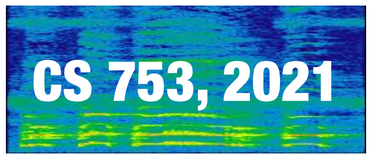
HMMs for Acoustic Modeling

Lecture 1a



Instructor: Preethi Jyothi, IITB

Recall: Statistical ASR

Let $\mathbf{O}$ be a sequence of acoustic features corresponding to a speech signal. That is, $\mathbf{O} = \{O_1, \ldots, O_T\}$, where $O_i \in \mathbb{R}^d$ refers to a d-dimensional acoustic feature vector and T is the length of the sequence.

Let $\mathbf{W}$ denote a word sequence. An ASR decoder solves the following problem:

$$
\begin{aligned}
\mathbf{W}^* &= \arg\max_W \Pr(\mathbf{W} \mid \mathbf{O}) \\
&= \arg\max_W \Pr(\mathbf{O} \mid \mathbf{W}) \Pr(\mathbf{W}) \\
&\approx \arg\max_W \sum_Q \Pr(\mathbf{O} \mid \mathbf{Q}) \Pr(\mathbf{Q} \mid \mathbf{W}) \Pr(\mathbf{W})
\end{aligned}
$$

Recall: Statistical ASR

Let $\mathbf{O}$ be a sequence of acoustic features corresponding to a speech signal. That is, $\mathbf{O} = \{O_1, \ldots, O_T\}$, where $O_i \in \mathbb{R}^d$ refers to a d-dimensional acoustic feature vector and T is the length of the sequence.

Let $\mathbf{W}$ denote a word sequence. An ASR decoder solves the following problem:

$$
\begin{aligned}
\mathbf{W}^* &= \arg\max_W \Pr(\mathbf{W} \mid \mathbf{O}) \\
&= \arg\max_W \Pr(\mathbf{O} \mid \mathbf{W}) \Pr(\mathbf{W}) \\
&\approx \arg\max_W \sum_Q \Pr(\mathbf{O} \mid \mathbf{Q}) \Pr(\mathbf{Q} \mid \mathbf{W}) \Pr(\mathbf{W})
\end{aligned}
$$

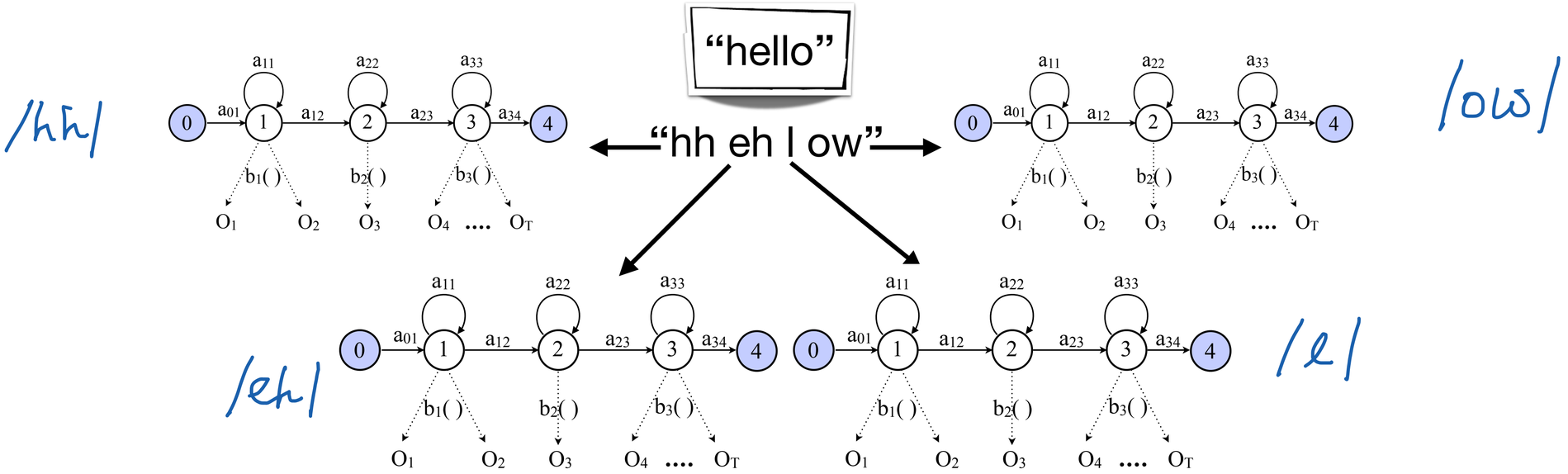
Acoustic Model

Generative Model of Speech

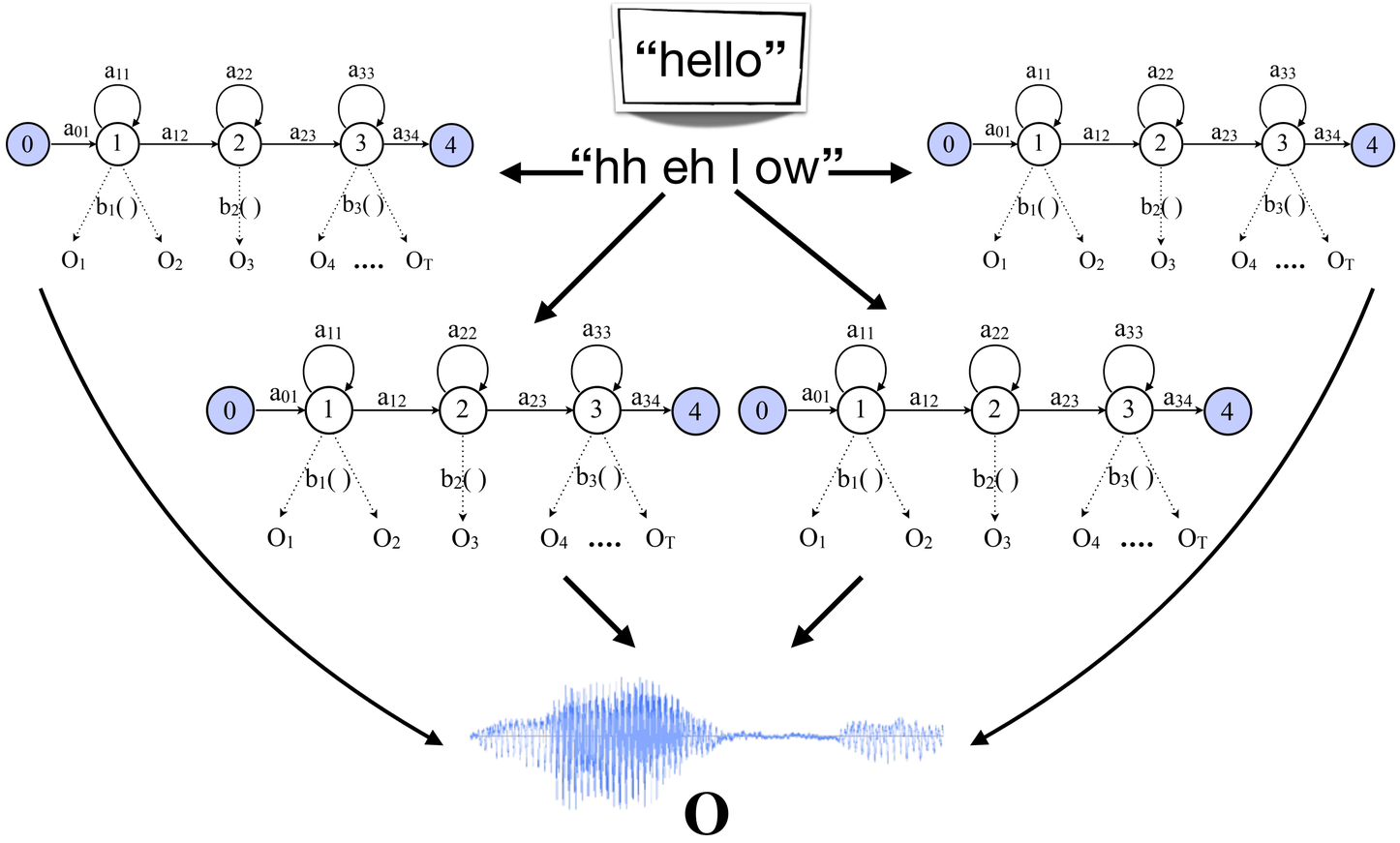
Generative Model of Speech

“hh eh l ow”

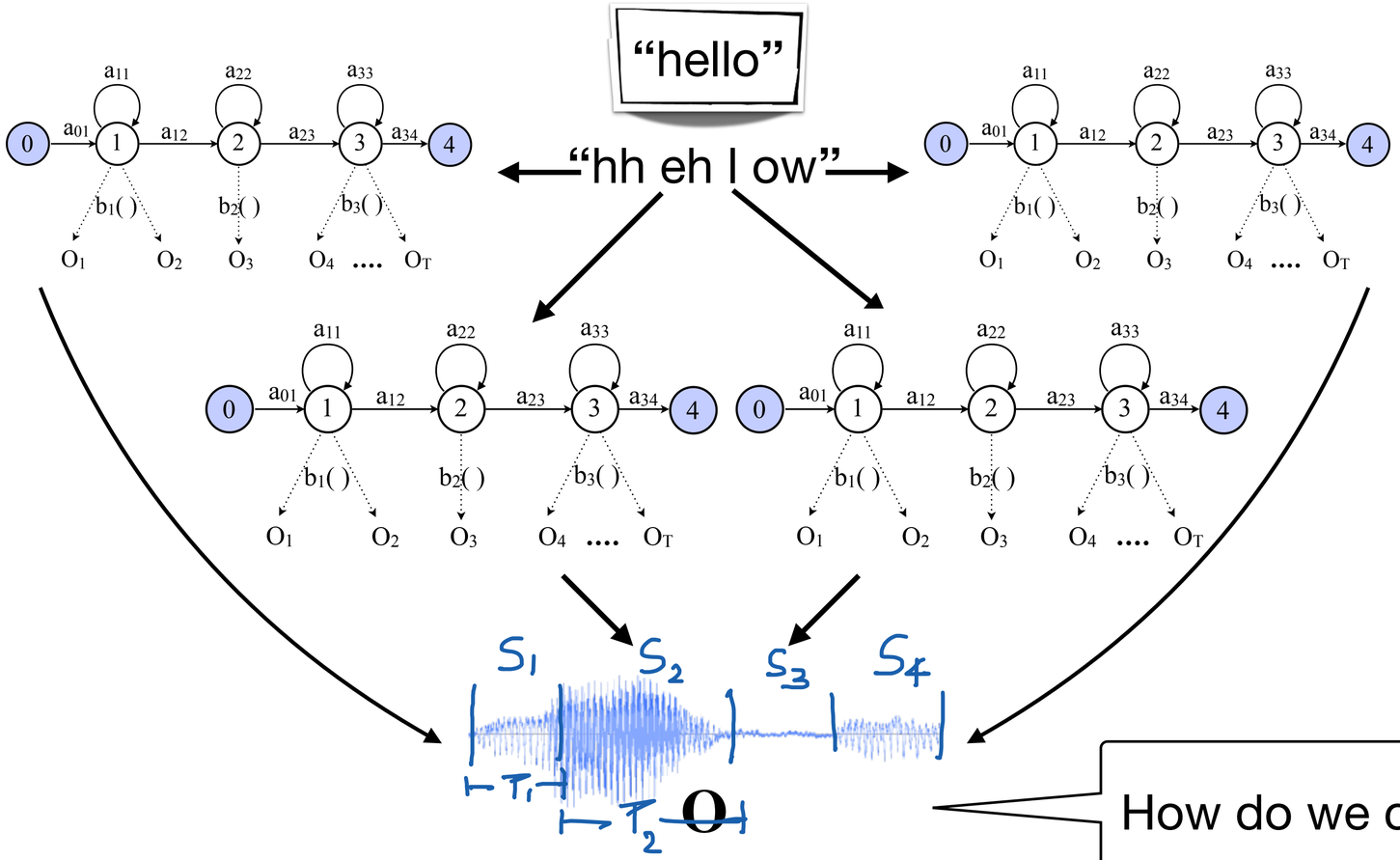
Generative Model of Speech



Generative Model of Speech



Generative Model of Speech



How do we compute $P(\mathbf{O} \mid \text{hello})$?

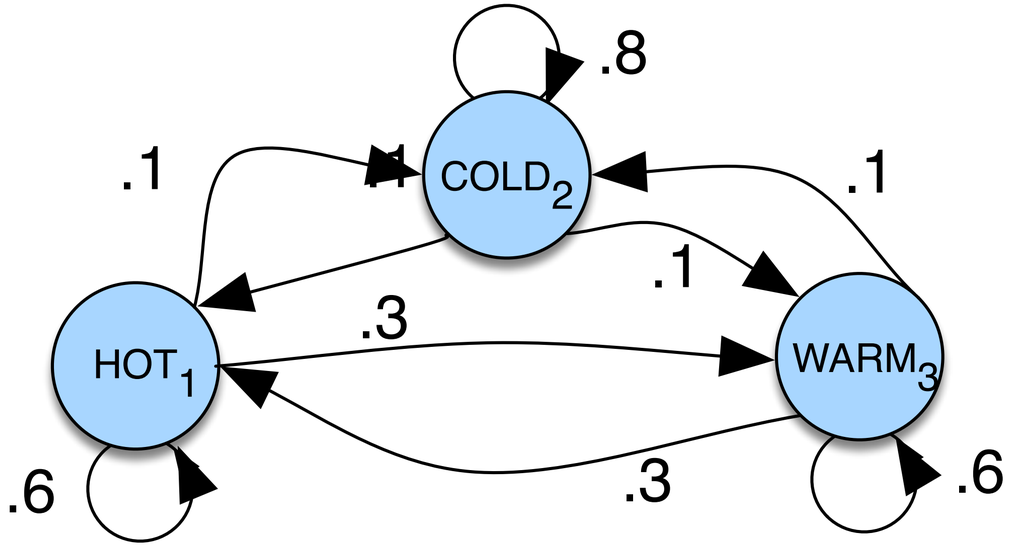
$P(\ \mid \text{hello}) = P(S_1 \mid hh)\ P(S_2 \mid eh)\ P(S_3 \mid l)\ P(S_4 \mid ow$

$P(O_1, \ldots O_{T_1} \mid hh)$ $P(O_{T_1+1}, \ldots O_{T_1+T_2} \mid eh)$ ...

What are Hidden Markov Models (HMMs)?

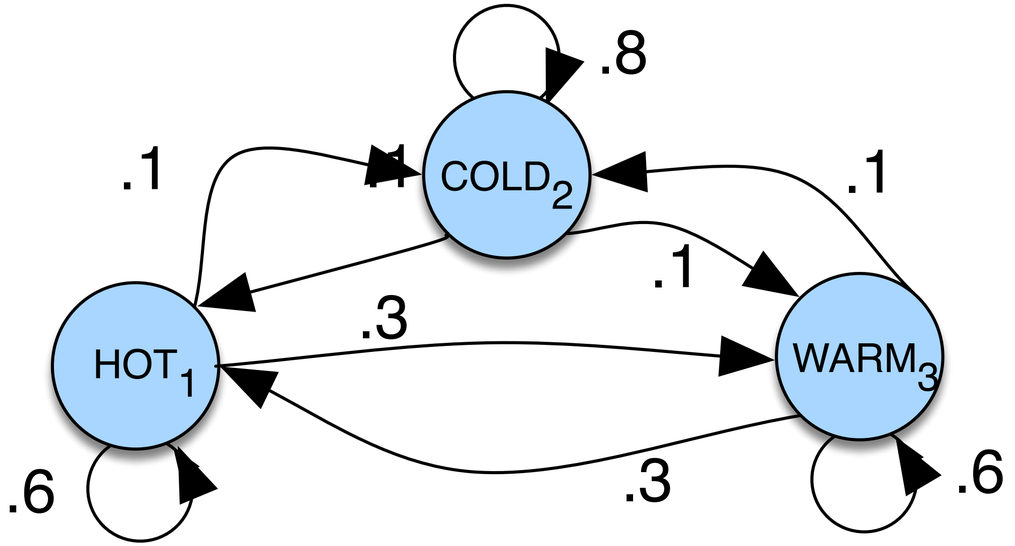
Following slides contain figures/material from “Hidden Markov Models”, “Speech and Language Processing”, D. Jurafsky and J. H. Martin, 2019. (https://web.stanford.edu/~jurafsky/slp3/A.pdf)

Markov Chains



$$\pi = [0.1, 0.7, 0.2]$$

Markov Chains



$\pi = [0.1, 0.7, 0.2]$

P(HHH)

P(WHH)

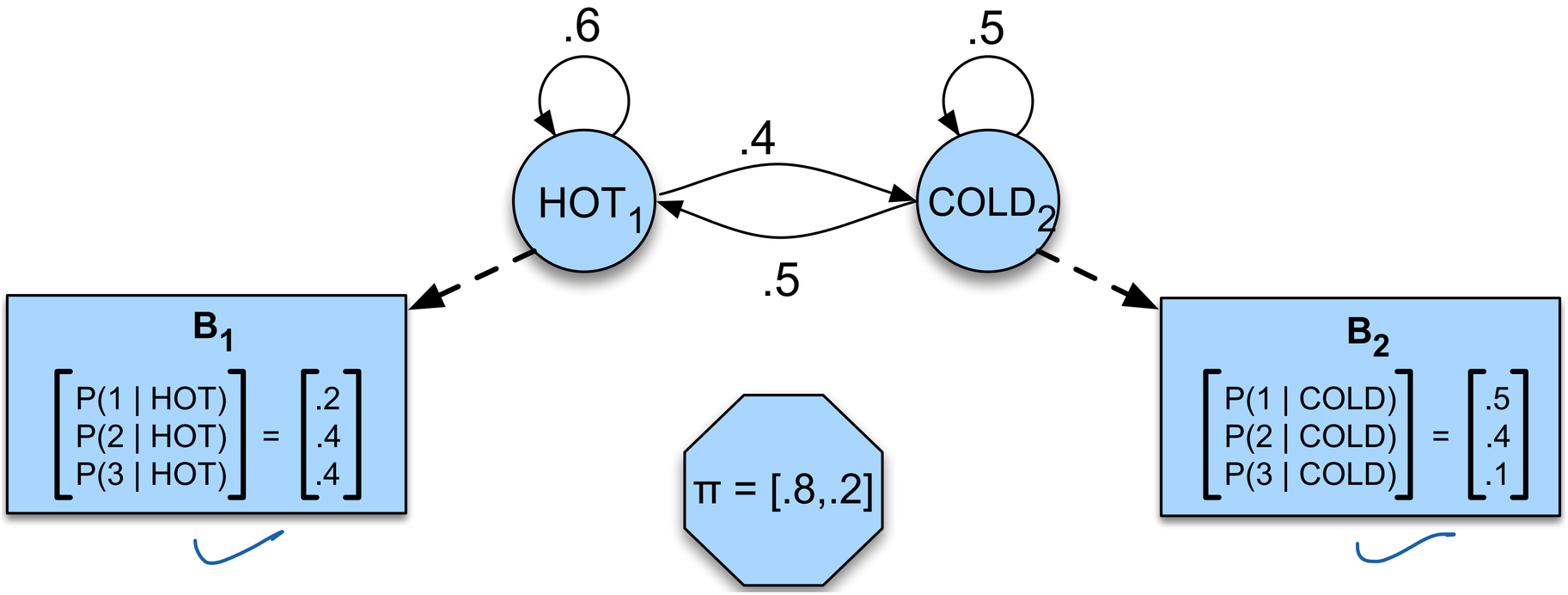
P(CHW)

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

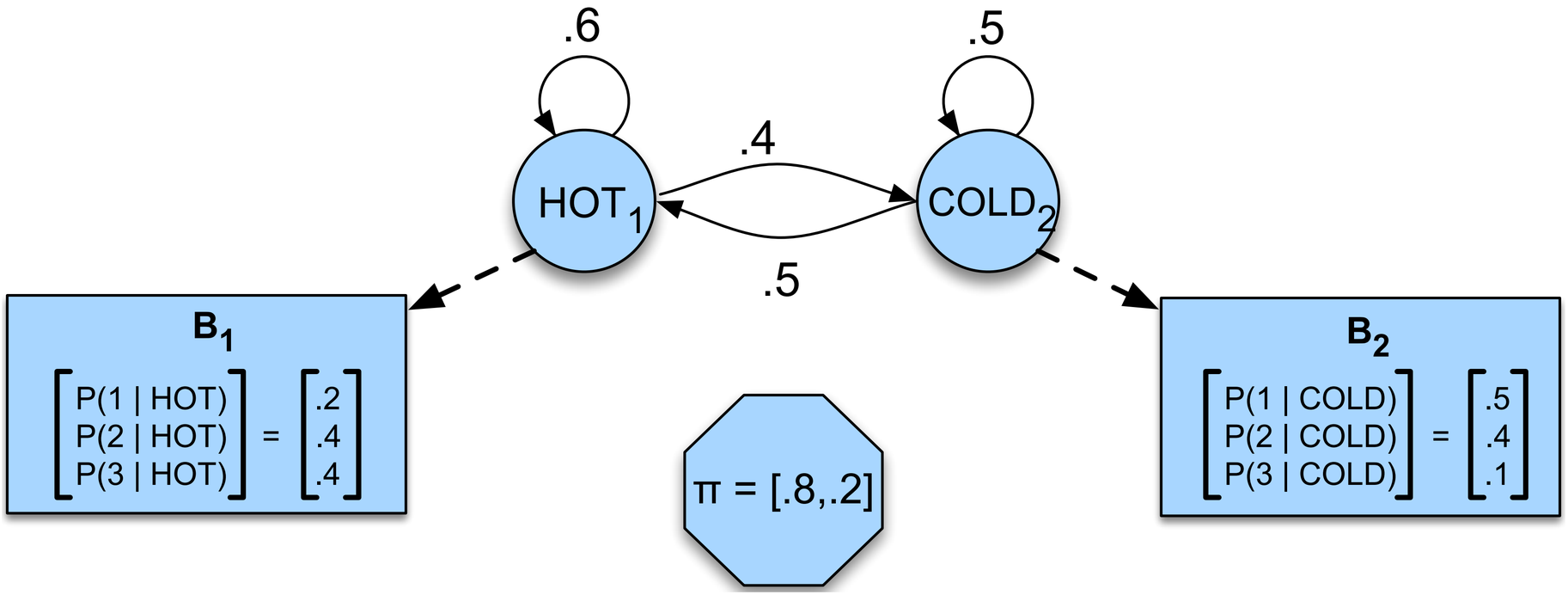
Hidden Markov Model

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, \ldots, v_V$
$B = b_i(o_t)$ $P(o_t \mid q_t = i)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

HMM Assumptions

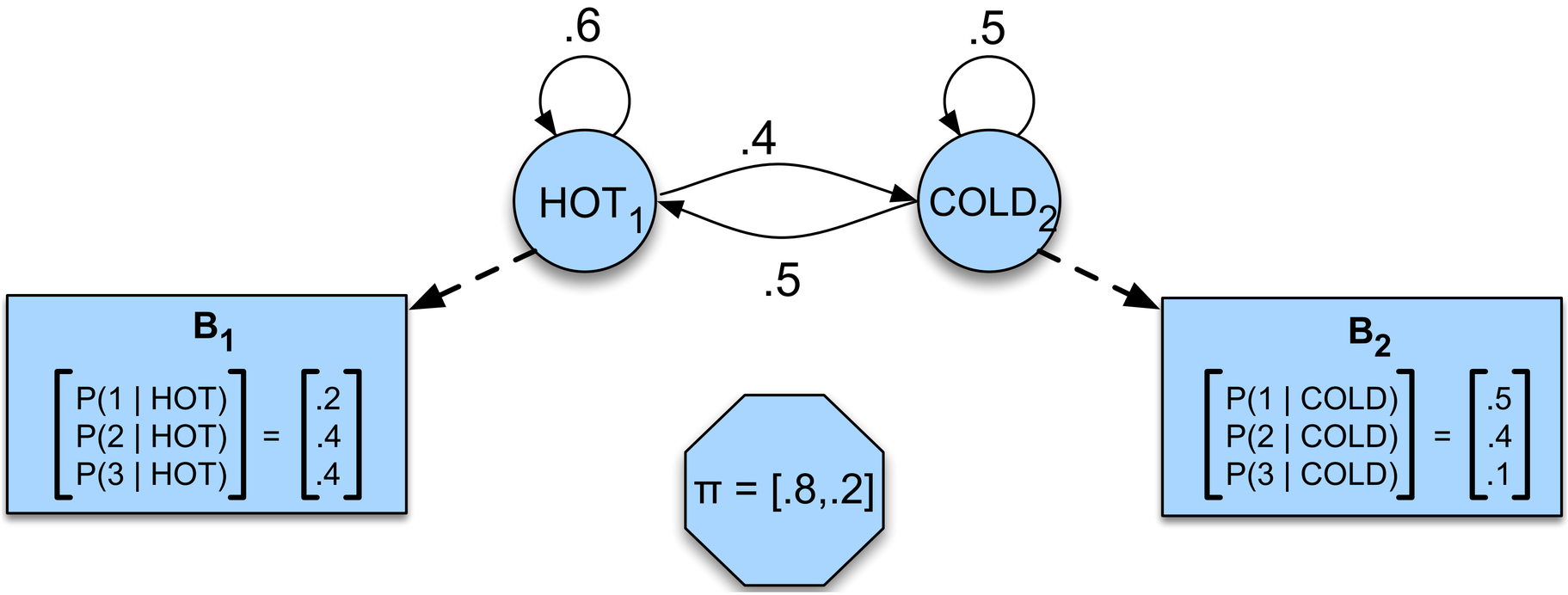


HMM Assumptions



Markov Assumption: $P(q_i|q_1 \ldots q_{i-1}) = P(q_i|q_{i-1})$

HMM Assumptions



Markov Assumption: $P(q_i|q_1 \ldots q_{i-1}) = P(q_i|q_{i-1})$

Output Independence: $P(o_i|q_1 \ldots q_i, \ldots, q_T, o_1, \ldots, o_{i-1}, o_{i+1}, \ldots, o_T) = P(o_i|q_i)$

Three problems for HMMs

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

"A tutorial on hidden Markov models and selected applications in speech recognition", Rabiner, 1989

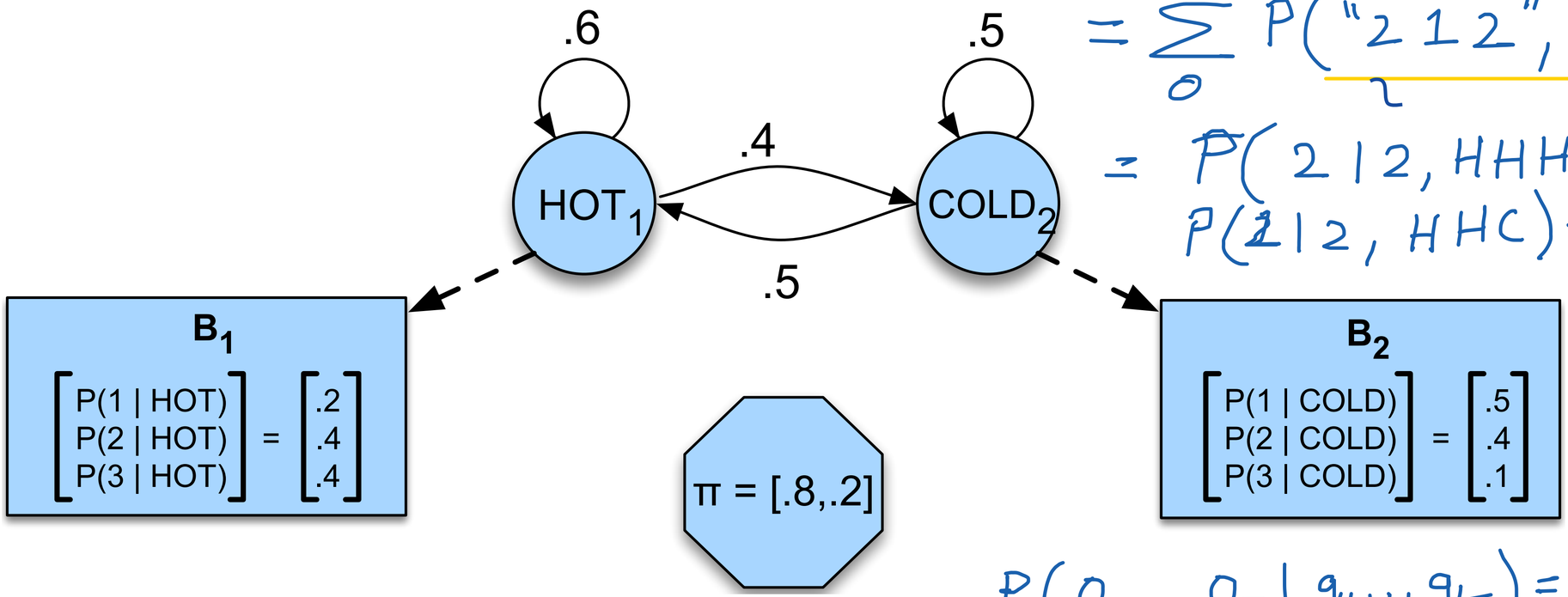
Three problems for HMMs

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

Computing Likelihood: Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O|\lambda)$.

"A tutorial on hidden Markov models and selected applications in speech recognition", Rabiner, 1989

What's the probability of O="2 1 2"?



$P(2\,1\,2) = \sum_{Q} P(\text{"2 1 2"}, Q)$

$= P(2\,1\,2, HHH) + P(2\,1\,2, HHC) + \ldots P(2\,1\,2, CCC)$

$P(O_1 \ldots O_T \mid q_1 \ldots q_T) = \prod_{t=1}^{T} P(O_t \mid q_t)$

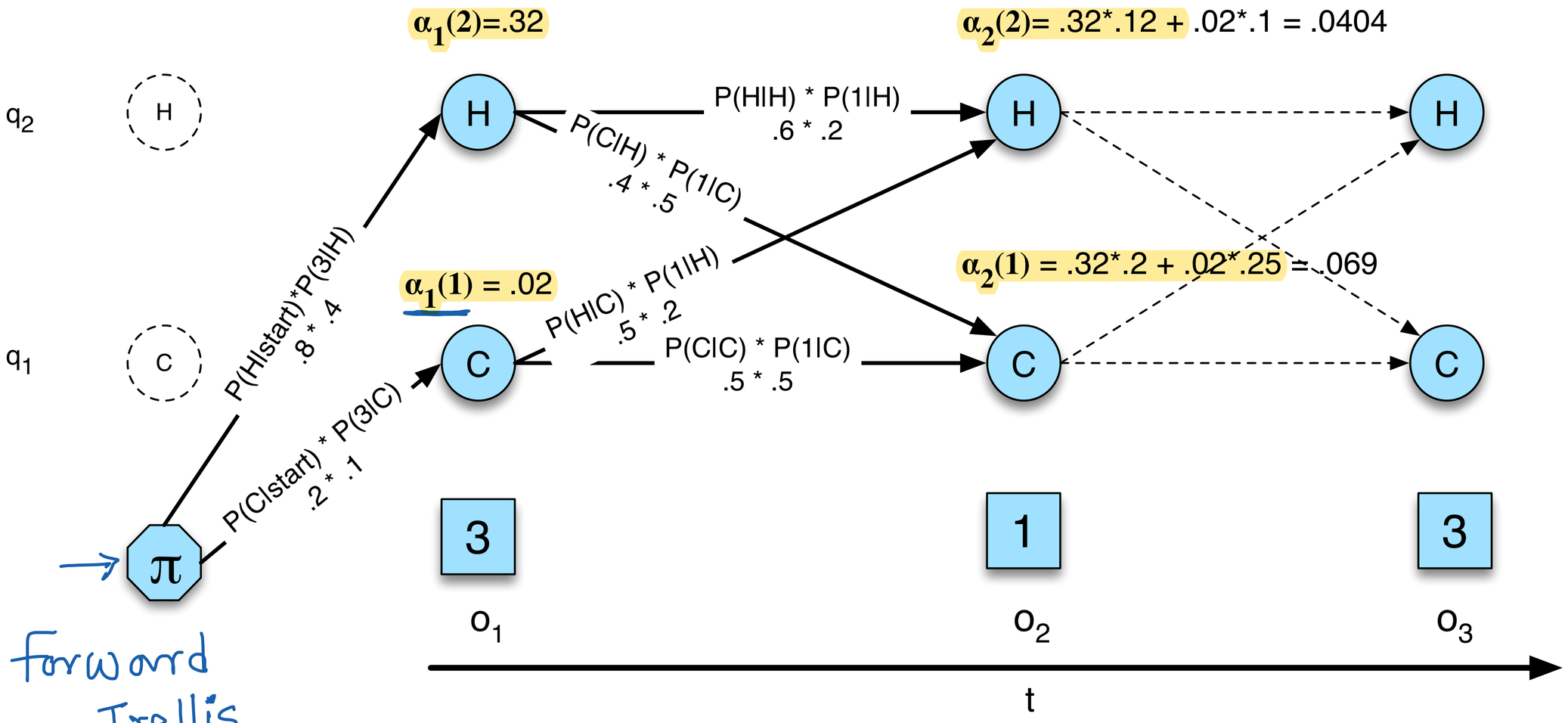
emission probs

$P(\text{"2 1 2"} \mid \lambda)$

$P(\text{"2 1 2"} \mid \text{"H C H"})$

$= P(2 \mid H)\, P(1 \mid C)\, P(2 \mid H)$

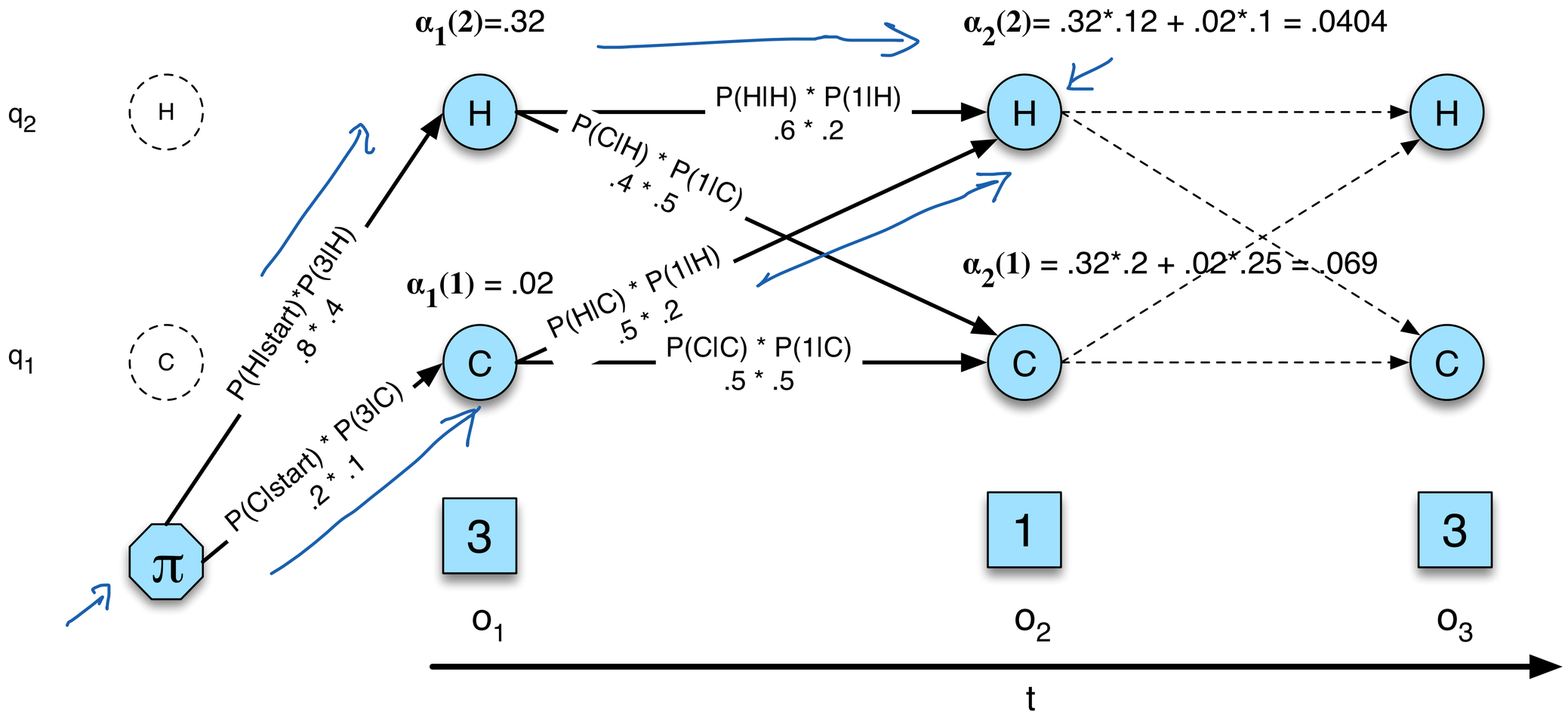
Forward Algorithm



Forward Algorithm

Forward Probability

$$\alpha_t(j) = P(o_1, o_2 \ldots o_t, q_t = j|\lambda)$$

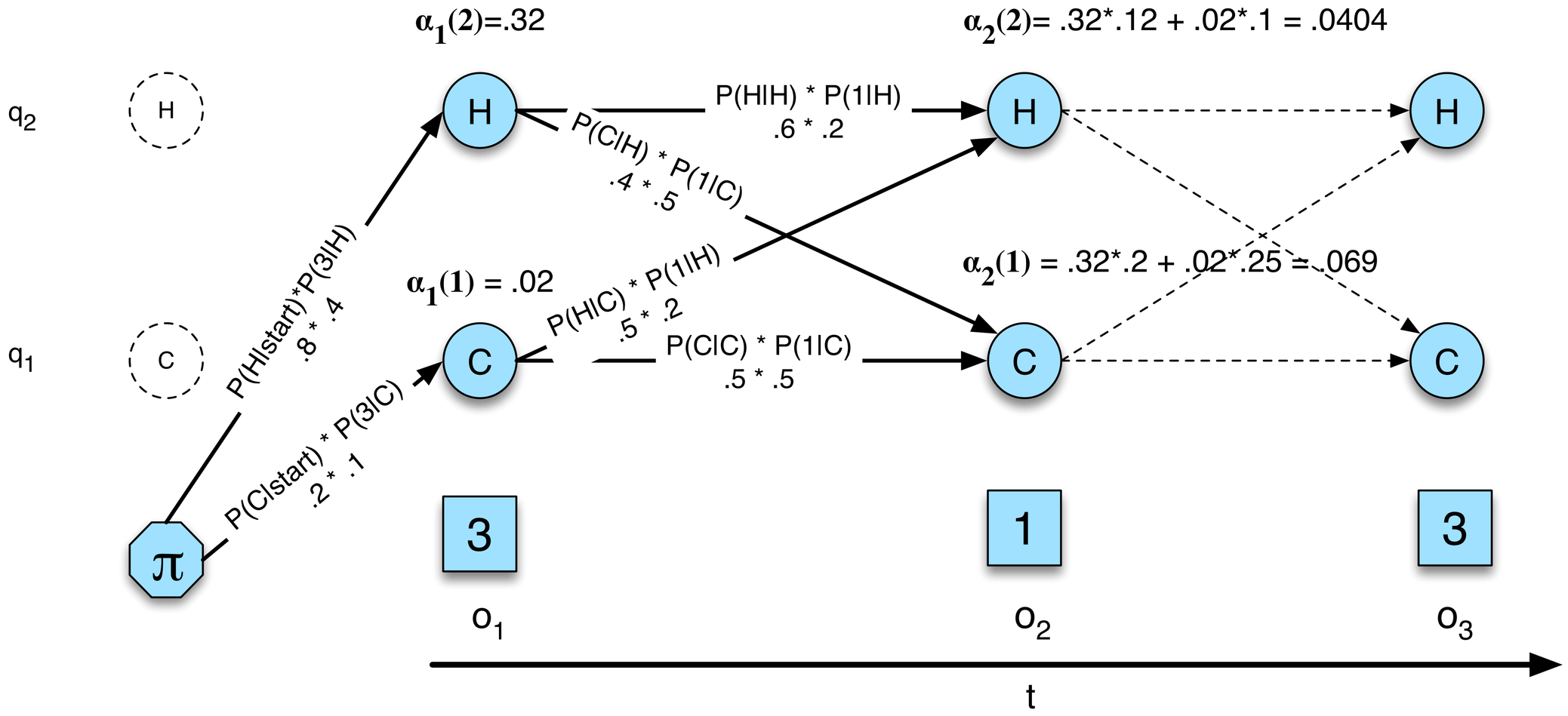


Forward Algorithm

Forward Probability

$$\alpha_t(j) = P(o_1, o_2 \ldots o_t, q_t = j|\lambda)$$

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t)$$



Forward Algorithm

Forward Probability

$$\alpha_t(j) = P(o_1, o_2 \ldots o_t, q_t = j|\lambda)$$

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t)$$

$$\alpha_t(j) = P(o_1, o_2, \ldots o_t, q_t = j \mid \lambda)$$

$$= \sum_i P(o_1, o_2, \ldots, o_t, q_t = j, q_{t-1} = i \mid \lambda)$$

$$= \sum_i P(o_t \mid o_{1:t-1}, q_t = j, q_{t-1} = i) \cdot P(o_{1:t-1}, q_t = j, q_{t-1} = i)$$

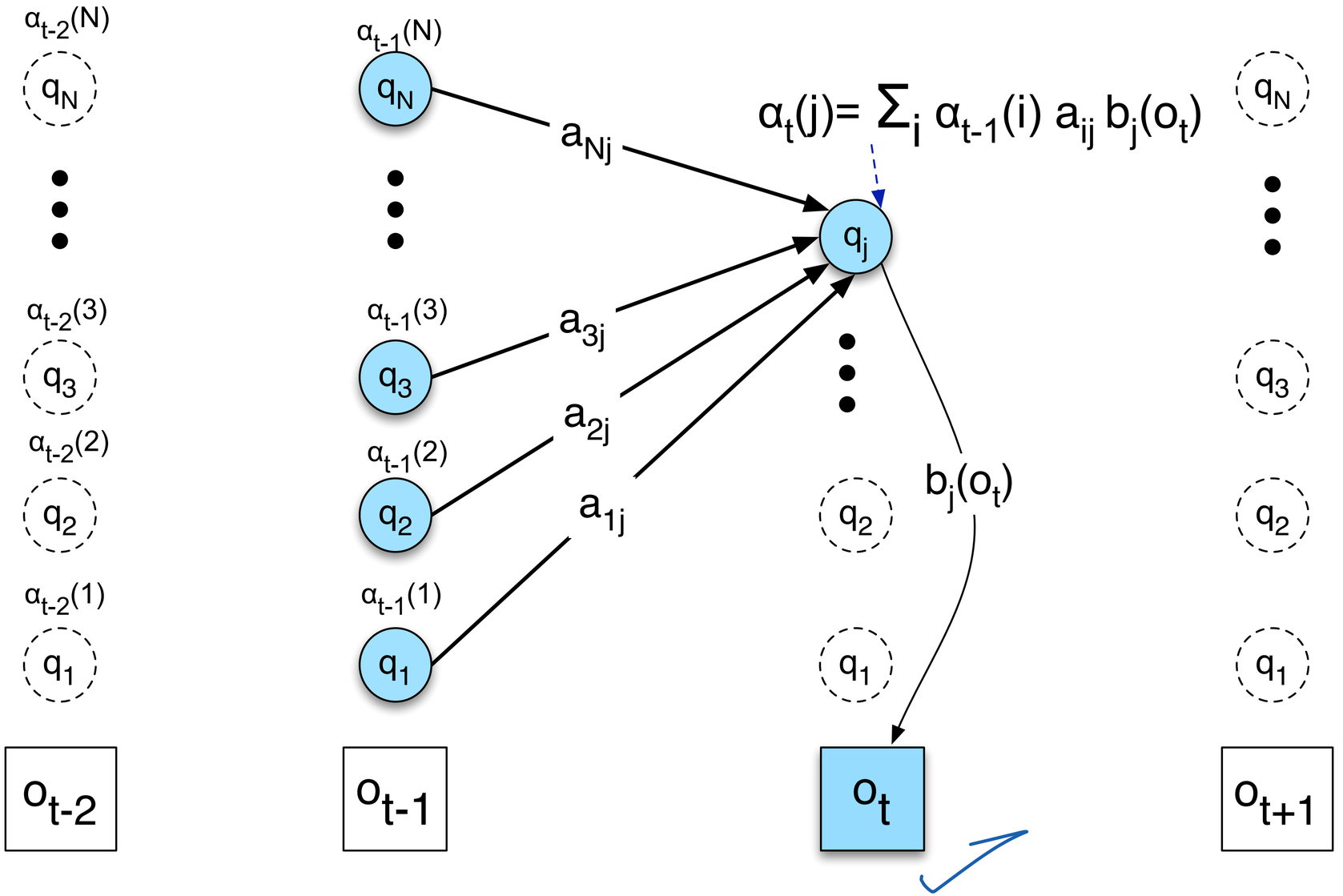
$$= \sum_i P(o_t \mid q_t = j) P(q_t = j \mid q_{t-1} = i, o_{1:t-1}) P(q_{t-1} = i, o_{1:t-1})$$

$$= \sum_i P(o_t \mid q_t = j) P(q_t = j \mid q_{t-1} = i) \alpha_{t-1}(i)$$

$P(o_t \mid q_t = j) \rightarrow b_j(o_t)$

$P(q_t = j \mid q_{t-1} = i) \rightarrow a_{ij}$

Visualizing the forward recursion



Forward Algorithm

1. Initialization:

$$\alpha_1(j) = \pi_j b_j(o_1) \quad 1 \leq j \leq N$$

2. Recursion:

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t); \quad 1 \leq j \leq N, 1 < t \leq T$$

3. Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Three problems for HMMs

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

Three problems for HMMs

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

Decoding: Given as input an HMM $\lambda = (A, B)$ and a sequence of observations $O = o_1, o_2, ..., o_T$, find the most probable sequence of states $Q = q_1 q_2 q_3 \ldots q_T$.

Viterbi Trellis

"Viterbi Algorithm"

$$v_t(j) = \max_{q_1,\ldots,q_{t-1}} P(q_1 \ldots q_{t-1}, o_1, o_2 \ldots o_t, q_t = j | \lambda)$$

Viterbi Path Probability

Viterbi Trellis

$$v_t(j) = \max_{q_1,\ldots,q_{t-1}} P(q_1 \ldots q_{t-1}, o_1, o_2 \ldots o_t, q_t = j|\lambda)$$

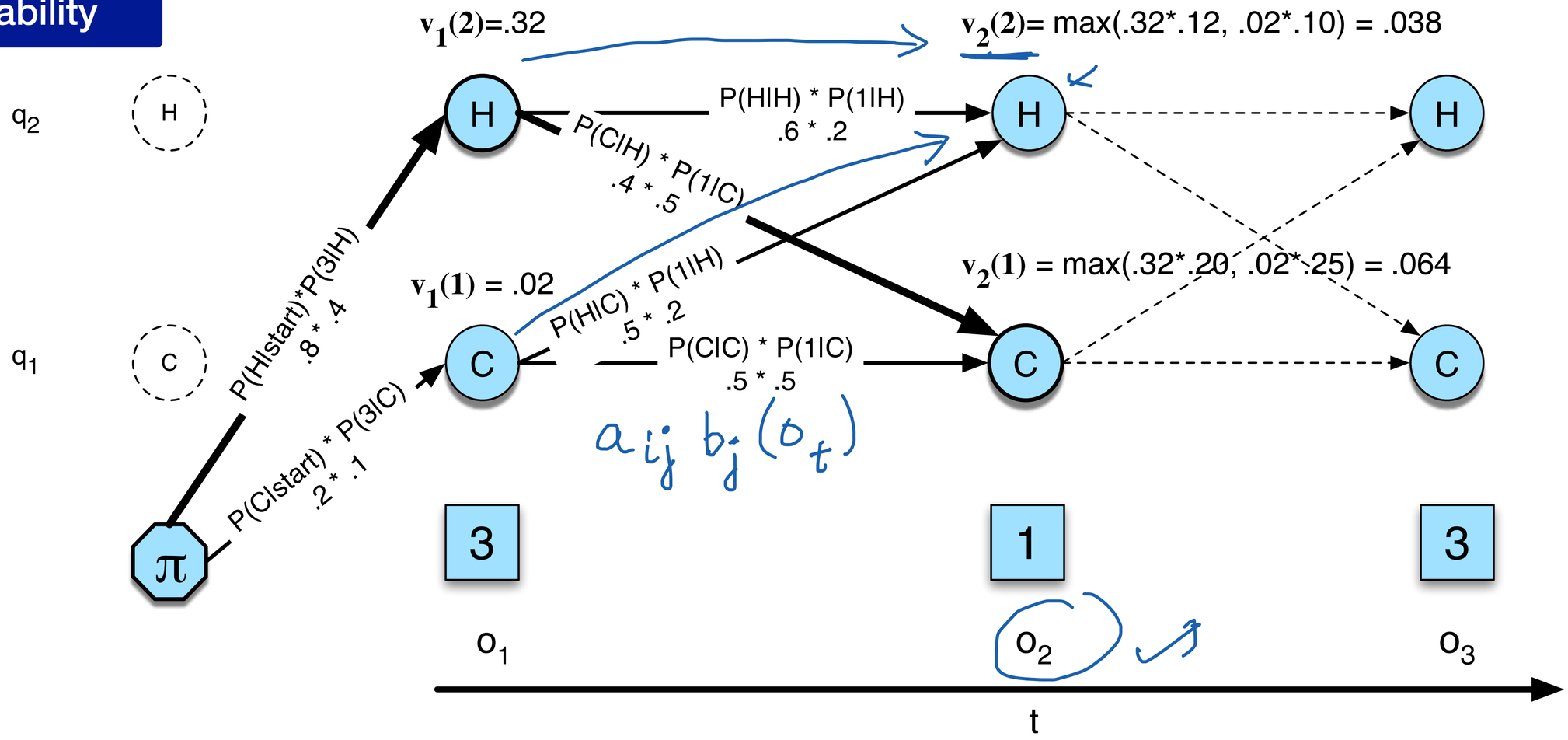
Viterbi Path Probability

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t)$$

Viterbi Trellis

$$v_t(j) = \max_{q_1,\ldots,q_{t-1}} P(q_1 \ldots q_{t-1}, o_1, o_2 \ldots o_t, q_t = j|\lambda) \qquad v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t)$$

Viterbi Path Probability



Viterbi recursion

1. **Initialization:**

$$v_1(j) = \pi_j b_j(o_1) \qquad 1 \le j \le N$$
$$bt_1(j) = 0 \qquad 1 \le j \le N$$

2. **Recursion**

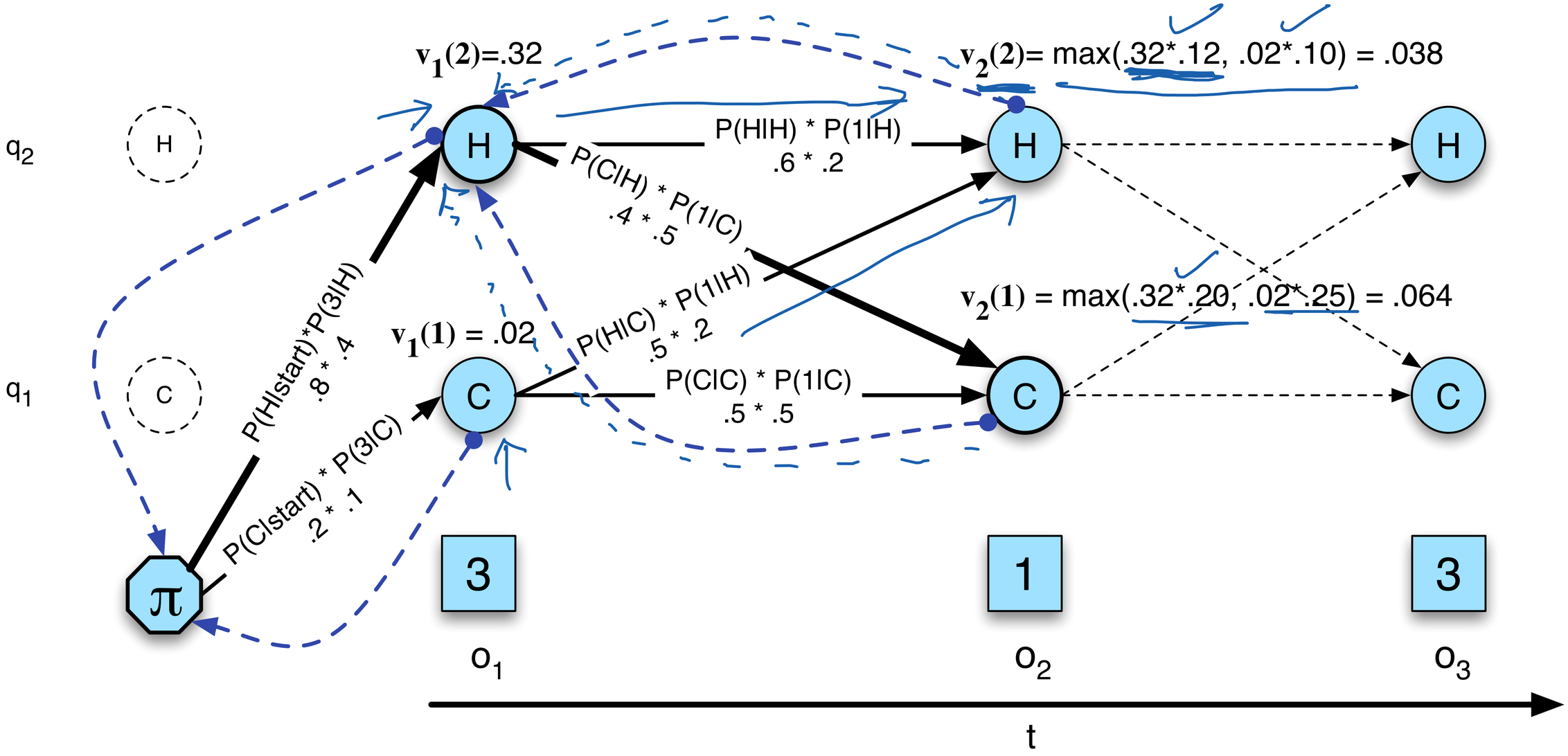
$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t); \quad 1 \le j \le N, 1 < t \le T$$
$$bt_t(j) = \operatorname*{argmax}_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t); \quad 1 \le j \le N, 1 < t \le T$$

3. **Termination:**

The best score: $P* = \max_{i=1}^{N} v_T(i)$

The start of backtrace: $q_T* = \operatorname*{argmax}_{i=1}^{N} v_T(i)$

Viterbi backtrace



Next Module: Learning in HMMs

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

Learning: Given an observation sequence O and the set of possible states in the HMM, learn the HMM parameters A and B.

Next Module: Learning in HMMs

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

Learning: Given an observation sequence O and the set of possible states in the HMM, learn the HMM parameters A and B.

HMM training: **Forward-backward** or **Baum-Welch** algorithm