*

*Carl Friedrich Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, 1809. However, Gauss claimed that he started working on this in 1795 (when he was only 18 years old).

Least squares estimation

- Find the best estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$. What is "best"?

> "The most probable value of the unknown quantities will be that in which the sum of the squares of the differences between the actually observed and the computed values multiplied by numbers that measure the degree of precision is a minimum."
>
> – Carl Friedrich Gauss, 1809*

- Define the measurement error (*residual*) based on the estimate as

$$\epsilon_z = \mathbf{z} - H\hat{\mathbf{x}}$$

- Following Gauss' advise, this gives rise to the objective

$$J = \epsilon_z^\top \epsilon_z$$

*Carl Friedrich Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, 1809. However, Gauss claimed that he started working on this in 1795 (when he was only 18 years old).

Least squares estimation

$$\begin{aligned} J &= (\mathbf{z} - H\hat{\mathbf{x}})^\top(\mathbf{z} - H\hat{\mathbf{x}}) \\ &= \mathbf{z}^\top\mathbf{z} - \hat{\mathbf{x}}^\top H^\top\mathbf{z} - \mathbf{z}^\top H\hat{\mathbf{x}} + \hat{\mathbf{x}}^\top H^\top H\hat{\mathbf{x}} \end{aligned}$$

- To minimize, we take the partial derivative w.r.t. $\hat{\mathbf{x}}$

$$\frac{\partial J}{\partial \hat{\mathbf{x}}} = -H^\top\mathbf{z} - \mathbf{z}^\top H + 2\hat{\mathbf{x}}^\top H^\top H$$

Least squares estimation

$$\begin{aligned} J &= (\mathbf{z} - H\hat{\mathbf{x}})^\top(\mathbf{z} - H\hat{\mathbf{x}}) \\ &= \mathbf{z}^\top\mathbf{z} - \hat{\mathbf{x}}^\top H^\top\mathbf{z} - \mathbf{z}^\top H\hat{\mathbf{x}} + \hat{\mathbf{x}}^\top H^\top H\hat{\mathbf{x}} \end{aligned}$$

- To minimize, we take the partial derivative w.r.t. $\hat{\mathbf{x}}$

$$\frac{\partial J}{\partial \hat{\mathbf{x}}} = -H^\top\mathbf{z} - \mathbf{z}^\top H + 2\hat{\mathbf{x}}^\top H^\top H$$

- Setting this equal to zero yields

$$\begin{aligned} H^\top H\hat{\mathbf{x}} &= H^\top\mathbf{z} \\ \hat{\mathbf{x}} &= \left(H^\top H\right)^{-1} H^\top\mathbf{z} \end{aligned}$$

- $H^\dagger = \left(H^\top H\right)^{-1} H^\top$ exists if H is full rank ($m \geq n$ & z_i are indep.)

Least squares estimation: Properties

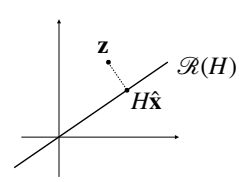
$$\hat{\mathbf{x}} = \left(H^\top H\right)^{-1} H^\top\mathbf{z}$$

- $\hat{\mathbf{x}}$ is a linear function of $\mathbf{z}$

> The least-squares estimator is not necessarily linear in the measurements

- $\hat{\mathbf{x}} = H^{-1}\mathbf{z}$ if H is square and full rank
- $H^\dagger = \left(H^\top H\right)^{-1} H^\top$ is the *pseudo-inverse* of H
- $H^\dagger$ is the *left inverse* of (full-rank, skinny) H

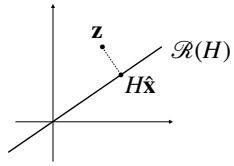
Least squares estimation



- Let $\mathscr{R}(H)$ be the *range* of H (span of columns of H)
- $\hat{\mathbf{x}}$ solves $H\hat{\mathbf{x}} = \mathbf{z}$ iff $\mathbf{z} \in \mathscr{R}(H)$
- $H\hat{\mathbf{x}}$ is the point in $\mathscr{R}(H)$ closest to $\mathbf{z}$, i.e., projection of $\mathbf{z}$ onto $\mathscr{R}(H)$

$$\mathscr{P}_{\mathscr{R}(H)}(\mathbf{z}) = H\hat{\mathbf{x}} = \underbrace{H\left(H^\top H\right)^{-1} H^\top}_{\textit{projection matrix}}\mathbf{z}$$

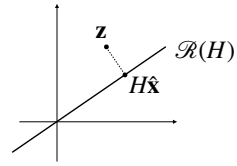
Least squares estimation



- The residual associated with $\hat{\mathbf{x}}$: $\epsilon_z = \mathbf{z} - H\left(H^\top H\right)^{-1} H^\top \mathbf{z}$

is orthogonal to $\mathscr{R}(H)$: i.e., $\epsilon_z^\top (H\mathbf{y}) = 0 \ \forall \mathbf{y} \in \mathbb{R}^n$

Least squares estimation

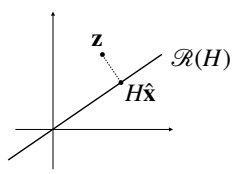


- The residual associated with $\hat{\mathbf{x}}$: $\epsilon_z = \mathbf{z} - H\left(H^\top H\right)^{-1} H^\top \mathbf{z}$

is orthogonal to $\mathscr{R}(H)$: i.e., $\epsilon_z^\top (H\mathbf{y}) = 0 \ \forall \mathbf{y} \in \mathbb{R}^n$

$$\mathbf{z}^\top \left(I - H\left(H^\top H\right)^{-1} H^\top\right) H\mathbf{y} = \mathbf{z}^\top \left(H - H\left(H^\top H\right)^{-1} H^\top H\right) \mathbf{y}$$

Least squares estimation



- The residual associated with $\hat{\mathbf{x}}$: $\epsilon_z = \mathbf{z} - H\left(H^\top H\right)^{-1} H^\top \mathbf{z}$

is orthogonal to $\mathscr{R}(H)$: i.e., $\epsilon_z^\top (H\mathbf{y}) = 0 \ \forall \mathbf{y} \in \mathbb{R}^n$

$$\begin{aligned}\mathbf{z}^\top \left(I - H\left(H^\top H\right)^{-1} H^\top\right) H\mathbf{y} &= \mathbf{z}^\top \left(H - H\left(H^\top H\right)^{-1} H^\top H\right) \mathbf{y} \\ &= \mathbf{z}^\top (H - HI)\,\mathbf{y} \quad \text{(since } H^\dagger \text{ is the left inverse)} \\ &= 0\end{aligned}$$

Exercise

- Let x be a scalar and assume that

$$\begin{aligned} z_1 &= x + v_1 \\ z_2 &= x + v_2 \\ &\vdots \\ z_m &= x + v_m \end{aligned}$$

which we can express as $\mathbf{z} = Hx + \mathbf{v}$ where $H = [1 \quad 1 \quad \cdots \quad 1]^\top$

- Derive the optimal estimator $\hat{x} = \left(H^\top H\right)^{-1} H^\top \mathbf{z}$

Exercise

- Let x be a scalar and assume that

$$\begin{aligned} z_1 &= x + v_1 \\ z_2 &= x + v_2 \\ &\vdots \\ z_m &= x + v_m \end{aligned}$$

which we can express as $\mathbf{z} = Hx + \mathbf{v}$ where $H = [1 \quad 1 \quad \cdots \quad 1]^\top$

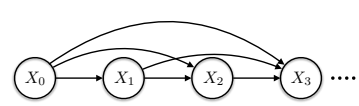
- Derive the optimal estimator $\hat{x} = \left(H^\top H\right)^{-1} H^\top \mathbf{z}$

$$\hat{x} = \frac{1}{m} \sum_{i=1}^{n} z_i$$

Full probability model

- Model joint distribution over sequence
- Assume full dependencies between values in sequence

$$P(X_0, X_1, \ldots, X_T) = P(X_0)P(X_1|X_0)P(X_2|X_1, X_0)P(X_3|X_2, X_1, X_0)\cdots$$



Markov assumption

Definition: A stochastic process exhibits the ***Markov property*** if future states are (conditionally) independent of the past states given present states:

$$P(X_{t+1}|X_t, X_{t-1}, \dots X_0) = P(X_{t+1}|X_t)$$

first-order Markov

Markov assumption

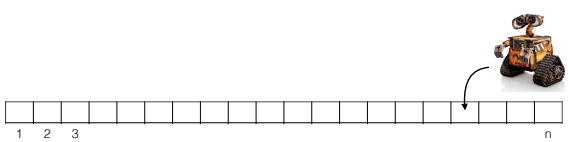
Definition: A stochastic process exhibits the ***Markov property*** if future states are (conditionally) independent of the past states given present states:

$$P(X_{t+1}|X_t, X_{t-1}, \dots X_0) = P(X_{t+1}|X_t)$$

first-order Markov

$$\begin{aligned} P(X_T, X_{T-1}, \dots, X_0) &= P(X_T|X_{T-1})P(X_{T-1}|X_{T-2}) \cdots P(X_1|X_0) \cdot P(X_0) \\ &= \prod_{t=0}^{t=T} P(X_t|X_{t-1}) \end{aligned}$$

Example: Robot localization



$$P(X_t = j|X_{t-1} = i) = \begin{cases} 0.2 & j = i+1 \\ 0.6 & j = i \\ 0.2 & j = i-1 \\ 0.0 & \text{else} \end{cases}$$

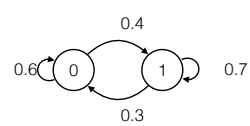
Example: Binary communication

- Consider a sequence of binary numbers

$$X_t \in \{0, 1\}$$

transmitted over a (noisy) channel

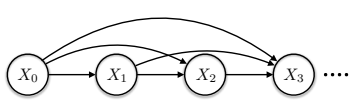
$$0010011001000101000001000011101101010100$$



Markov assumption

- Assume (first-order) Markov assumption holds

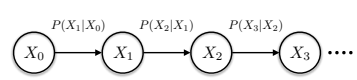
$$P(X_0, X_1, \dots, X_T) = P(X_0)P(X_1|X_0)P(X_2|X_1, X_0)P(X_3|X_2, X_1, X_0) \cdots$$



Markov assumption

- Assume (first-order) Markov assumption holds

$$P(X_0, X_1, \dots, X_T) = P(X_0)P(X_1|X_0)P(X_2|X_1)P(X_3|X_2) \cdots$$

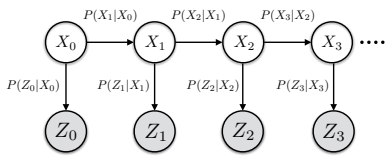


Hidden Markov Model

- We have noisy observations Z_t of each state X_t

$$Z_t = X_t + \text{noise}$$

- Hidden Markov Model (HMM): Markov models with state dependent observations



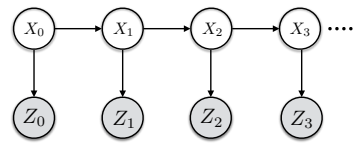
"transition distribution"

$P(X_t|X_{t-1})$

"observation (emission) distribution"

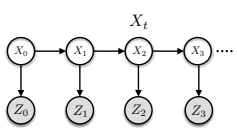
$P(Z_t|X_t)$

Hidden Markov Model



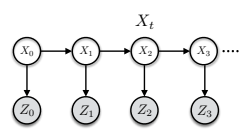
- How do we evaluate the conditional likelihood at individual instances in time?
  - Forward-backward algorithm
- How do we infer the most likely sequence of hidden states given the observation sequence?
  - Viterbi algorithm

Inference in HMMs



$$P(X_t|Z^T) = \frac{P(X_t, Z^T)}{P(Z^T)}$$

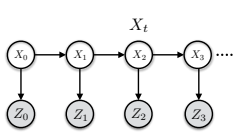
Inference in HMMs



$$P(X_t|Z^T) = \frac{P(X_t, Z^T)}{P(Z^T)}$$

$$P(X_t, Z^T) = P(X_t, Z_0, Z_1, \ldots, Z_t, \ldots, Z_T)$$

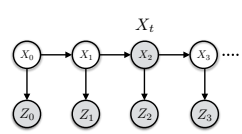
Inference in HMMs



$$P(X_t|Z^T) = \frac{P(X_t, Z^T)}{P(Z^T)}$$

$$\begin{aligned} P(X_t, Z^T) &= P(X_t, Z_0, Z_1, \ldots, Z_t, \ldots, Z_T) \\ &= P(Z_{t+1}, \ldots, Z_T | X_t, Z_0, Z_1, \ldots, Z_t) P(X_t, Z_0, Z_1, \ldots Z_t) \end{aligned}$$

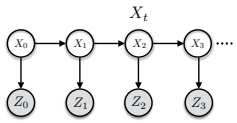
Inference in HMMs



$$P(X_t|Z^T) = \frac{P(X_t, Z^T)}{P(Z^T)}$$

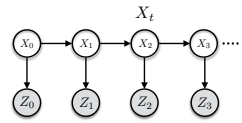
$$\begin{aligned} P(X_t, Z^T) &= P(X_t, Z_0, Z_1, \ldots, Z_t, \ldots, Z_T) \\ &= P(Z_{t+1}, \ldots, Z_T | X_t, Z_0, Z_1, \ldots, Z_t) P(X_t, Z_0, Z_1, \ldots Z_t) \\ &\overset{\text{CI}}{=} P(Z_{t+1}, \ldots, Z_T | X_t) P(X_t, Z_0, Z_1, \ldots, Z_t) \end{aligned}$$

Inference in HMMs



$$P(X_t|Z^T) = \frac{P(X_t, Z^T)}{P(Z^T)}$$
$$= \frac{P(X_t, Z_0, Z_1, \ldots, Z_t)P(Z_{t+1}, \ldots, Z_T|X_t)}{P(Z^T)}$$

Inference in HMMs

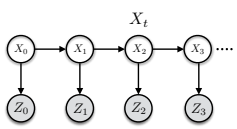


$$P(X_t|Z^T) = \frac{P(X_t, Z^T)}{P(Z^T)}$$
$$= \frac{P(X_t, Z_0, Z_1, \ldots, Z_t)P(Z_{t+1}, \ldots, Z_T|X_t)}{P(Z^T)}$$

$$\alpha(X_t) = P(X_t, Z_0, Z_1, \ldots, Z_t)$$

$$\beta(X_t) = P(Z_{t+1}, \ldots, Z_T|X_t)$$

Inference in HMMs

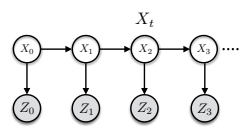


$$\alpha(X_t) = P(X_t, Z_0, Z_1, \ldots, Z_t)$$

$$\beta(X_t) = P(Z_{t+1}, \ldots, Z_T|X_t)$$

$$P(X_t|Z^T) = \frac{\alpha(X_t)\beta(X_t)}{P(Z^T)}$$

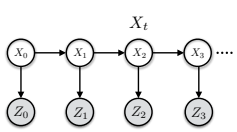
Inference in HMMs



$$\sum_{X_t} P(X_t|Z^T)$$

For any $t \in \{1, 2, \ldots, T\}$

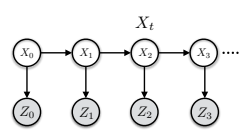
Inference in HMMs



$$\sum_{X_t} P(X_t|Z^T) = \sum_{X_t} \frac{\alpha(X_t)\beta(X_t)}{P(Z^T)} = 1$$

For any $t \in \{1, 2, \ldots, T\}$

Inference in HMMs



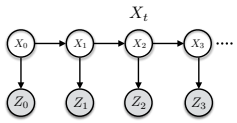
$$\sum_{X_t} P(X_t|Z^T) = \sum_{X_t} \frac{\alpha(X_t)\beta(X_t)}{P(Z^T)} = 1 \Rightarrow$$

Probability of the observation sequence

$$P(Z^T) = \sum_{X_t} \alpha(X_t)\beta(X_t)$$

For any $t \in \{1, 2, \ldots, T\}$

Forward-backward algorithm

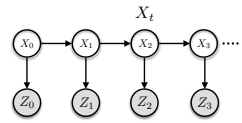


$$\alpha(X_t) = P(X_t, Z_0, Z_1, \ldots, Z_t)$$

$$\beta(X_t) = P(Z_{t+1}, \ldots, Z_T | X_t)$$

$$P(X_t | Z^T) = \frac{\alpha(X_t)\beta(X_t)}{P(Z^T)}$$

Forward-backward algorithm



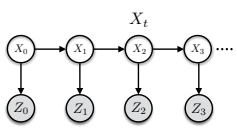
- Forward probabilities (messages):

$$\alpha_t(i) = P(Z_0, Z_1, \ldots, Z_t, X_t = i)$$

- Backward probabilities (messages):

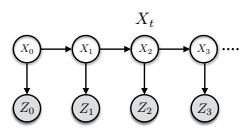
$$\beta_t(i) = P(Z_{t+1}, \ldots, Z_T | X_t = i)$$

Forward probabilities



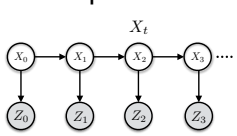
$$\alpha_t(i) = P(Z_0, \ldots, Z_t, X_t = i)$$

Forward probabilities



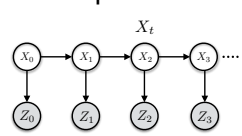
$$\begin{aligned}\alpha_t(i) &= P(Z_0, \ldots, Z_t, X_t = i) \\ &= \sum_{X_{t-1}} P(Z_0, \ldots, Z_t, X_{t-1}, X_t = i)\end{aligned}$$

Forward probabilities



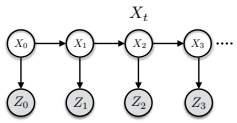
$$\begin{aligned}\alpha_t(i) &= P(Z_0, \ldots, Z_t, X_t = i) \\ &= \sum_{X_{t-1}} P(Z_0, \ldots, Z_t, X_{t-1}, X_t = i) \\ &= \sum_{X_{t-1}} P(Z_t | Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i) P(Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i)\end{aligned}$$

Forward probabilities



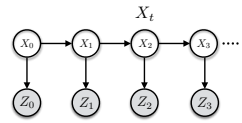
$$\begin{aligned}\alpha_t(i) &= P(Z_0, \ldots, Z_t, X_t = i) \\ &= \sum_{X_{t-1}} P(Z_0, \ldots, Z_t, X_{t-1}, X_t = i) \\ &= \sum_{X_{t-1}} P(Z_t | Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i) P(Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i) \\ &= \sum_{X_{t-1}} P(Z_t | \overset{\text{CI}}{X_t = i}) P(X_t = i | Z_0, \ldots, Z_{t-1}, X_{t-1}) P(Z_0, \ldots, Z_{t-1}, X_{t-1})\end{aligned}$$

Forward probabilities



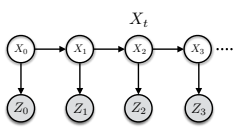
$$
\begin{aligned}
\alpha_t(i) &= P(Z_0, \ldots, Z_t, X_t = i) \\
&= \sum_{X_{t-1}} P(Z_0, \ldots, Z_t, X_{t-1}, X_t = i) \\
&= \sum_{X_{t-1}} P(Z_t | Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i) P(Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i) \\
&= \sum_{X_{t-1}} \underbrace{P(Z_t | X_t = i)}_{\text{CI}} P(X_t = i | Z_0, \ldots, Z_{t-1}, X_{t-1}) P(Z_0, \ldots, Z_{t-1}, X_{t-1}) \\
&= P(Z_t | X_t = i) \sum_{X_{t-1}} \underbrace{P(X_t = i | X_{t-1})}_{\text{Markov assumption}} P(Z_0, \ldots, Z_{t-1}, X_{t-1})
\end{aligned}
$$

Forward probabilities



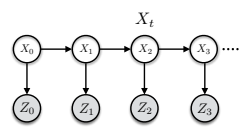
$$
\begin{aligned}
\alpha_t(i) &= P(Z_0, \ldots, Z_t, X_t = i) \\
&= \sum_{X_{t-1}} P(Z_0, \ldots, Z_t, X_{t-1}, X_t = i) \\
&= \sum_{X_{t-1}} P(Z_t | Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i) P(Z_0, \ldots, Z_{t-1}, X_{t-1}, X_t = i) \\
&= \sum_{X_{t-1}} \underbrace{P(Z_t | X_t = i)}_{\text{CI}} P(X_t = i | Z_0, \ldots, Z_{t-1}, X_{t-1}) P(Z_0, \ldots, Z_{t-1}, X_{t-1}) \\
&= P(Z_t | X_t = i) \sum_{X_{t-1}} \underbrace{P(X_t = i | X_{t-1})}_{\text{Markov assumption}} P(Z_0, \ldots, Z_{t-1}, X_{t-1}) \\
&= P(Z_t | X_t = i) \sum_{X_{t-1}} P(X_t = i | X_{t-1}) \alpha_{t-1}(X_{t-1})
\end{aligned}
$$

Forward probabilities



$\alpha_t(X_t)$ follows from $\alpha_{t-1}(X_{t-1})$ ⇨ Compute $\alpha_t(X_t)$ by working <u>forward</u> in time

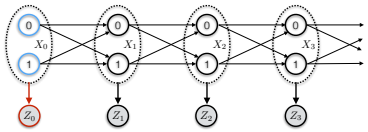
Forward probabilities



$\alpha_t(X_t)$ follows from $\alpha_{t-1}(X_{t-1})$ ⇨ Compute $\alpha_t(X_t)$ by working <u>forward</u> in time

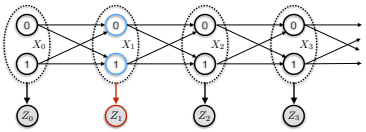
$$
\boxed{
\begin{aligned}
\alpha_0(i) &= P(Z_0, X_0 = i) = P(Z_0 | X_0 = i) P(X_0 = i) \\
\alpha_t(i) &= P(Z_t | X_t = i) \sum_{X_{t-1}} P(X_t = i | X_{t-1}) \alpha_{t-1}(X_{t-1})
\end{aligned}
}
$$

Forward probabilities: Example



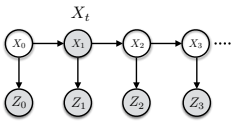
$$
\begin{aligned}
\alpha_0(0) &= P(Z_0, X_0 = 0) = P(Z_0 | X_0 = 0) P(X_0 = 0) \\
\alpha_0(1) &= P(Z_0, X_0 = 1) = P(Z_0 | X_0 = 0) P(X_0 = 1)
\end{aligned}
$$

Forward probabilities: Example



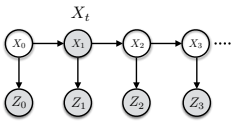
$$
\begin{aligned}
\alpha_1(0) &= P(Z_1 | X_1 = 0) \cdot \Big[P(X_1 = 0 | X_0 = 0) \alpha_0(0) + P(X_1 = 0 | X_0 = 1) \alpha_0(1) \Big] \\
\alpha_1(1) &= P(Z_1 | X_1 = 1) \cdot \Big[P(X_1 = 1 | X_0 = 0) \alpha_0(0) + P(X_1 = 1 | X_0 = 1) \alpha_0(1) \Big]
\end{aligned}
$$

Backward probabilities



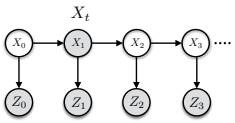
$$\beta_t(i) = P(Z_{t+1}, \dots Z_T | X_t = i)$$

Backward probabilities



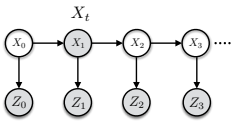
$$\begin{aligned}\beta_t(i) &= P(Z_{t+1}, \dots Z_T | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1}, \dots Z_T, X_{t+1} | X_t = i)\end{aligned}$$

Backward probabilities



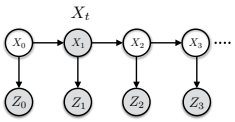
$$\begin{aligned}\beta_t(i) &= P(Z_{t+1}, \dots Z_T | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1}, \dots Z_T, X_{t+1} | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1}, \dots Z_T | X_{t+1}, X_t = i) P(X_{t+1} | X_t = i)\end{aligned}$$

Backward probabilities



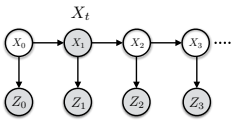
$$\begin{aligned}\beta_t(i) &= P(Z_{t+1}, \dots Z_T | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1}, \dots Z_T, X_{t+1} | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1}, \dots Z_T | X_{t+1}, X_t = i) P(X_{t+1} | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1} \overset{\text{CI}}{|} X_{t+1}) P(Z_{t+2}, \dots Z_T | X_{t+1}) P(X_{t+1} | X_t = i)\end{aligned}$$

Backward probabilities



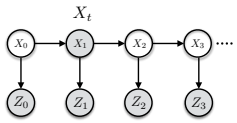
$$\begin{aligned}\beta_t(i) &= P(Z_{t+1}, \dots Z_T | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1}, \dots Z_T, X_{t+1} | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1}, \dots Z_T | X_{t+1}, X_t = i) P(X_{t+1} | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1} \overset{\text{CI}}{|} X_{t+1}) P(Z_{t+2}, \dots Z_T | X_{t+1}) P(X_{t+1} | X_t = i) \\ &= \sum_{X_{t+1}} P(Z_{t+1} | X_{t+1}) \beta_{t+1}(X_{t+1}) P(X_{t+1} | X_t = i)\end{aligned}$$

Backward probabilities



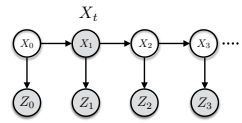
$$\beta_t(i) = \sum_{X_{t+1}} P(Z_{t+1} | X_{t+1}) P(X_{t+1} | X_t = i) \beta_{t+1}(X_{t+1})$$

Backward probabilities



$\beta_t(X_t)$ follows from $\beta_{t+1}(X_{t+1})$ ⇨ Compute $\beta_t(X_t)$ by working <u>backward</u> in time

Backward probabilities

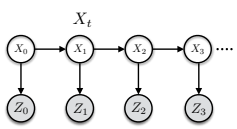


$\beta_t(X_t)$ follows from $\beta_{t+1}(X_{t+1})$ ⇨ Compute $\beta_t(X_t)$ by working <u>backward</u> in time

$$\beta_T(i) = P(Z_{T+1}, Z_{T+2}, \ldots | X_T = i) = 1$$

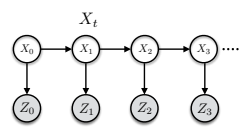
$$\beta_t(i) = \sum_{X_{t+1}} P(Z_{t+1}|X_{t+1}) P(X_{t+1}|X_t = i) \beta_{t+1}(X_{t+1})$$

Observation sequence likelihood



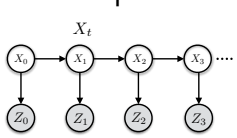
$$P(Z^T) = \sum_{X_t} \alpha(X_t)\beta(X_t)$$

Observation sequence likelihood



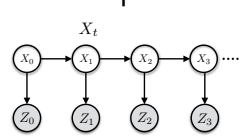
$$P(Z^T) = \sum_{X_t} \alpha(X_t)\beta(X_t) = \sum_{X_T} \alpha(X_T)\beta(X_T)$$

Observation sequence likelihood



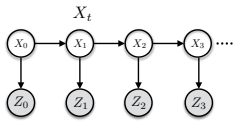
$$P(Z^T) = \sum_{X_t} \alpha(X_t)\beta(X_t) = \sum_{X_T} \alpha(X_T)\overbrace{\beta(X_T)}^{\beta(X_T)=1}$$

Observation sequence likelihood



$$P(Z^T) = \sum_{X_t} \alpha(X_t)\beta(X_t) = \sum_{X_T} \alpha(X_T)\overbrace{\beta(X_T)}^{\beta(X_T)=1} = \sum_{X_T} \alpha(X_T)$$

Observation sequence likelihood



$$P(Z^T) = \sum_{X_t} \alpha(X_t)\beta(X_t) = \sum_{X_T} \alpha(X_T)\overbrace{\beta(X_T)}^{\beta(X_T)=1} = \sum_{X_T} \alpha(X_T)$$

Probability of the observation sequence

$$P(Z^T) = \sum_{X_T} \alpha(X_T)$$

Forward-backward algorithm

$$\alpha_0(i) = P(Z_0, X_0 = i) = P(Z_0|X_0 = i)P(X_0 = i)$$

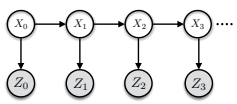
$$\alpha_t(i) = P(Z_t|X_t = i) \sum_{X_{t-1}} P(X_t = i|X_{t-1})\alpha_{t-1}(X_{t-1})$$

$$\beta_T(i) = P(Z_{T+1}, Z_{T+2}, \ldots | X_T = i) = 1$$

$$\beta_t(i) = \sum_{X_{t+1}} P(Z_{t+1}|X_{t+1})P(X_{t+1}|X_t = i)\beta_{t+1}(X_{t+1})$$

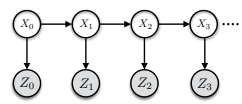
- Time complexity: Linear in t and quadratic in $\mathrm{Card}(\mathcal{X})$
- Space complexity: Constant

MAP state estimates



- **Goal**: Estimate most likely sequence of hidden states for a given observation sequence

MAP state estimates

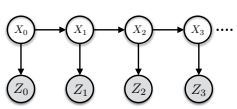


- **Goal**: Estimate most likely sequence of hidden states for a given observation sequence
- **Approach**: Individually determine MAP estimate of each X_t

$$X_t^* = \arg\max_{X_t} P(X_t|Z^T) = \arg\max_{X_t} \left\{ \frac{\alpha(X_t)\beta(X_t)}{P(Z^T)} \right\}$$

$$X^{T^*} = \{X_0^*, X_1^*, \ldots, X_T^*\}$$

MAP state estimates



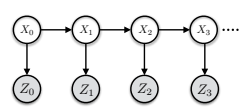
- **Goal**: Estimate most likely sequence of hidden states for a given observation sequence
- A

Problem: Sequence of most likely states may not be most likely sequence

$$\{X_0^*, X_1^*, \ldots, X_T^*\} \neq \{X_0, X_1, \ldots, X_T\}^*$$

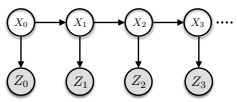
$$X^{T^*} = \{X_0^*, X_1^*, \ldots, X_T^*\}$$

MAP sequence estimate



$$\{X_0, X_1, \ldots, X_T\}^* = \arg\max_{X_1, X_2, \ldots, X_T} P(X_0, X_1, \ldots, X_T|Z^T)$$

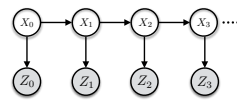
MAP sequence estimate



$$\{X_0, X_1, \ldots, X_T\}^* = \underset{X_1, X_2, \ldots, X_T}{\arg\max}\ P(X_0, X_1, \ldots, X_T | Z^T)$$

This is a large space to optimize over (e.g., 2^T)

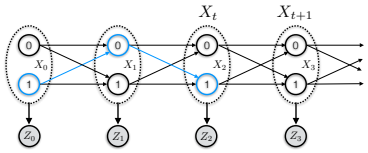
MAP sequence estimate



$$\{X_0, X_1, \ldots, X_T\}^* = \underset{X_1, X_2, \ldots, X_T}{\arg\max}\ P(X_0, X_1, \ldots, X_T | Z^T)$$

This is a large space to optimize over (e.g., 2^T) ⇨ Exploit conditional independence

MAP sequence estimate

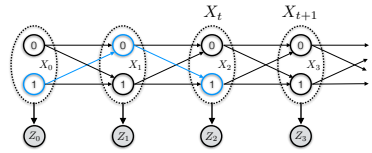


- **Key Idea**: The most likely sequence up to $X_{t+1} = i$ is formed from the most likely sequence to $X_t = j$, where

$$P(X_{t+1} = i | X_t = j) P(X_t = j)^* \geq P(X_{t+1} = i | X_t = k) P(X_t = k)^*$$

Likelihood for MAP sequence to $X_t = k$

MAP sequence estimate



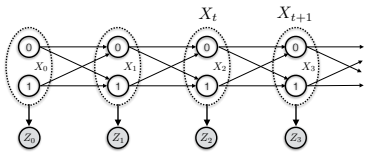
- **Key Idea**: The most likely sequence up to $X_{t+1} = i$ is formed from the most likely sequence to $X_t = j$, where

$$P(X_{t+1} = i | X_t = j) P(X_t = j)^* \geq P(X_{t+1} = i | X_t = k) P(X_t = k)^*$$

Likelihood for MAP sequence to $X_t = k$

⇨ Use dynamic programming to estimate MAP sequence incrementally

Viterbi algorithm



$$\underset{X_0, \ldots, X_T}{\arg\max}\ P(X_0, \ldots, X_T | Z^T) = \underset{X_0, \ldots, X_T}{\arg\max}\ P(X_0, \ldots, X_T, Z^T)$$

Viterbi algorithm

$$X^{T^*} = \underset{X_0, \ldots, X_T}{\arg\max}\ P(Z_0, \ldots, Z_T, X_0, \ldots, X_T)$$

Viterbi algorithm

$$
\begin{aligned}
X^{T^*} &= \arg\max_{X_0,\ldots,X_T} P(Z_0,\ldots,Z_T,X_0,\ldots,X_T) \\
&= \arg\max_{X_0,\ldots,X_T} \Big\{ P(Z_T|X_T)P(X_T|X_{T-1})P(Z_{T-1}|X_{T-1})P(X_{T-1}|X_{T-2})\cdots \\
&\qquad \cdots P(Z_t|X_t)P(X_t|X_{t-1})\cdots P(Z_0|X_0)P(X_0)\Big\}
\end{aligned}
$$

Viterbi algorithm

$$
\begin{aligned}
X^{T^*} &= \arg\max_{X_0,\ldots,X_T} P(Z_0,\ldots,Z_T,X_0,\ldots,X_T) \\
&= \arg\max_{X_0,\ldots,X_T} \Big\{ P(Z_T|X_T)P(X_T|X_{T-1})P(Z_{T-1}|X_{T-1})P(X_{T-1}|X_{T-2})\cdots \\
&\qquad \cdots P(Z_t|X_t)P(X_t|X_{t-1})\cdots P(Z_0|X_0)P(X_0)\Big\} \\
&= \arg\max_{X_T} \Big\{ P(Z_T|X_T)\cdot \max_{X_{T-1}}\Big(P(X_T|X_{T-1})P(Z_{T-1}|X_{T-1})\cdot \\
&\qquad \max_{X_{T-2}}\big(P(X_{T-1}|X_{T-2})P(Z_{T-2}|X_{T-2})\cdots\big)\Big)\Big\}
\end{aligned}
$$

Viterbi algorithm

$$
\delta_t(i) = \max_{X_0,\ldots,X_{t-1}} P(Z_0,\ldots,Z_t,X_0,\ldots,X_{t-1},X_t=i)
$$

Viterbi algorithm

$$
\begin{aligned}
\delta_t(i) &= \max_{X_0,\ldots,X_{t-1}} P(Z_0,\ldots,Z_t,X_0,\ldots,X_{t-1},X_t=i) \\
&= \max_{X_0,\ldots,X_{t-1}} \Big\{ P(Z_t|X_t=i)P(X_t=i|X_{t-1})P(Z_0,\ldots,Z_{t-1},X_0,\ldots X_{t-1})\Big\}
\end{aligned}
$$

Viterbi algorithm

$$
\begin{aligned}
\delta_t(i) &= \max_{X_0,\ldots,X_{t-1}} P(Z_0,\ldots,Z_t,X_0,\ldots,X_{t-1},X_t=i) \\
&= \max_{X_0,\ldots,X_{t-1}} \Big\{ P(Z_t|X_t=i)P(X_t=i|X_{t-1})P(Z_0,\ldots,Z_{t-1},X_0,\ldots X_{t-1})\Big\} \\
&= P(Z_t|X_t=i)\max_j\Big\{ P(X_t=i|X_{t-1}=j)P(Z_0,\ldots,Z_{t-1},X_0,\ldots X_{t-1}=j)\Big\}
\end{aligned}
$$

Viterbi algorithm

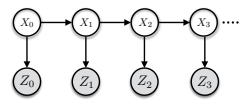
$$
\begin{aligned}
\delta_t(i) &= \max_{X_0,\ldots,X_{t-1}} P(Z_0,\ldots,Z_t,X_0,\ldots,X_{t-1},X_t=i) \\
&= \max_{X_0,\ldots,X_{t-1}} \Big\{ P(Z_t|X_t=i)P(X_t=i|X_{t-1})P(Z_0,\ldots,Z_{t-1},X_0,\ldots X_{t-1})\Big\} \\
&= P(Z_t|X_t=i)\max_j\Big\{ P(X_t=i|X_{t-1}=j)P(Z_0,\ldots,Z_{t-1},X_0,\ldots X_{t-1}=j)\Big\} \\
&= P(Z_t|X_t=i)\max_j\Big\{ P(X_t=i|X_{t-1}=j)\delta_{t-1}(j)\Big\}
\end{aligned}
$$

Viterbi algorithm

$$
\begin{aligned}
\delta_t(i) &= \max_{X_0,\ldots,X_{t-1}} P(Z_0,\ldots,Z_t,X_0,\ldots,X_{t-1},X_t=i) \\
&= \max_{X_0,\ldots,X_{t-1}} \Big\{ P(Z_t|X_t=i)P(X_t=i|X_{t-1})P(Z_0,\ldots,Z_{t-1},X_0,\ldots X_{t-1}) \Big\} \\
&= P(Z_t|X_t=i) \max_j \Big\{ P(X_t=i|X_{t-1}=j)P(Z_0,\ldots,Z_{t-1},X_0,\ldots X_{t-1}=j) \Big\} \\
&= P(Z_t|X_t=i) \max_j \Big\{ P(X_t=i|X_{t-1}=j)\delta_{t-1}(j) \Big\}
\end{aligned}
$$

Given the most likely sequence to $X_{t-1}=j$,
we can compute the most likely sequence to $X_t=i$

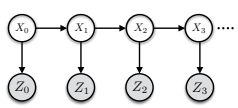
Viterbi algorithm



$\delta_t(i)$ follows from $\delta_{t-1}(j)$ ⇨ Compute $\delta_t(i)$ by working <u>forward</u> in time

$$
\boxed{
\begin{aligned}
\delta_0(i) &= P(X_0=i,Z_0) = P(Z_0|X_0=i)P(X_0=i) \\
\delta_t(i) &= P(Z_t|X_t=i) \cdot \max_j \{P(X_t=i|X_{t-1}=j)\delta_{t-1}(j)\}
\end{aligned}
}
$$

Viterbi algorithm



- Initialization: $\forall\, s \in \mathcal{X}$
  - $\delta_0(i) = P(Z_0|X_0=i)P(X_0=i)$
  - $\mathrm{Pre}_0(i) = \texttt{null}$
- Repeat: for $t = 1,\ldots,T$
  - $\delta_t(i) = P(Z_t|X_t=i) \cdot \max_j \big\{ P(X_t=i|X_{t-1}=j)\,\delta_{t-1}(j) \big\}$
  - $\mathrm{Pre}_t(i) = \arg\max_j \big\{ P(X_t=i|X_{t-1}=j)\,\delta_{t-1}(j) \big\}$
- Select most likely terminal state: $X_T^* = \arg\max_i \delta_T(i)$
- Backtrack: for $t = T-1,\ldots,0$
  - $X_t^* = \mathrm{Pre}_{t+1}(X_{t+1}^*)$

Viterbi algorithm

- Time complexity: Linear in t and quadratic in Card($\mathcal{X}$)
- Space complexity: Linear in t