Hidden Markov Model (HMM) Review

Hidden Markov Model (HMM) Review

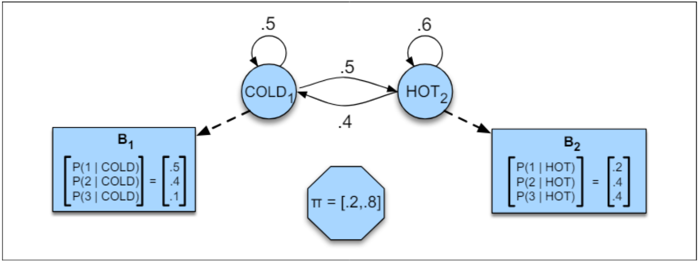
(Example from Prof. Jurafsky's book)

Hidden Markov Model (HMM) Review

Defining HMMs

Defining HMMs - Ice Cream Example

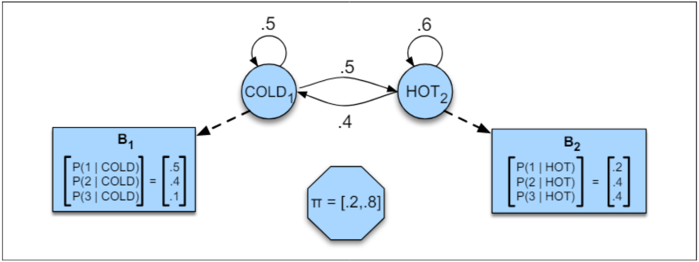
Famous problem by Jason Eisner (2002) where you want to predict if a day was COLD or HOT (your hidden states) based on records of the # of ice creams (your known evidence) Eisner ate that day.



- $S = \{s_1 ... s_N\}$, N states (2 states: cold or hot)
- $A = a_{11} ... a_{ij} ... a_{NN}$, transition probabilities (e.g. cold $\rightarrow$ hot?)
- $B = b_i(o_t)$, emission probabilities (e.g. 3 ice creams $\rightarrow$ hot?)
- $\pi = \{\pi_1 ... \pi_N\}$, initial probabilities (e.g. start $\rightarrow$ hot?)

Defining HMMs - Ice Cream Example

Famous problem by Jason Eisner (2002) where you want to predict if a day was COLD or HOT (your hidden states) based on records of the # of ice creams (your known evidence) Eisner ate that day.



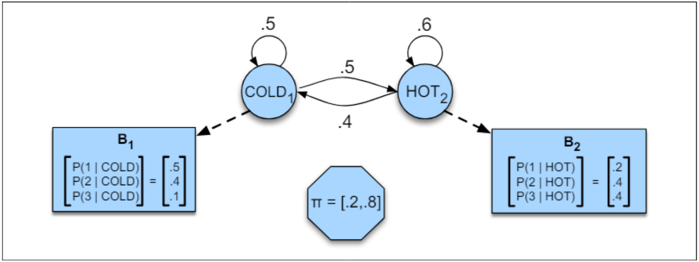
HMM problem to motivate the Forward Algorithm:

- Given HMM λ (like above), what is the probability $P(O|\lambda)$ of a specific observation sequence O (evidence e.g. 3 1 3)?

Hidden Markov Model (HMM) Review

The Forward Algorithm

The Forward Algorithm - Ice Cream Example



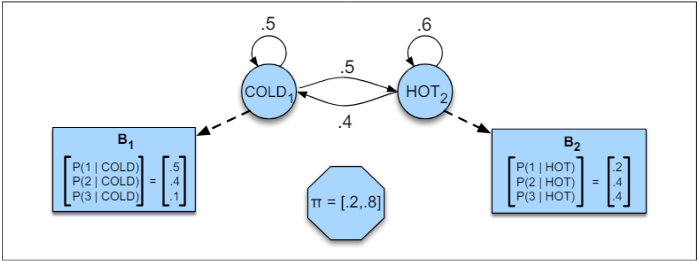
HMM problem to motivate the Forward Algorithm:

- Given HMM λ (like above), what is the probability $P(O|\lambda)$ of a specific observation sequence O (evidence e.g. 3 1 3)?

First, consider an easier problem: suppose our states are not hidden (we just have a "Markov model") and we have $Q =$ (hot hot cold).

What is the probability (aka likelihood) of $O = 3, 1, 3$?

The Forward Algorithm - Ice Cream Example

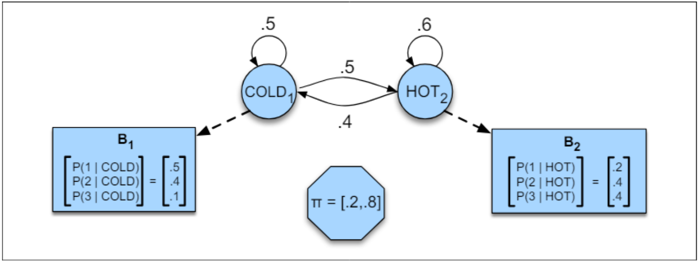


First, consider an easier problem: suppose our states are not hidden (we just have a "Markov model") and we have $Q =$ (hot hot cold).

What is the probability (aka likelihood) of $O = 3, 1, 3$?

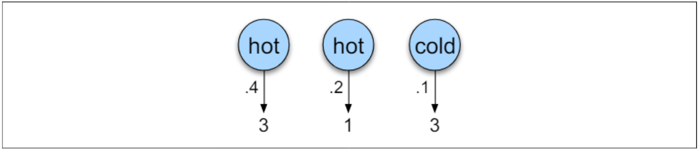
$$P(O|Q) = \prod_{t}^{T} P(o_t|q_t)$$

$$P(3\ 1\ 3|\text{hot hot cold}) = P(3|\text{hot})P(1|\text{hot})P(3|\text{cold})$$

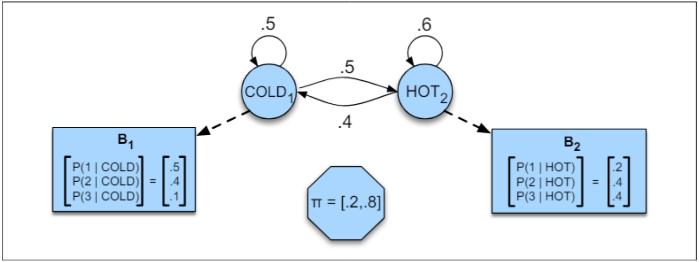


Simplification: probability $O = 3, 1, 3$ given $Q =$ (hot hot cold)?

$$P(3\ 1\ 3|\text{hot hot cold}) = P(3|\text{hot})P(1|\text{hot})P(3|\text{cold})$$



The Forward Algorithm - Ice Cream Example



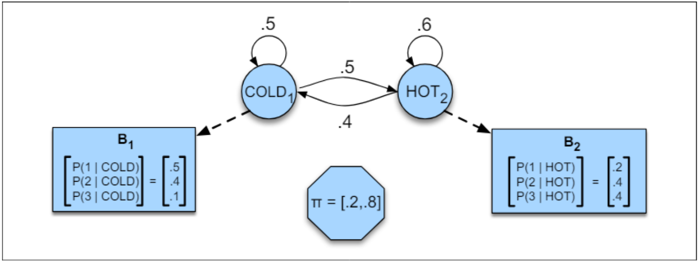
Simplification: probability $O = 3, 1, 3$ given $Q = (\text{hot hot cold})$?

$$P(3\ 1\ 3|\text{hot hot cold}) = P(3|\text{hot})P(1|\text{hot})P(3|\text{cold})$$

Back to the original problem: we don't know the actual weather sequence – it's a HIDDEN Markov model!

What is the probability of 3 1 3 given the HMM?

The Forward Algorithm - Ice Cream Example

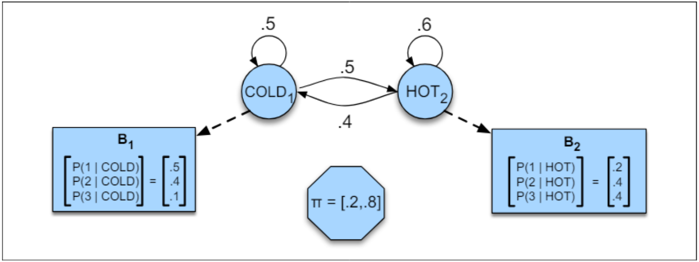


Back to the original problem: we don't know the actual weather sequence – it's a HIDDEN Markov model!

What is the probability of 3 1 3 given the HMM?

Brute Force: Sum over all possible weather sequences:
P(3 1 3, cold cold cold)? P(3 1 3, hot cold cold)?
P(3 1 3, hot hot cold)? etc...?
Then add them all together...

The Forward Algorithm - Ice Cream Example



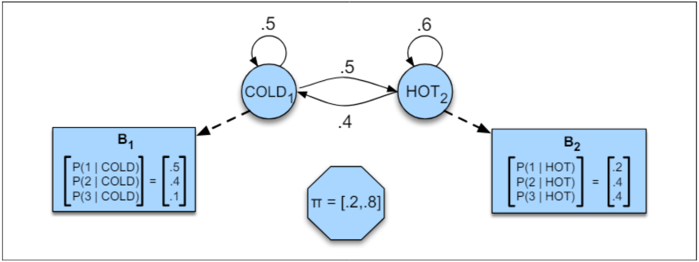
What is the probability of 3 1 3 given the HMM?

Brute Force: Sum over all possible weather sequences:
P(3 1 3, cold cold cold)? P(3 1 3, hot cold cold)? etc...?

$P(O, Q)$ is the joint probability:

$$P(O, Q) = P(O|Q)P(Q) = \prod_t^T P(o_t|q_t) \prod_t^T P(q_t|q_{t-1})$$

The Forward Algorithm - Ice Cream Example



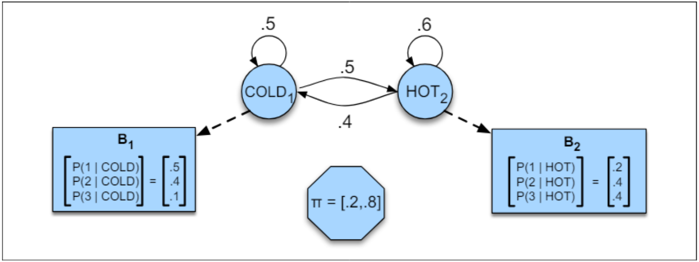
$P(O, Q)$ is the joint probability:

$$P(O, Q) = P(O|Q)P(Q) = \prod_{t}^{T} P(o_t|q_t) \prod_{t}^{T} P(q_t|q_{t-1})$$

Example: joint probability of $O = 3\ 1\ 3$ and $Q =$ hot hot cold

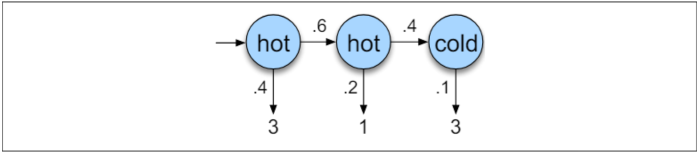
$P(O, Q) = P(3|\text{hot})P(1|\text{hot})P(3|\text{cold})P(\text{hot}|\text{start})P(\text{hot}|\text{hot})P(\text{cold}|\text{hot})$

The Forward Algorithm - Ice Cream Example

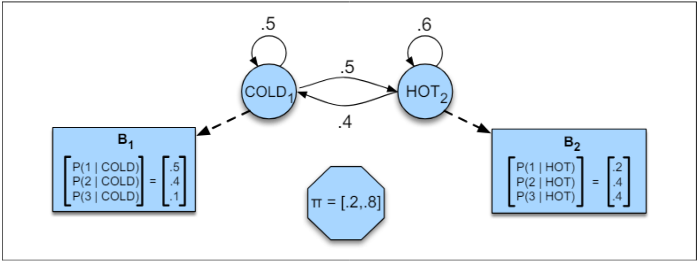


Example: joint probability of $O = 3\ 1\ 3$ and $Q =$ hot hot cold

$P(O, Q) = P(3|\text{hot})P(1|\text{hot})P(3|\text{cold})P(\text{hot}|\text{start})P(\text{hot}|\text{hot})P(\text{cold}|\text{hot})$



The Forward Algorithm - Ice Cream Example



What is the probability of 3 1 3 given the HMM?

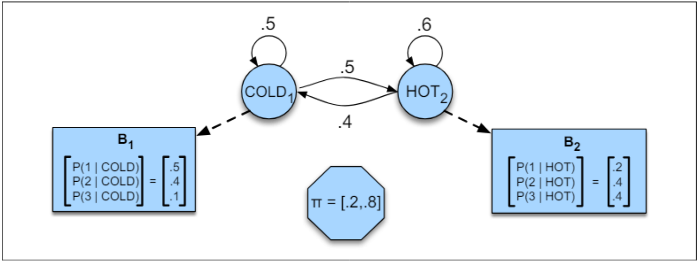
Brute Force: Sum over all possible weather sequences:

P(3 1 3, cold cold cold) + P(3 1 3, hot cold cold) + P(3 1 3, hot hot cold) + ...

This is a N^T operation with N states and T observations!

Not efficient for more complex problems!

The Forward Algorithm - Ice Cream Example

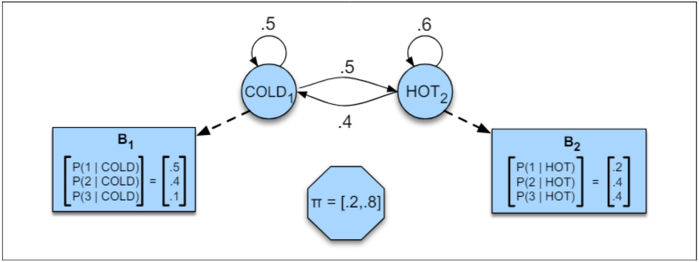


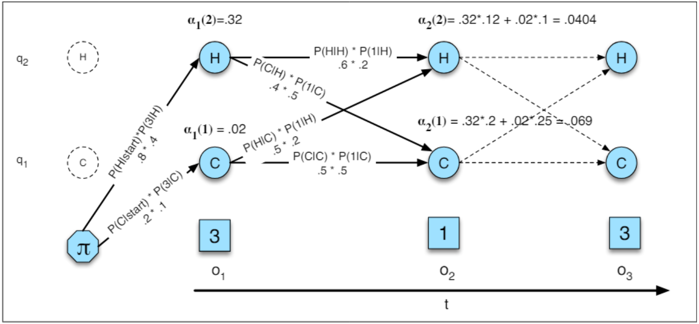
What is the probability of 3 1 3 given the HMM?

P(3 1 3, cold cold cold) + P(3 1 3, hot cold cold) + P(3 1 3, hot hot cold) + ...

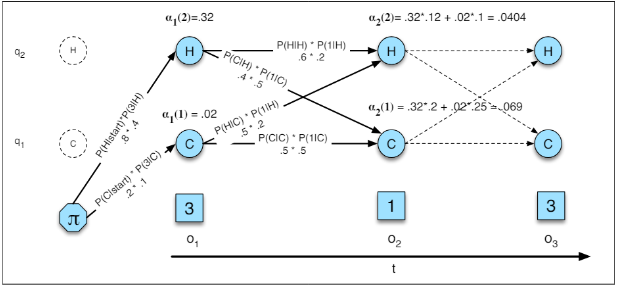
This is a N^T operation with N states and T observations!

Forward Algorithm does this in $O(N^2T)$ via dynamic programming!


.5
.6
.5
COLD1
HOT2
.4
B1
P(1 | COLD)
P(2 | COLD)
P(3 | COLD)
=
.5
.4
.1
π = [.2,.8]
B2
P(1 | HOT)
P(2 | HOT)
P(3 | HOT)
=
.2
.4
.4


α1(2)=.32
α2(2)= .32*.12 + .02*.1 = .0404
q2
q1
H
C
P(H|H) * P(1|H)
.6 * .2
P(C|H) * P(1|C)
.4 * .5
α2(1) = .32*.2 + .02*.25 = .069
α1(1) = .02
P(H|C) * P(1|H)
.5 * .2
P(C|C) * P(1|C)
.5 * .5
P(H|start)*P(3|H)
.8 * .4
P(C|start) * P(3|C)
.2 * .1
π
3
1
3
o1
o2
o3
t

The Forward Algorithm - Ice Cream Example



Formally, for each cell $\alpha_t(j)$ in our lattice structure, we compute

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

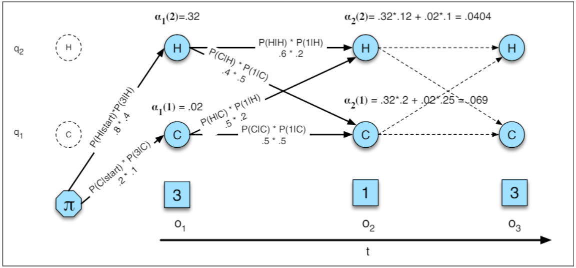
and the probability of sequence 3 1 3 is at the end

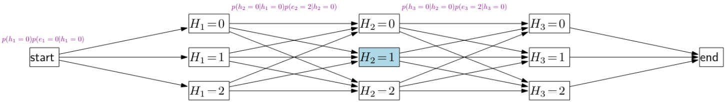
$$P(O|\lambda) = \sum_i^N \alpha_T(i)$$

Hidden Markov Model (HMM) Review

Relating back to Lecture

The Problem of Filtering

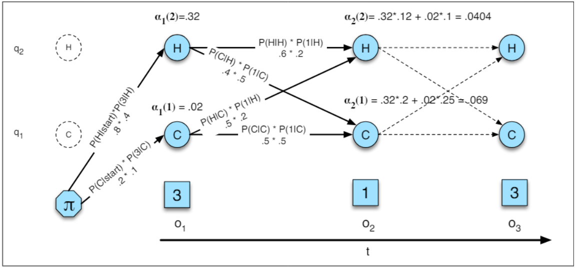


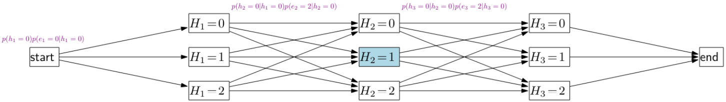


Problem of Filtering: what is the distribution of a hidden state H_i based on the observations aka evidence (E in lecture) so far?

Check your understanding: what is the distribution of q_2 in the ice cream example given observations O: $o_1 = 3$ and $o_2 = 1$?

The Problem of Filtering

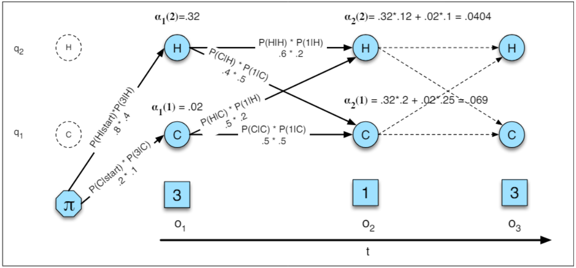


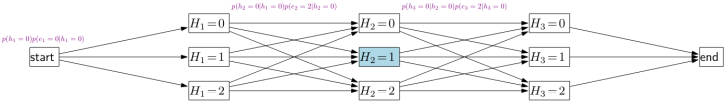


Problem of Filtering: what is the distribution of q_2 in the ice cream example given observations O: $o_1 = 3$ and $o_2 = 1$?

$P(q_2 = \text{H} \mid o_1, o_2) = \frac{0.0404}{0.0404+0.069}$, $P(q_2 = \text{C} \mid o_1, o_2) = \frac{0.069}{0.0404+0.069}$

The Problem of Smoothing

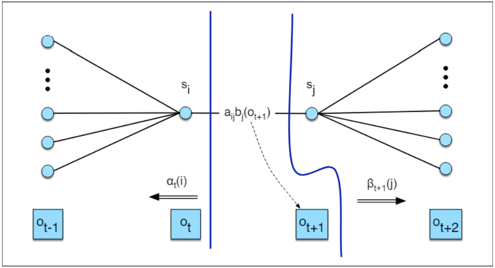




Problem of Smoothing: what is the distribution of a hidden state H_i based ALL observations aka evidence from start to end?

Forward Algorithm is not enough! What if hypothetically a later transition is 0?

The Problem of Smoothing

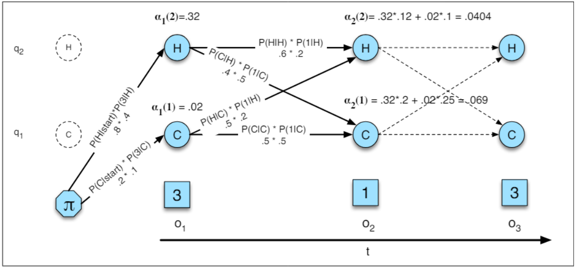


For Smoothing, need Forward AND Backward passes!

- Forward: compute $\alpha_t(i)$ or F from lecture.
- Backward: compute $\beta_t(i)$ or B from lecture.
- Define $S = FB$, that is for each cell in the lattice, multiply the forward and backward results together.

What happens now if there is a 0 along the backward pass?

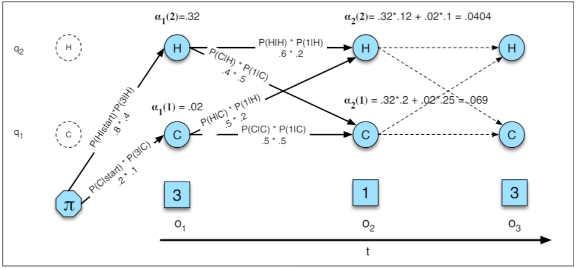
The Problem of Smoothing



- Forward: compute $\alpha_t(i)$ or F from lecture.
- Backward: compute $\beta_t(i)$ or B from lecture.
- Define $S = FB$

Suppose $\beta_2(1) = 0.03, \beta_2(2) = 0.02$ (made up numbers).
What is the distribution of q_2 given all observations O?

The Problem of Smoothing



Suppose $\beta_2(1) = 0.03, \beta_2(2) = 0.02$ (made up numbers).

What is the distribution of q_2 given all observations O?

$$P(q_2 = \text{H} \mid O) = \frac{0.0404 * 0.02}{0.0404 * 0.02 + 0.069 * 0.03}$$

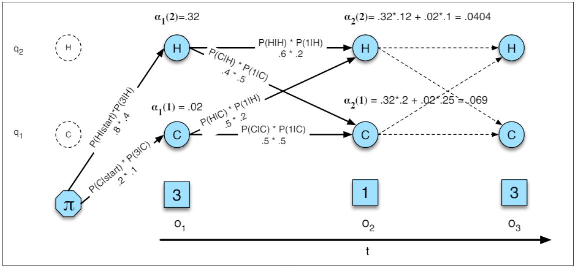
$$P(q_2 = \text{C} \mid O) = \frac{0.069 * 0.03}{0.0404 * 0.02 + 0.069 * 0.03}$$

Check back on the lecture slides to make sure you see the parallel!

Hidden Markov Model (HMM) Review

Particle Filtering

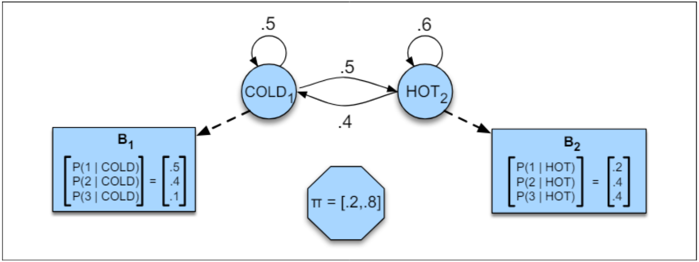
Motivation for Particle Filtering



For T observations and N possible states (i.e. $|\text{domain}| = N$), the Forward-Backward Algorithm is $O(2 * N^2 T) \rightarrow O(N^2 T)$.

This can still be slow if N is large! Or consider if the domain is based on a continuous function, e.g. instead of just hot or cold, we have to consider a spectrum of floating point temperatures $[0, 100]$.

Particle Filtering

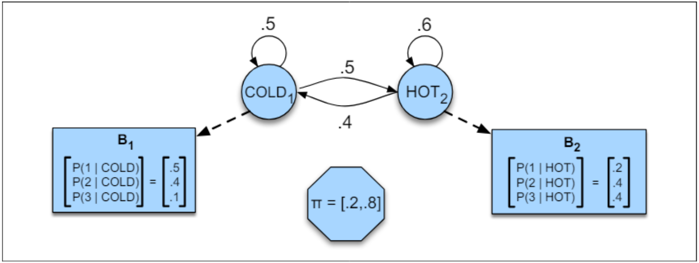


Big idea of Particle Filtering: introduce sampling!

1. First, we propose assignments aka particles to each hidden state by sampling from the transition probabilities.

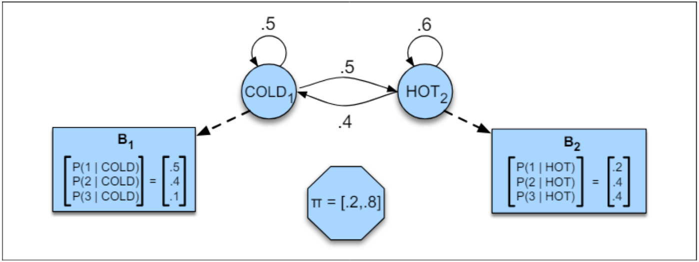
Example: proposing a value for q_1 involves sampling from $P(H|\text{start}) = 0.8$ and $P(C|\text{start}) = 0.2$, i.e. we have an 80% chance to pick hot, 20% chance to pick cold.

Particle Filtering



1. First, we propose assignments aka particles to each hidden state by sampling from the transition probabilities.
2. Second, we weight each assignment by the emission probabilities.

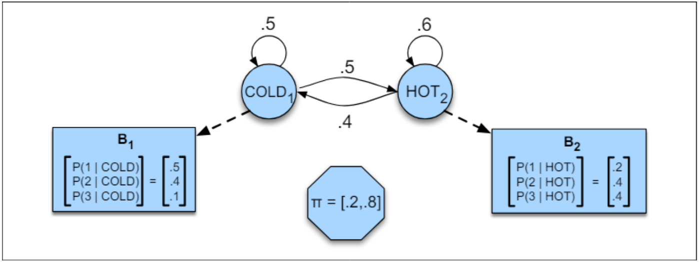
Particle Filtering



2 Second, we weight each assignment by the emission probabilities.

Example: suppose we have 3 particles of $q_1 = H, q_1 = H, q_1 = C$, and we have the observation $o_1 = 1$.
Then the weights of our particles are $P(1|H) = 0.2$, $P(1|H) = 0.2$, $P(1|C) = 0.5$ respectively.

Particle Filtering



1. First, we propose assignments aka particles to each hidden state by sampling from the transition probabilities.
2. Second, we weight each assignment by the emission probabilities.
3. Third, we resample new assignments from the particles based on the weight distributions.

3 Third, we resample new assignments from the particles based on the weight distributions.

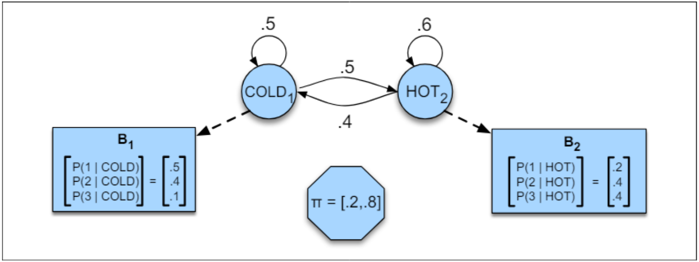
Example: suppose we have 3 particles of $q_1 = H, q_1 = H, q_1 = C$, and we have the observation $o_1 = 1$.
Then the weights of our particles are $P(1|H) = 0.2$, $P(1|H) = 0.2$, $P(1|C) = 0.5$ respectively.

Now to resample, we have the distribution:

- $P(q_1 \rightarrow H) = \frac{0.2}{0.2+0.2+0.5}$
- $P(q_1 \rightarrow H) = \frac{0.2}{0.2+0.2+0.5}$
- $P(q_1 \rightarrow C) = \frac{0.5}{0.2+0.2+0.5}$

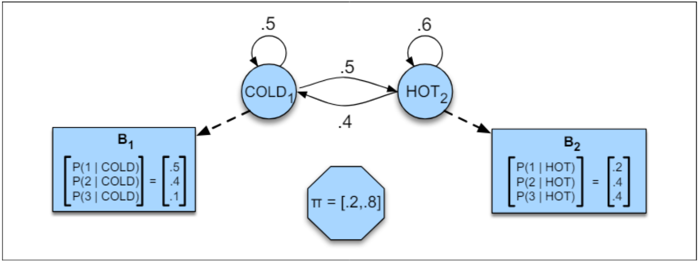
Notice how even though our initial proposal had a higher chance to pick $q_1 = H$, we now have a higher chance to get $q_1 = C$!
The resampling takes into account the observations!

Particle Filtering



Suppose after all that, we have new assignments for our 3 particles: $q_1 = C, q_1 = C, q_1 = H$...
And repeated the propose process for q_2 to get: $(q_1, q_2) = (C, H)$; $(q_1, q_2) = (C, C)$; $(q_1, q_2) = (H, C)$ with $o_2 = 3$...

Particle Filtering



And repeated the propose process for q_2 to get: $(q_1, q_2) = (C, H)$; $(q_1, q_2) = (C, C)$; $(q_1, q_2) = (H, C)$ with $o_2 = 3$...

The weight process then assigns the particles:

- $(q_1, q_2) = (C, H)$: $P(3|H) = 0.4$
- $(q_1, q_2) = (C, C)$: $P(3|C) = 0.1$
- $(q_1, q_2) = (H, C)$: $P(3|C) = 0.1$

Particle Filtering

The weight process then assigns the particles:

- $(q_1, q_2) = (C, H)$: $P(3|H) = 0.4$
- $(q_1, q_2) = (C, C)$: $P(3|C) = 0.1$
- $(q_1, q_2) = (H, C)$: $P(3|C) = 0.1$

And the resample process then samples from the above 3 options, that is:

- $(q_1, q_2) = (C, H)$ has a $4/6$ chance of being picked.
- The other two each have a $1/6$ chance of being picked.

And so a possible resampling result might yield the particles:
$(q_1, q_2) = (C, H), (C, H)$, and (C, C).
And you'd repeat the process with q_3...