

# States of States of Knowledge II

Reading for this lecture:

CMR articles *PRATISP*, *IACP*, *IACPLCOCS*

# States of States of Knowledge

Causal states:

Conditions of knowledge that lead to optimal prediction

Last lecture:

Mixed states as a hierarchy of states for optimal prediction

# States of States of Knowledge

## Agenda:

- Review
- Mixed State Presentation
- How to Calculate
- Efficient Block Entropies
- Synchronization Information

# States of States of Knowledge

## Mixed State Presentations

**Mixed states** are state distributions induced by seeing a word:

1. Let  $w \in \mathcal{A}^*$  such that  $|w| = L$ .
2. Let  $\mu_t(\lambda)$  be a RV for the state distribution  $S_t$  at time  $t$ .

Then:

$$\Pr(\mu_{t+L}(w) = \sigma) \equiv \Pr(S_{t+L} = \sigma | X_t^L = w, \mu_t(\lambda))$$

Comments:

- Mixed states  $\mu_t(w)$  are RVs.
- Conditioned on event  $w$  and random variable  $\mu_t(w)$ .
- Typically, we take  $t = 0$  and  $\mu_0(\lambda) = \pi$ .

# States of States of Knowledge

## Mixed State Presentations ...

### Mixed states as vectors:

Write  $\mu_{t+L}(w)$  as vector in  $|\mathcal{V}|$ -dimensional vector space:

$$\begin{aligned}\mu_{t+L}(w) &\equiv \Pr(S_{t+L} | X_t^L = w, \mu_t(\lambda)) \\ &= \frac{\Pr(X_t^L = w, S_{t+L} | \mu_t(\lambda))}{\Pr(X_t^L = w | \mu_t(\lambda))} \\ &= \frac{\mu_t(\lambda) T^{(w)}}{\mu_t(\lambda) T^{(w)} \mathbf{1}}\end{aligned}$$

# States of States of Knowledge

## Mixed State Presentations ...

### Interpretation of mixed states:

- Uncertainty in state given word and starting distribution.
- $H[\mu_L(w)] = 0$ : State is known with probability 1.
- Mixed states with zero entropy are basis vectors (“pure” states).
- Distributions are mixtures over pure states.
- Points in a geometric space.

# States of States of Knowledge

## Mixed State Presentations ...

## Interpretation of mixed states ...

- Predictive equivalence relation:

$$\overleftarrow{x} \sim \overleftarrow{x}' \iff \Pr(\overrightarrow{X} | \overleftarrow{x}) = \Pr(\overrightarrow{X} | \overleftarrow{x}')$$

- Mixed-state equivalence relation:

$$w \sim_{\text{MSP}} w' \iff \mu(w) = \mu(w')$$

# States of States of Knowledge

## Mixed State Presentations ...

### MSP Dynamic:

- Words have a natural dynamic:  $w \xrightarrow{s} ws$
- Dynamic over mixed states gives evolution of state uncertainty:

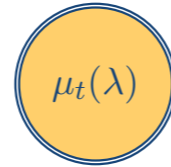
$$\begin{array}{ccc} w & \xrightarrow{s} & ws \\ \downarrow & & \downarrow \\ \mu(w) & \xrightarrow{s} & \mu(ws) \end{array}$$

- Mixed state unifilarity inherited from dynamic over words.
- Calculating the MSP is a method to **unifilarize** a presentation.



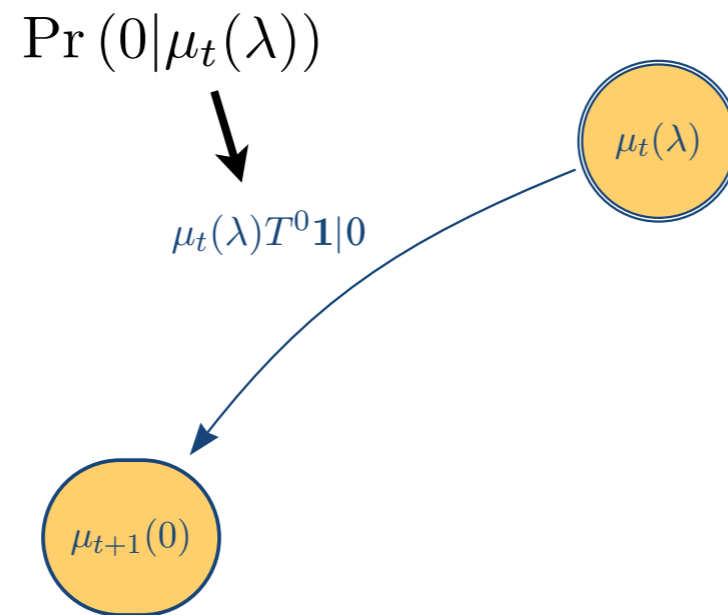
# States of States of Knowledge

Mixed State Presentations ... How to calculate


$$\mu_t(\lambda)$$

# States of States of Knowledge

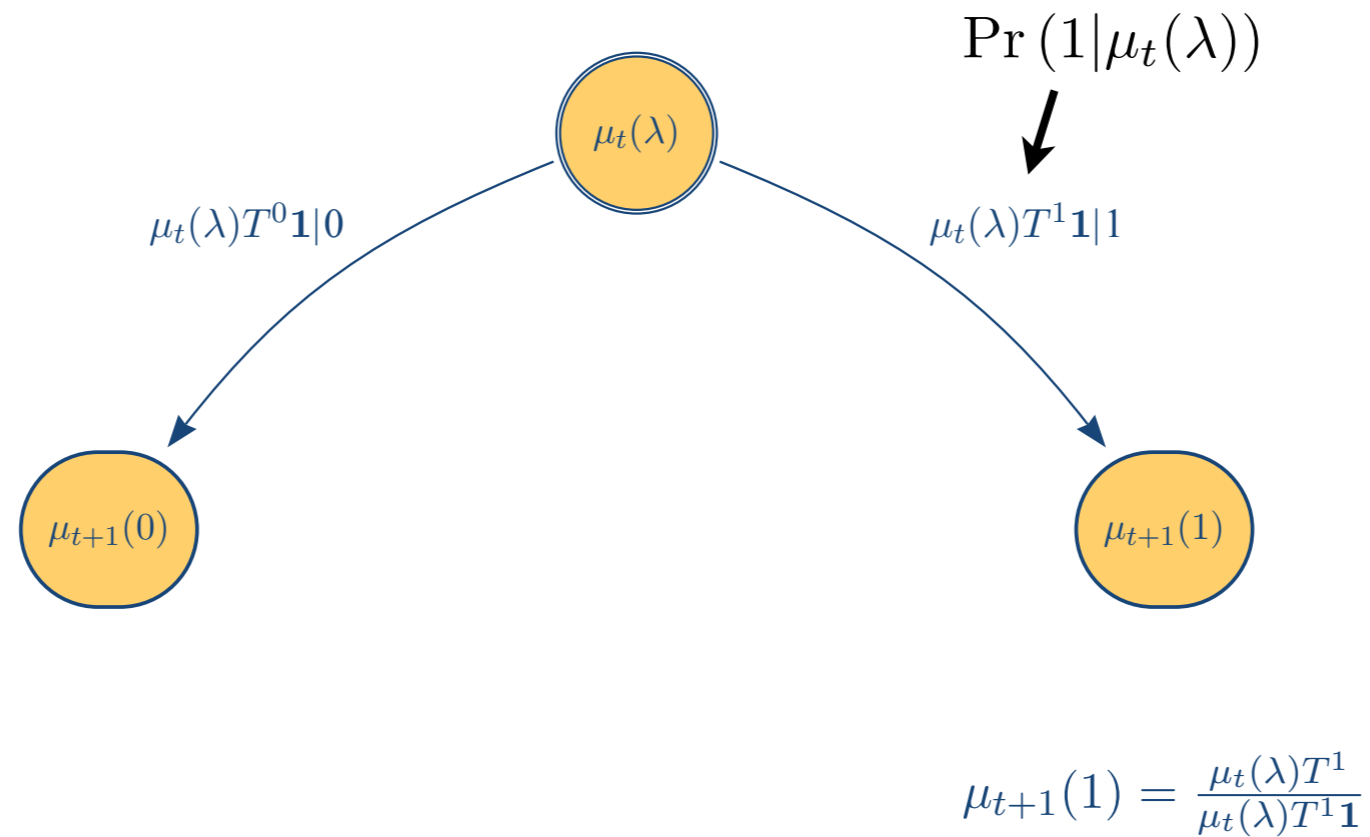
## Mixed State Presentations ... How to calculate



$$\mu_{t+1}(0) = \frac{\mu_t(\lambda)T^0}{\mu_t(\lambda)T^0\mathbf{1}}$$

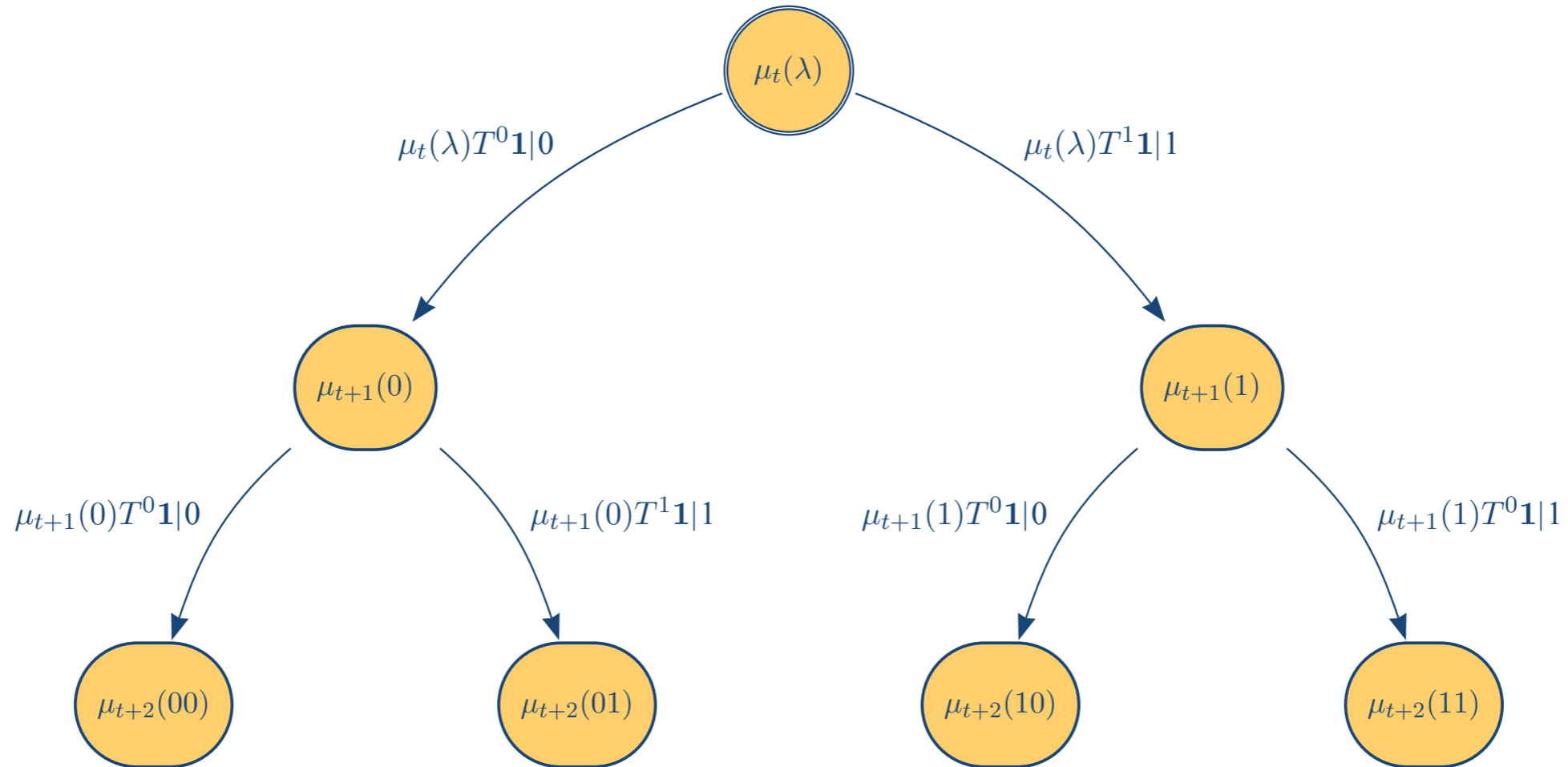
# States of States of Knowledge

## Mixed State Presentations ... How to calculate



# States of States of Knowledge

## Mixed State Presentations ... How to calculate



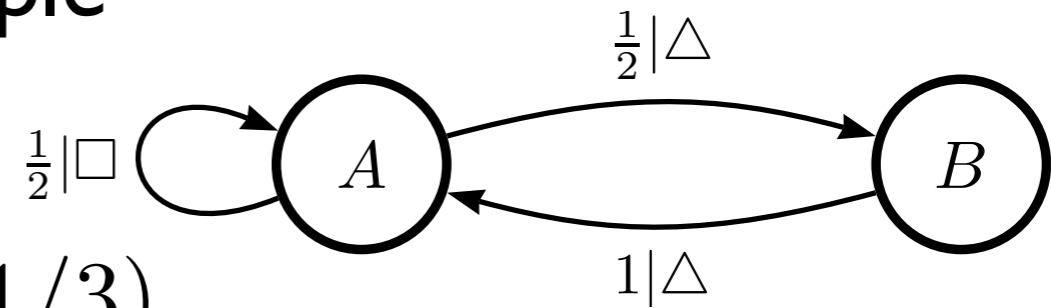
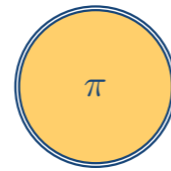
Different words can induce the same uncertainty in state

# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$

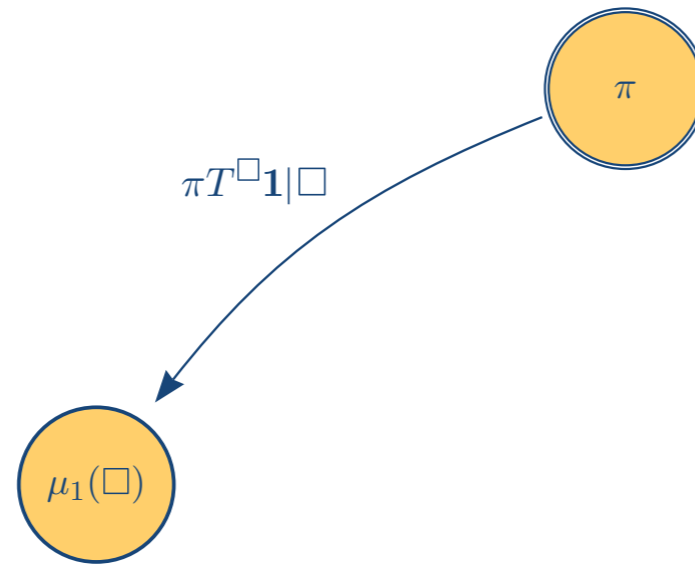
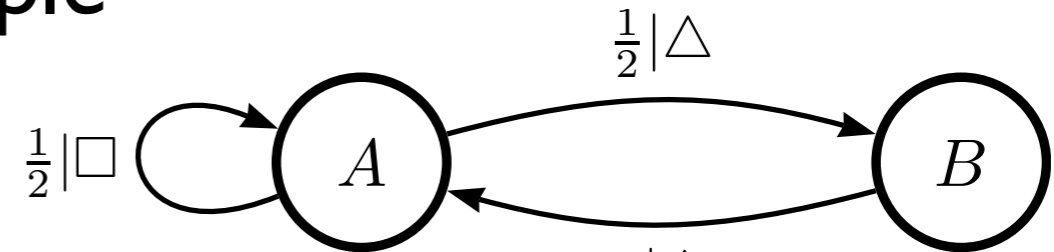


# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$



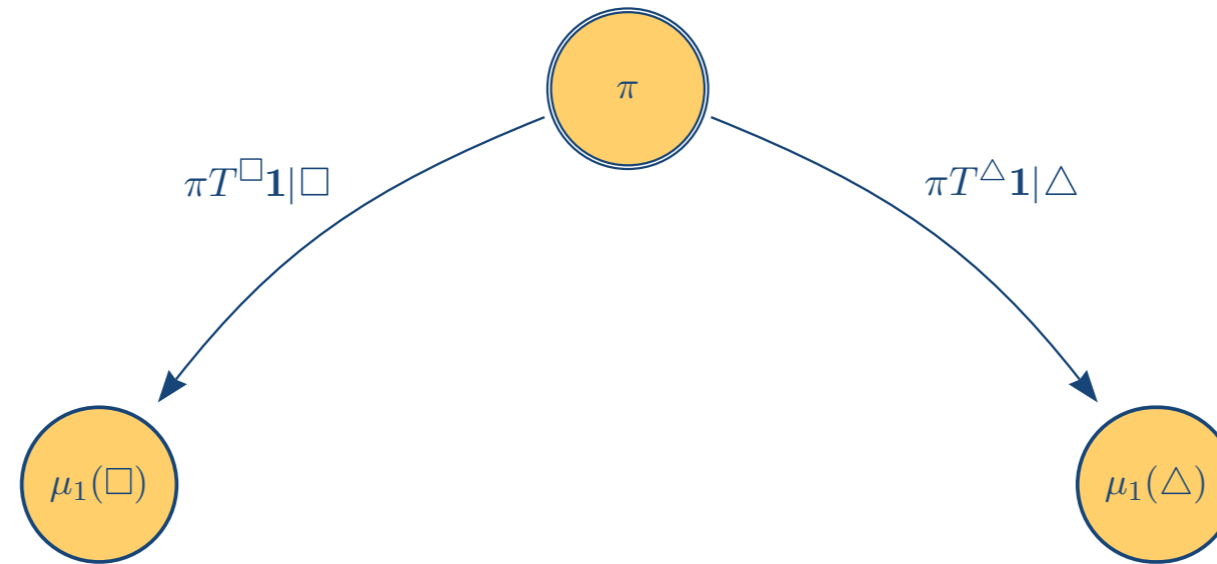
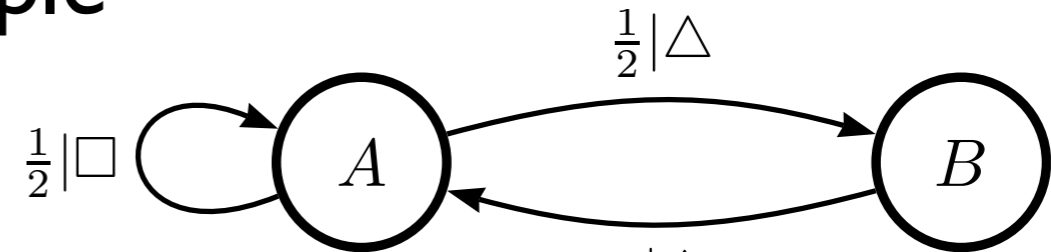
$$\mu_1(\square) = \frac{\pi T^{\square}}{\pi T^{\square} \mathbf{1}}$$

# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$



$$\mu_1(\square) = \frac{\pi T^{\square} \mathbf{1}}{\pi T^{\square} \mathbf{1}}$$

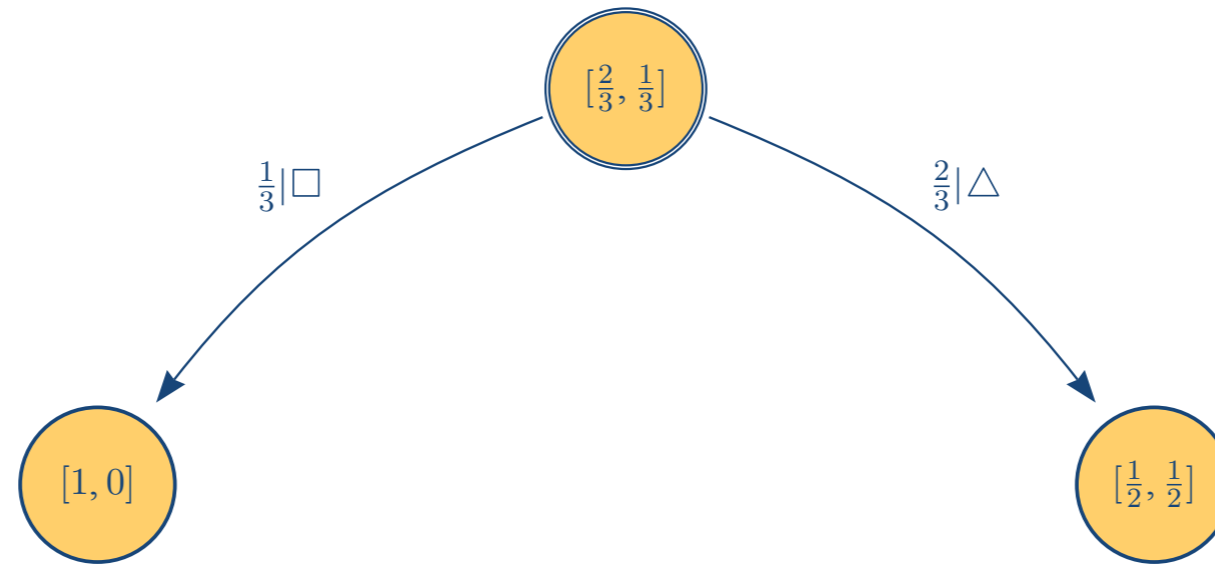
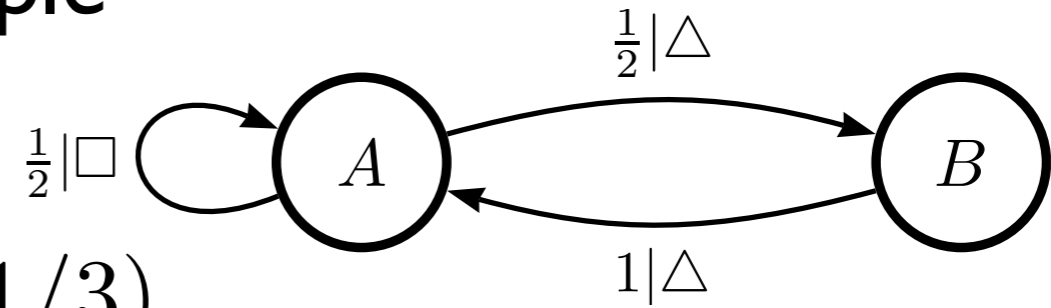
$$\mu_1(\Delta) = \frac{\pi T^{\Delta} \mathbf{1}}{\pi T^{\Delta} \mathbf{1}}$$

# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$



$$\mu_1(\square) = [1, 0]$$

$$\mu_1(\triangle) = [1/2, 1/2]$$

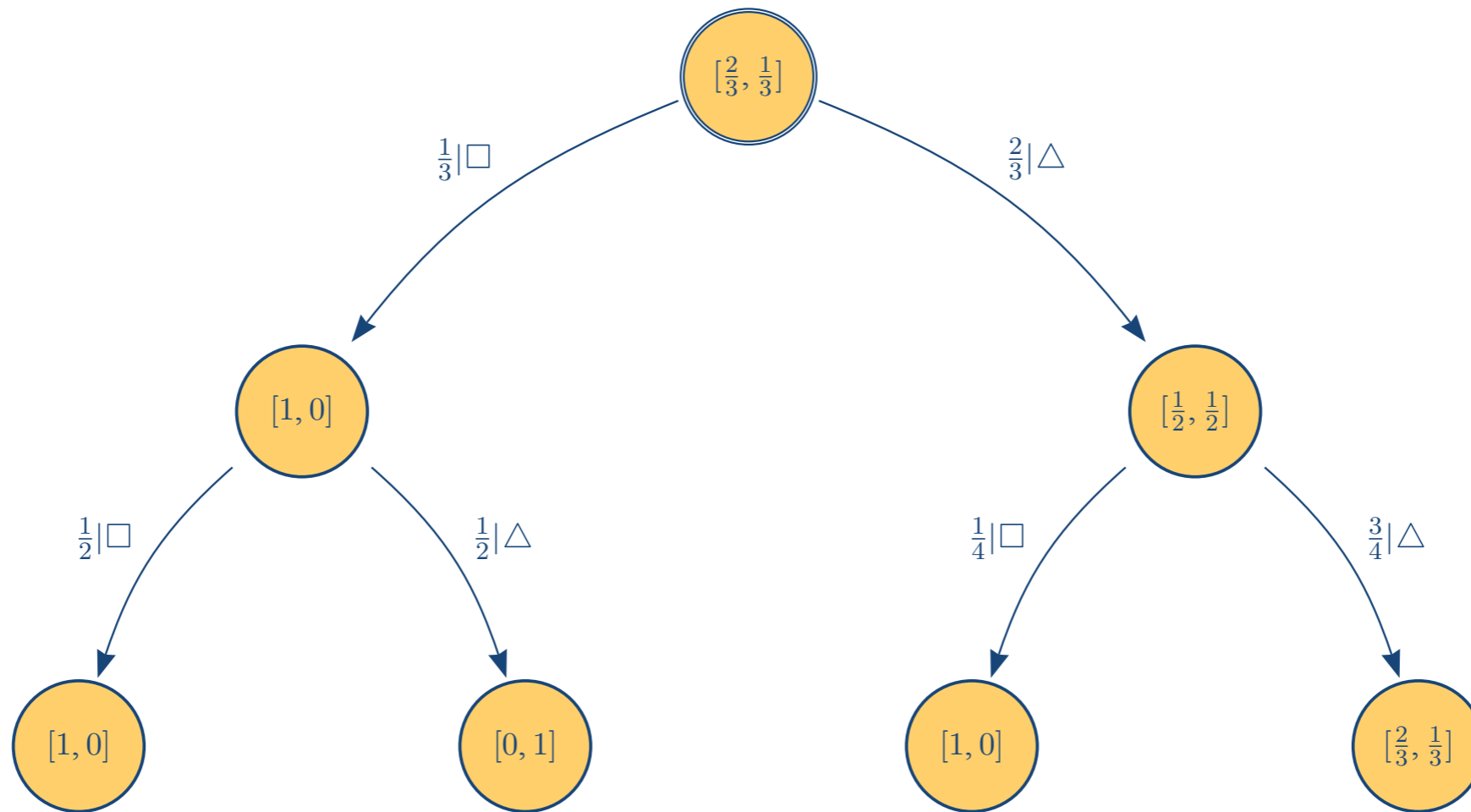
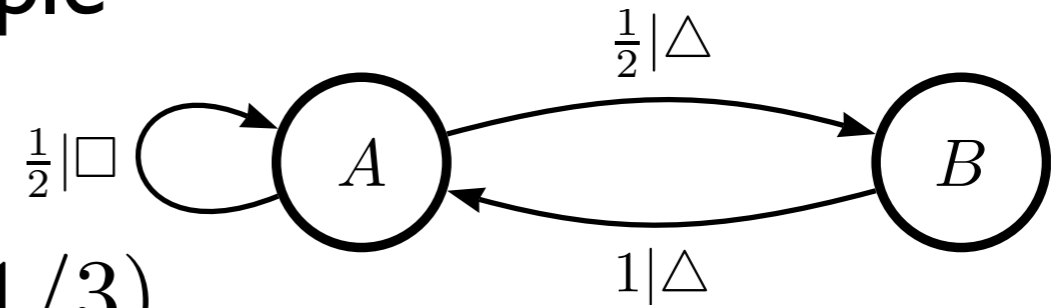


# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$

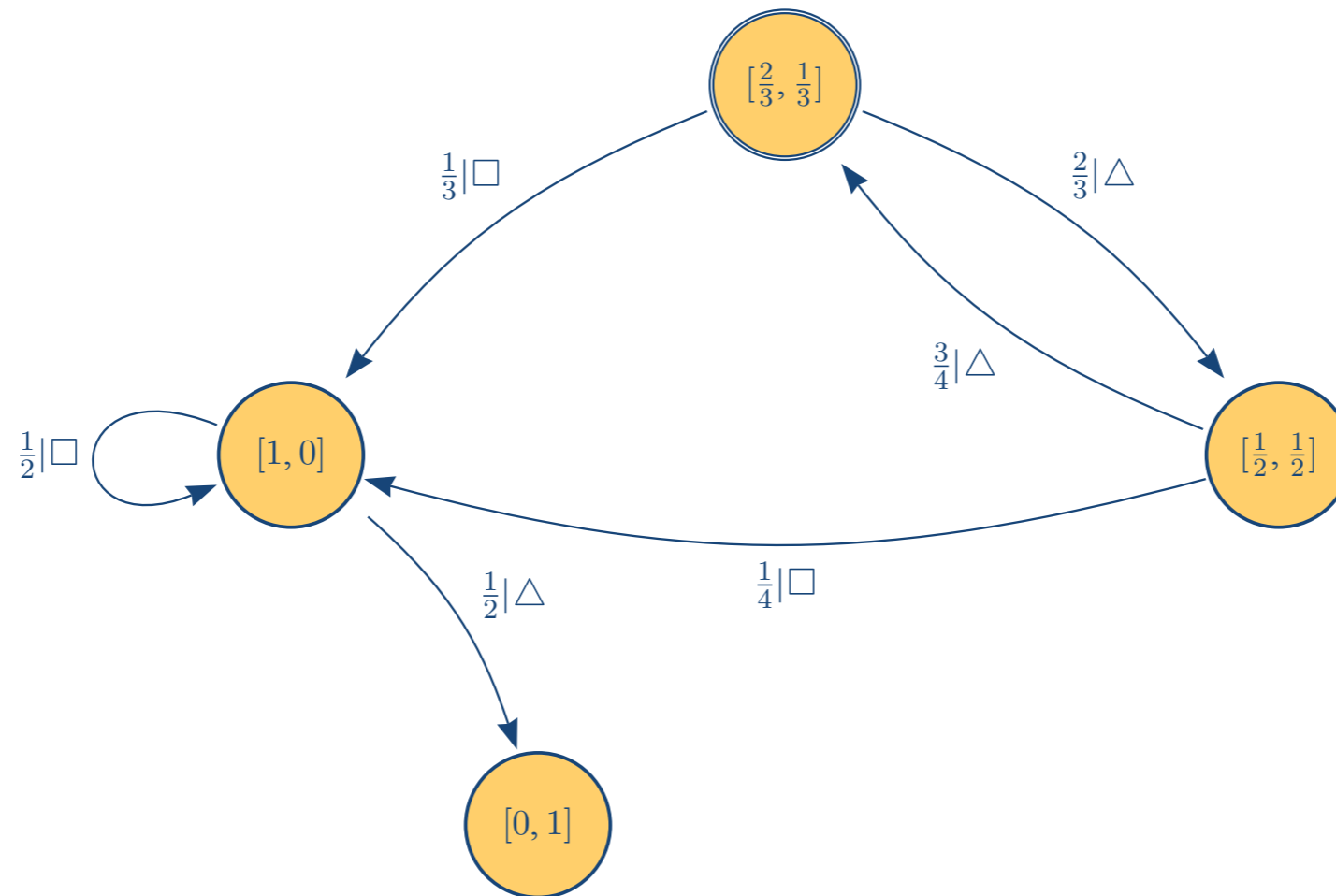
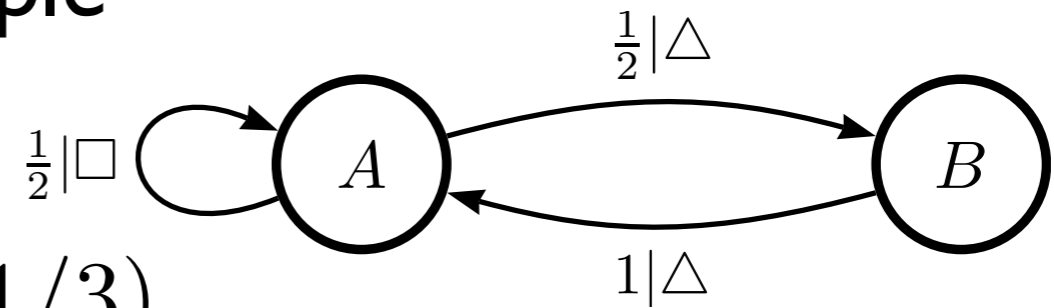


# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$

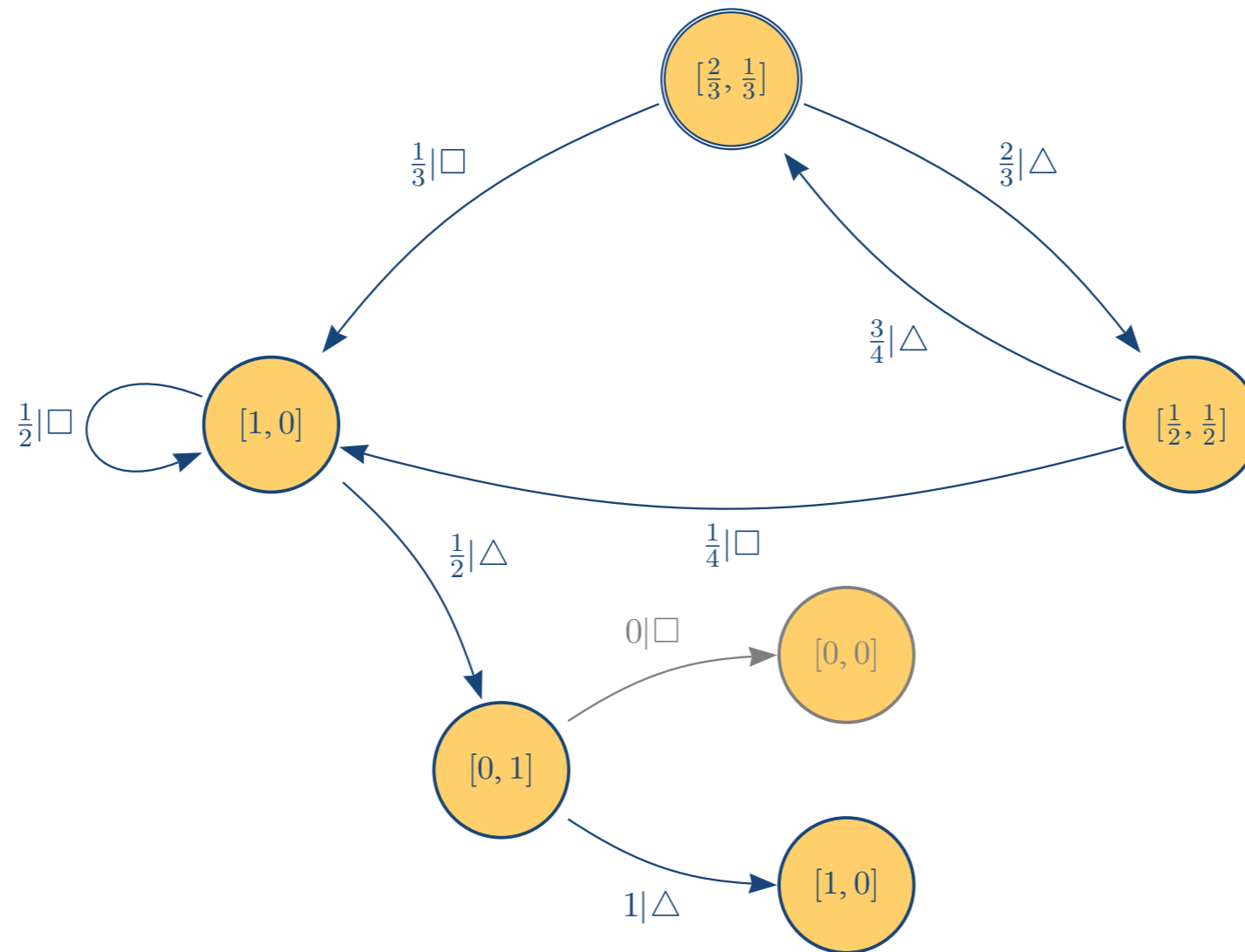
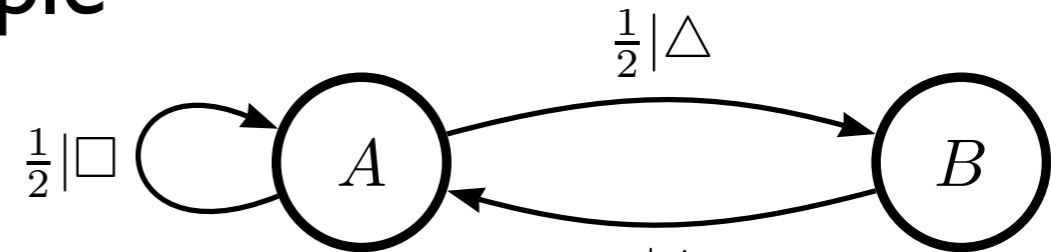


# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$

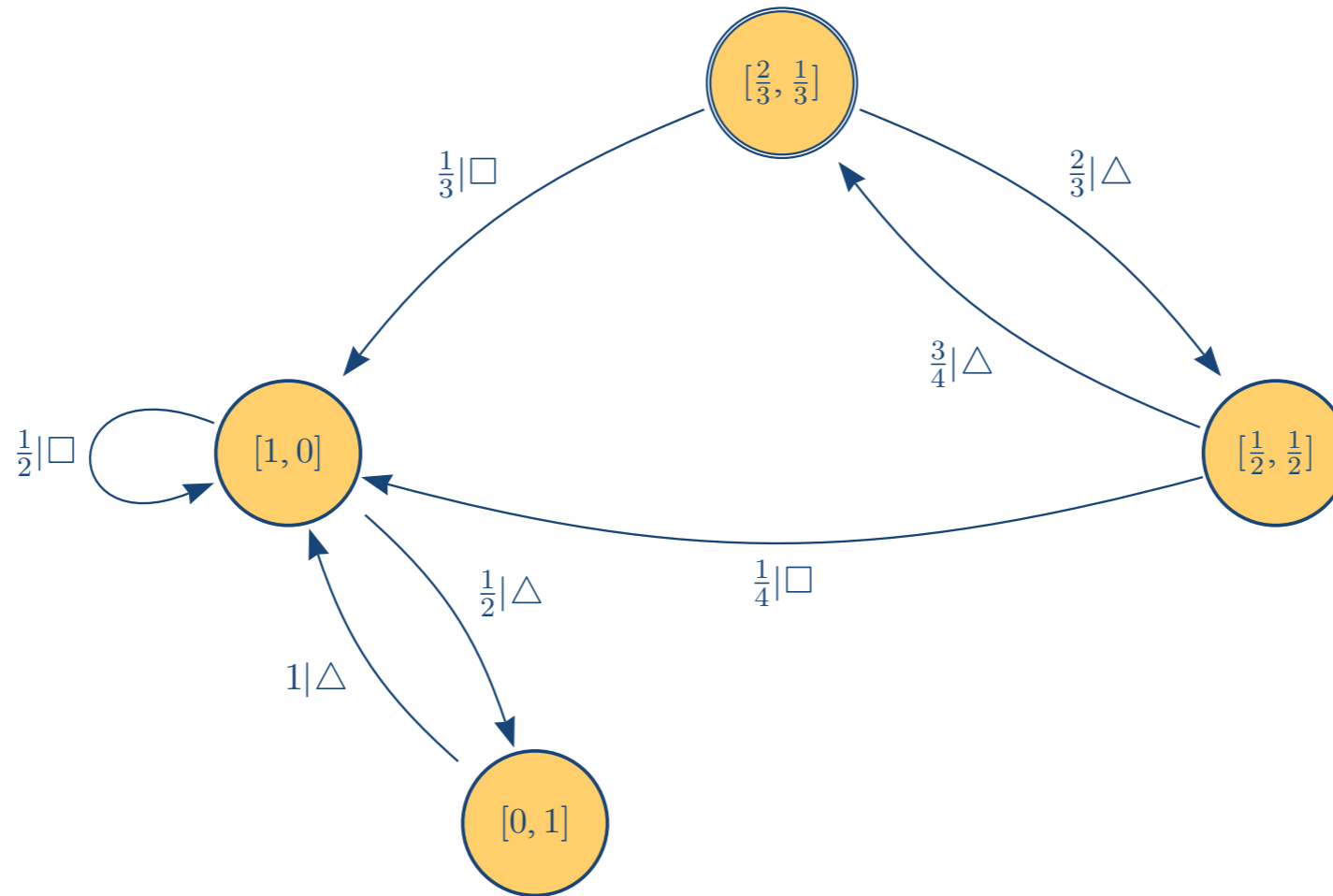
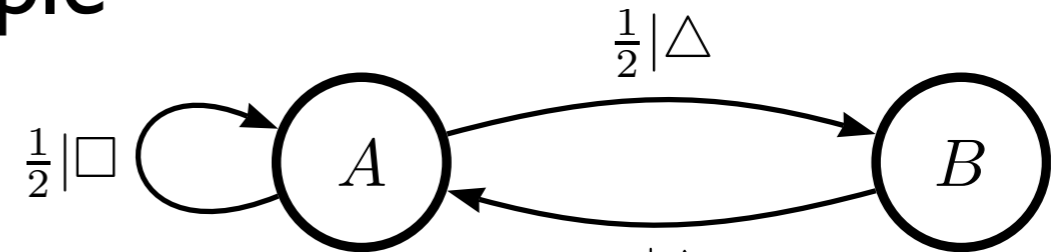


# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$

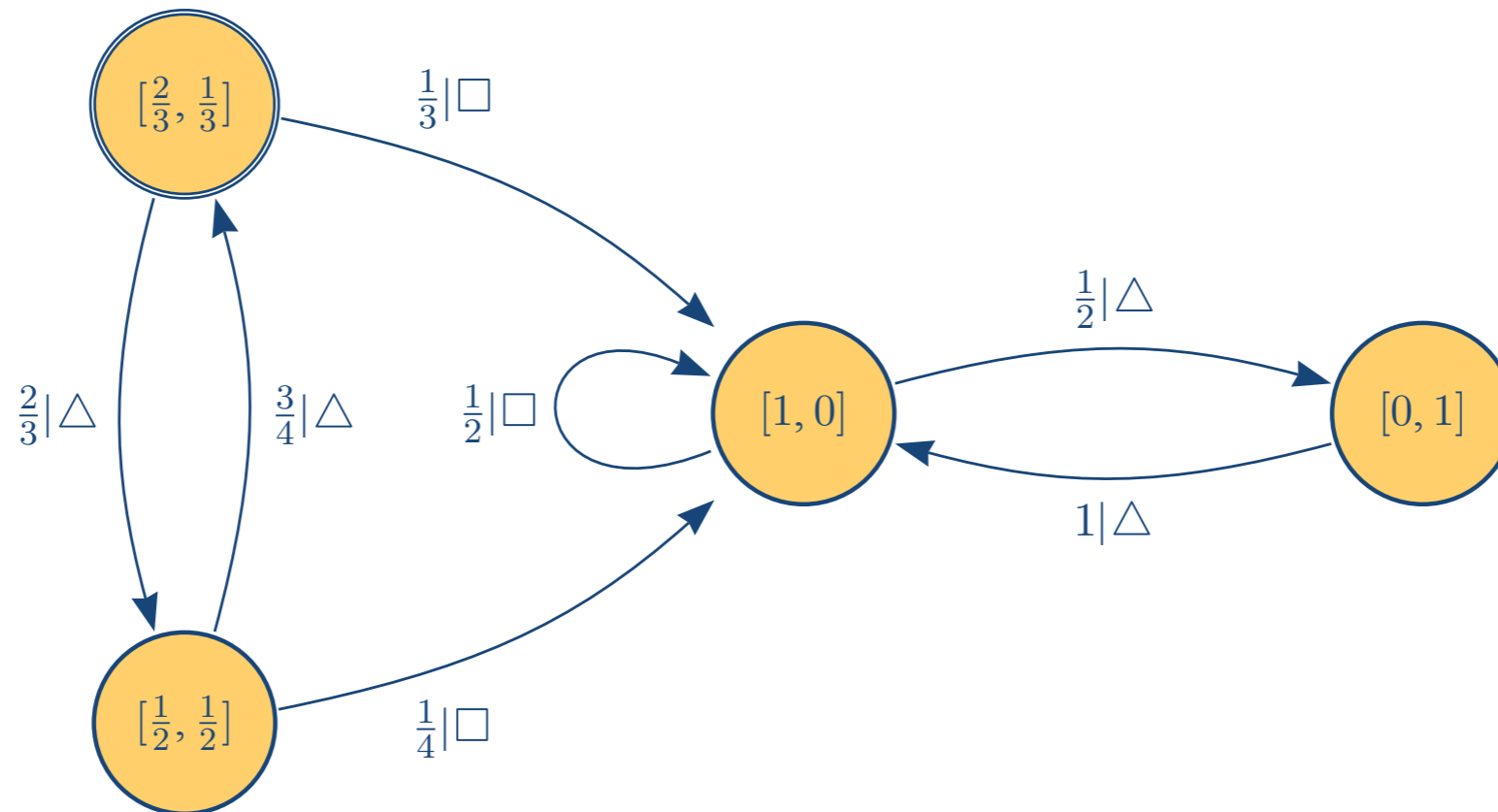
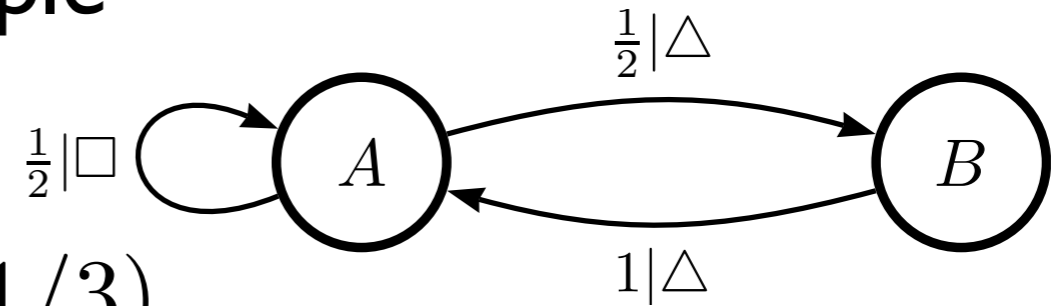


# States of States of Knowledge

## Mixed State Presentations ... Example

Even Process

$$\mu_0(\lambda) \equiv \pi = (2/3, 1/3)$$



# States of States of Knowledge

## Mixed State Presentation

### Definition:

- Let  $M = (\mathcal{A}, \mathcal{V}, \mathcal{D})$  be a presentation of a process  $\mathcal{P}$ , with
- Alphabet  $\mathcal{A} \rightarrow X_t = x \in \mathcal{A}$
- States  $\mathcal{V} \rightarrow S_t = \sigma \in \mathcal{V}$
- Dynamic  $\mathcal{D} = \{T^{(x)} : x \in \mathcal{A}\}$   
where  $T_{\sigma\sigma'}^{(x)} \equiv \Pr(X_t = x, S_{t+1} = \sigma' | S_t = \sigma)$

# States of States of Knowledge

## Mixed State Presentation ...

### Definition: **Mixed State Presentation (MSP)**

- The MSP of  $M$  starting from  $\mu_t(\lambda)$  is:

$$\mathcal{U}(M, \mu_t(\lambda)) = (\mathcal{A}, \mathcal{V}_{\mathcal{U}}, \mathcal{D}_{\mathcal{U}}, \mu_t(\lambda))$$

- Same alphabet:  $x \in \mathcal{A}$
- States are mixed states:  $\mathcal{V}_{\mathcal{U}} = \{\mu_{t+|w|}(w) : w \in \mathcal{A}^*\}$
- Induced dynamic:  $\mathcal{D}_{\mathcal{U}} = \{T^{(x)} : x \in \mathcal{A}\}$

where  $T_{\eta\eta'}^{(x)} \equiv \Pr(X_t = x, R_{t+1} = \eta' | R_t = \eta)$

$$\Pr(X_t = x, R_{t+1} = \eta | R_t = \mu_t(\lambda)) \equiv \begin{cases} \mu_t(\lambda) T^{(x)} \mathbf{1} & \eta = \mu_{t+1}(x) \\ 0 & \eta \neq \mu_{t+1}(x) \end{cases}$$

- Start state:  $\mu_t(\lambda)$

# States of States of Knowledge

## Mixed State Presentation ...

### Comments:

- For  $|w| = L$  and  $x \in \mathcal{A}$ , mixed states are defined recursively:

$$\mu_{t+L+1}(wx) = \frac{\mu_{t+L}(w)T^x}{\mu_{t+L}(w)T^x \mathbf{1}}$$

- By construction, the MSP is unifilar.
- States of  $\mathcal{U}(M, \mu_t(\lambda))$  are mixed states of  $M$ .
- Mixed states of mixed states? Definitely:

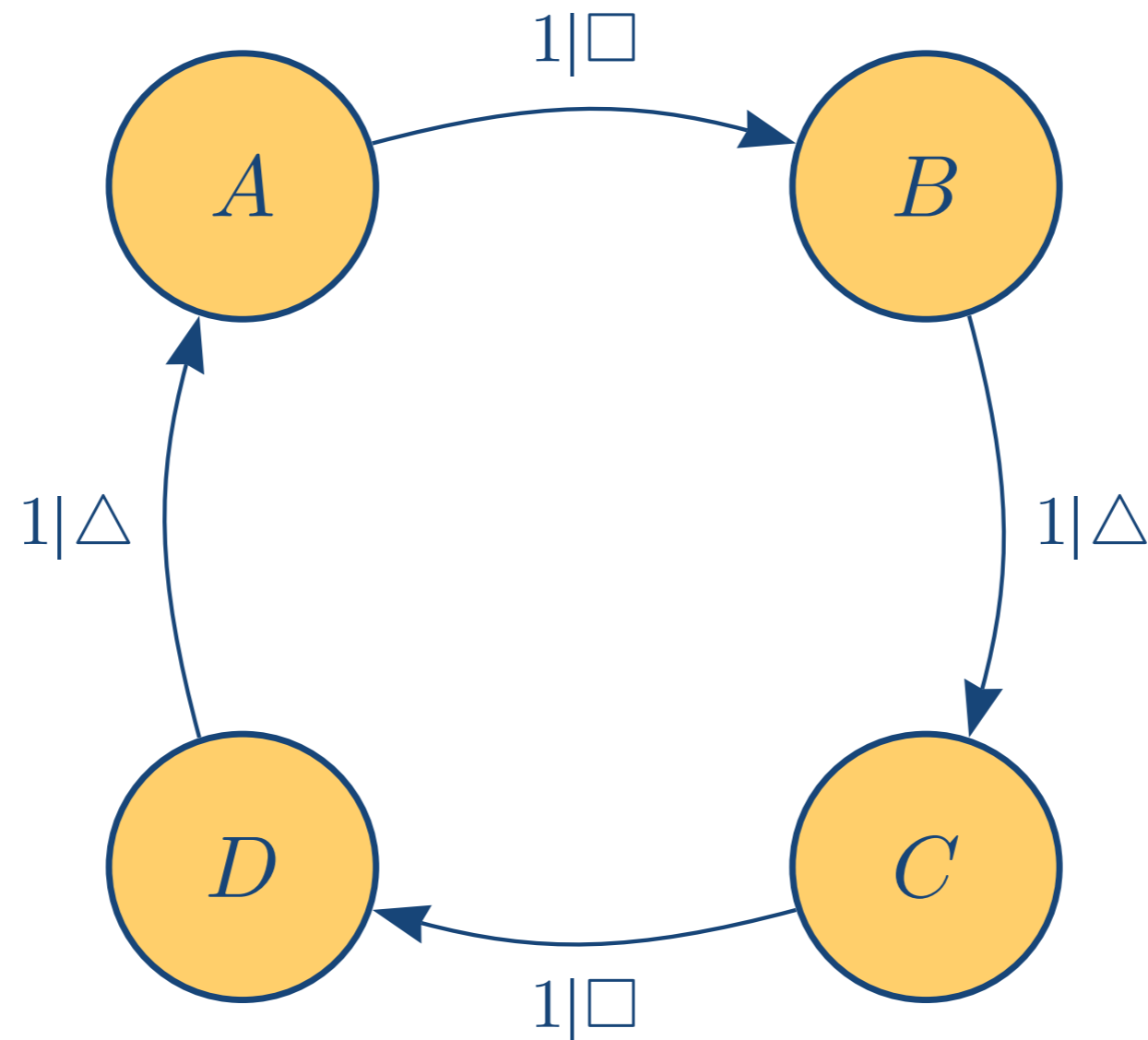
$$\mathcal{U}\left(\mathcal{U}(M, \mu_t(\lambda)), \mathbf{1}_{\{\mu_t(\lambda)\}}\right) \cong \mathcal{U}(M, \mu_t(\lambda))$$

- MSP calculation is a way to convert nonunifilar to unifilar.



# States of States of Knowledge

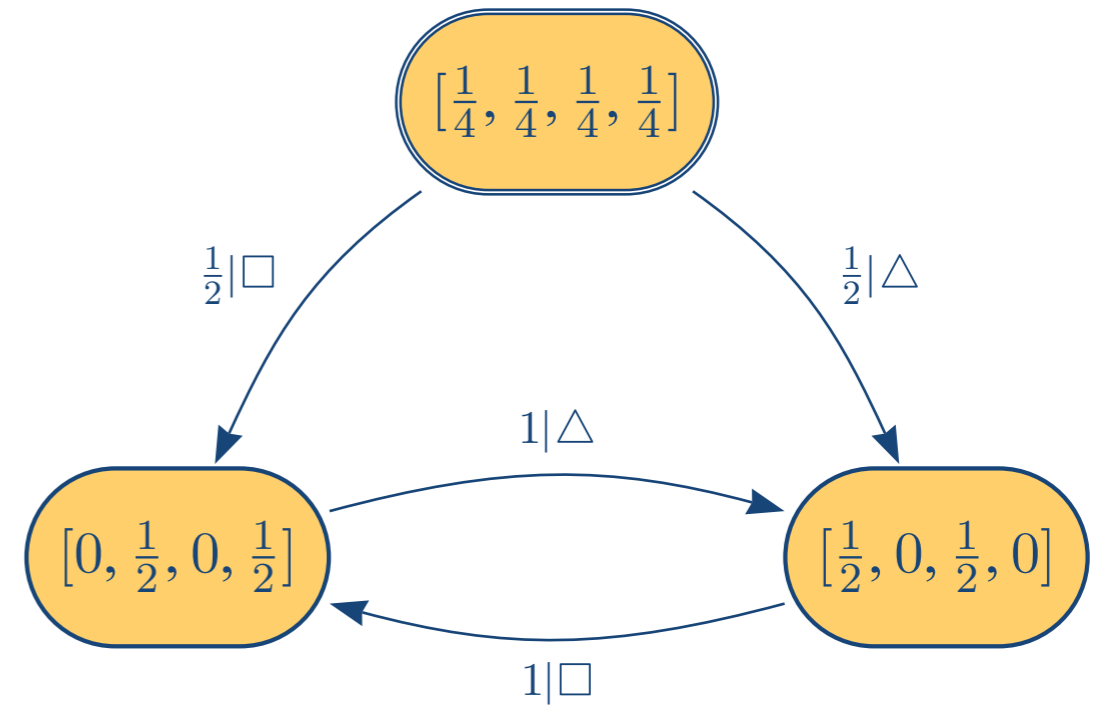
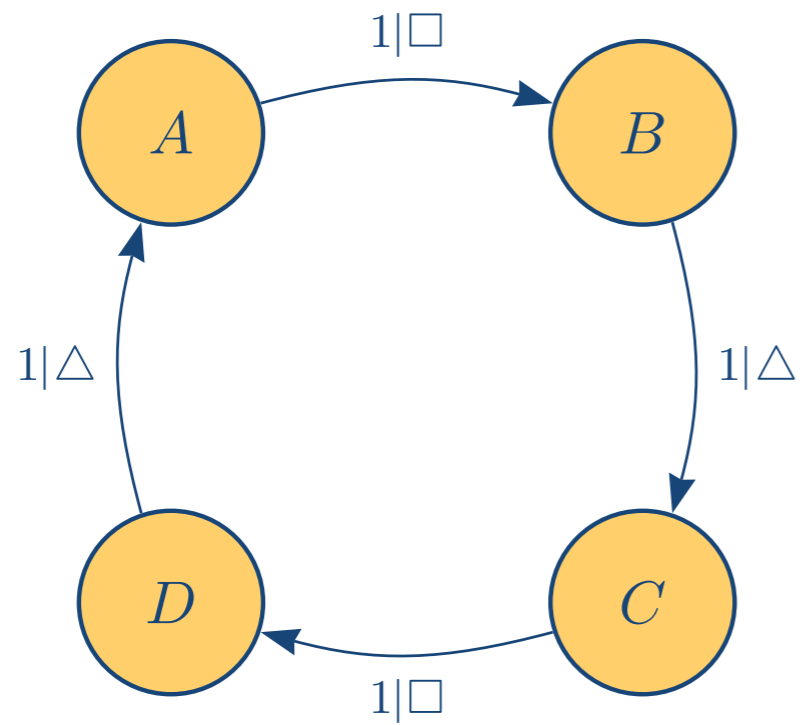
Transient mixed states:



## 4-State presentation of the Period-2 Process

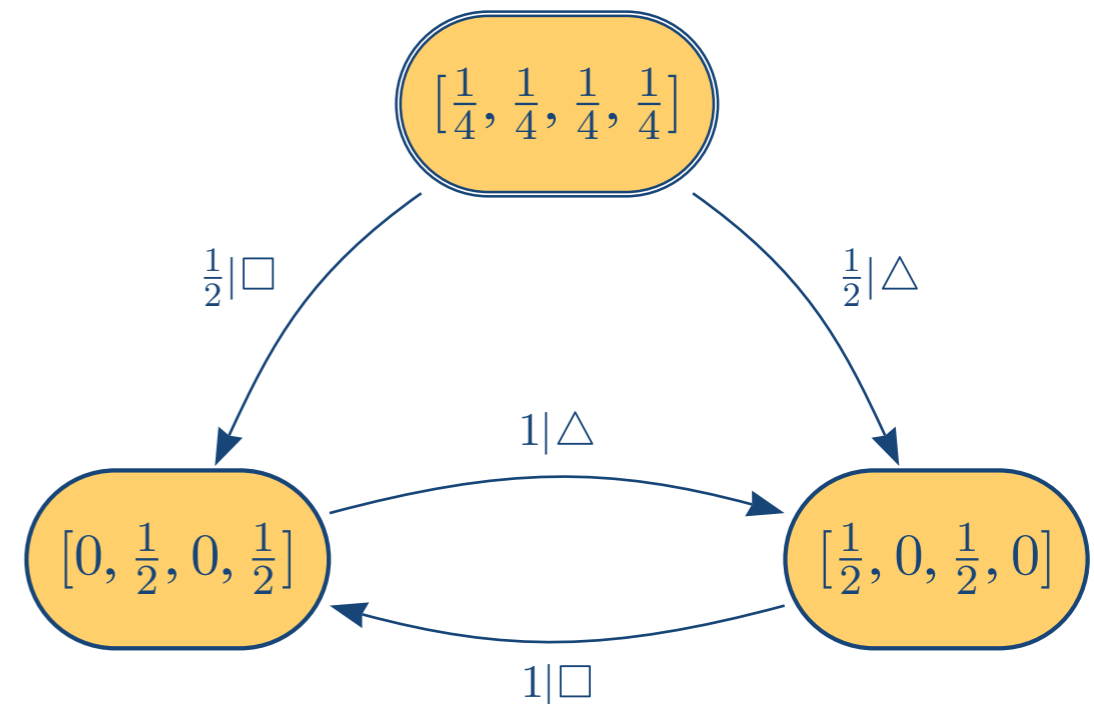
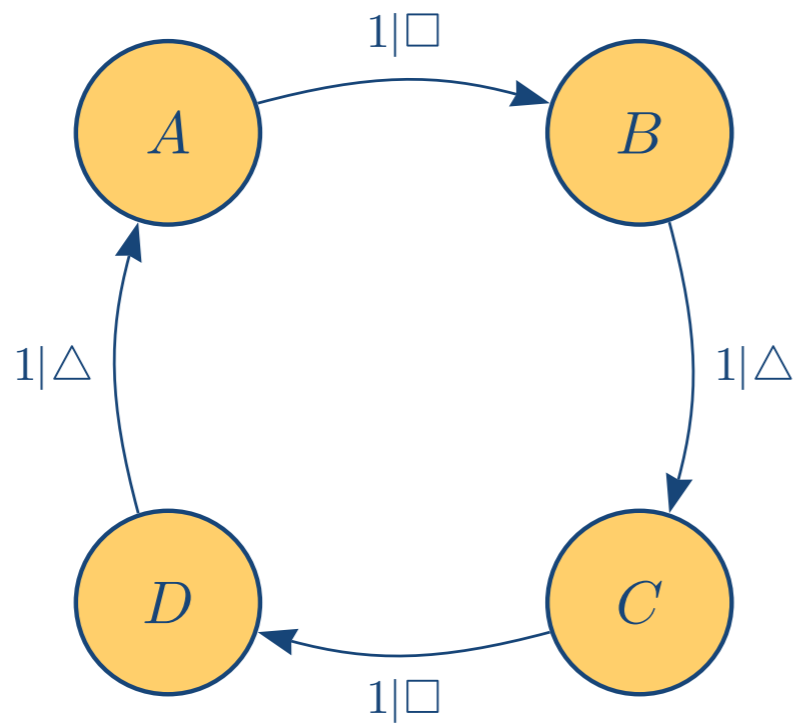
# States of States of Knowledge

Transient mixed states ...



# States of States of Knowledge

## Transient mixed states ...



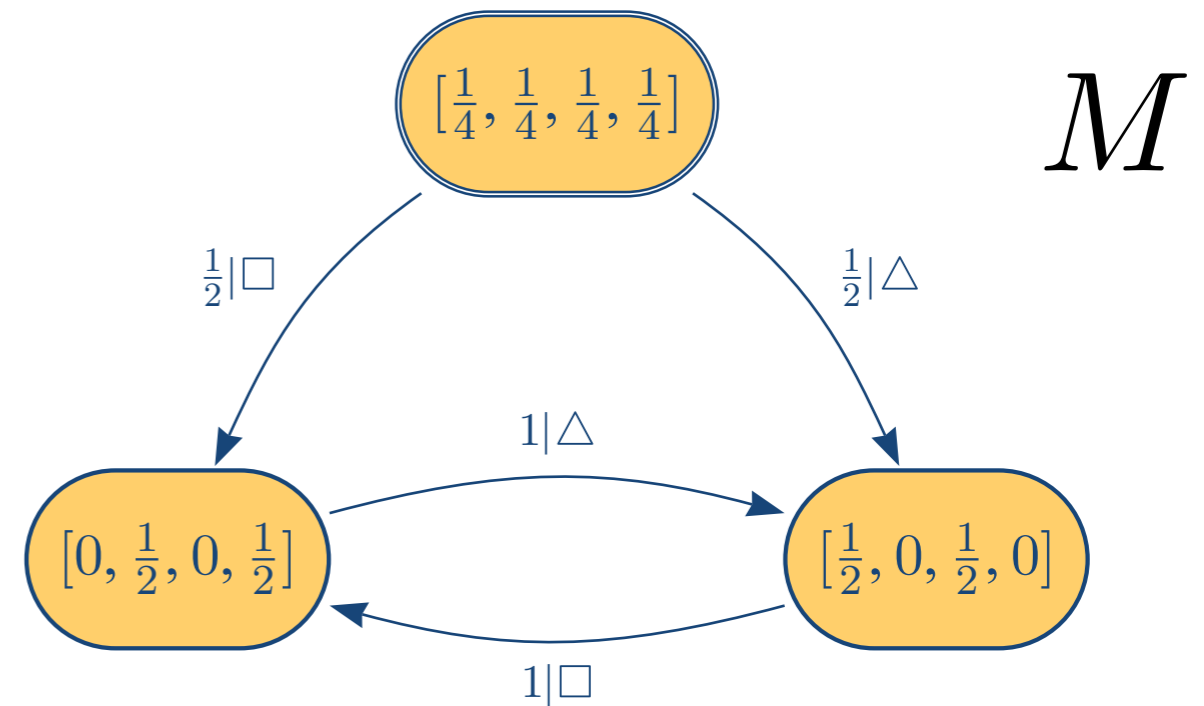
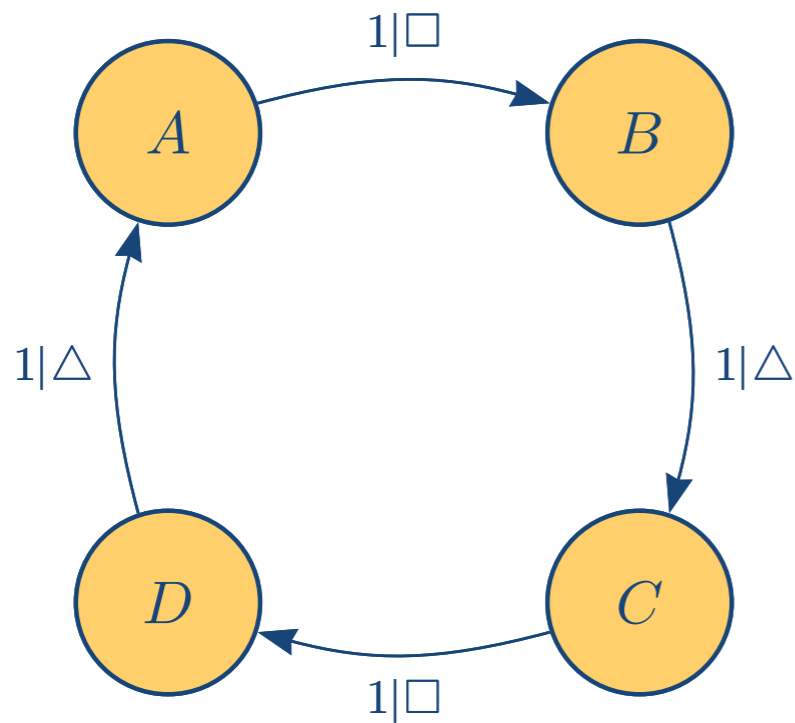
- **Transient mixed states:** States of uncertainty eventually never revisited.
- **Recurrent mixed states:** States of uncertainty repeatedly visited.
- MSP tells us we can never know in which *presentation state* we are.
- 4-state presentation is not “exactly” synchronizable (no finite word syncs).
- Exercise: Assume  $M$  is unifilar & sync’ble, show that once sync’d, always sync’d:

$$\exists w \in \mathcal{A}^*, H[\mu_{t+L}(w)] = 0$$

$$\Rightarrow \forall ww' \in \text{Allowed}, H[\mu_{t+L+L'}(ww')] = 0$$

# States of States of Knowledge

Transient mixed states ...



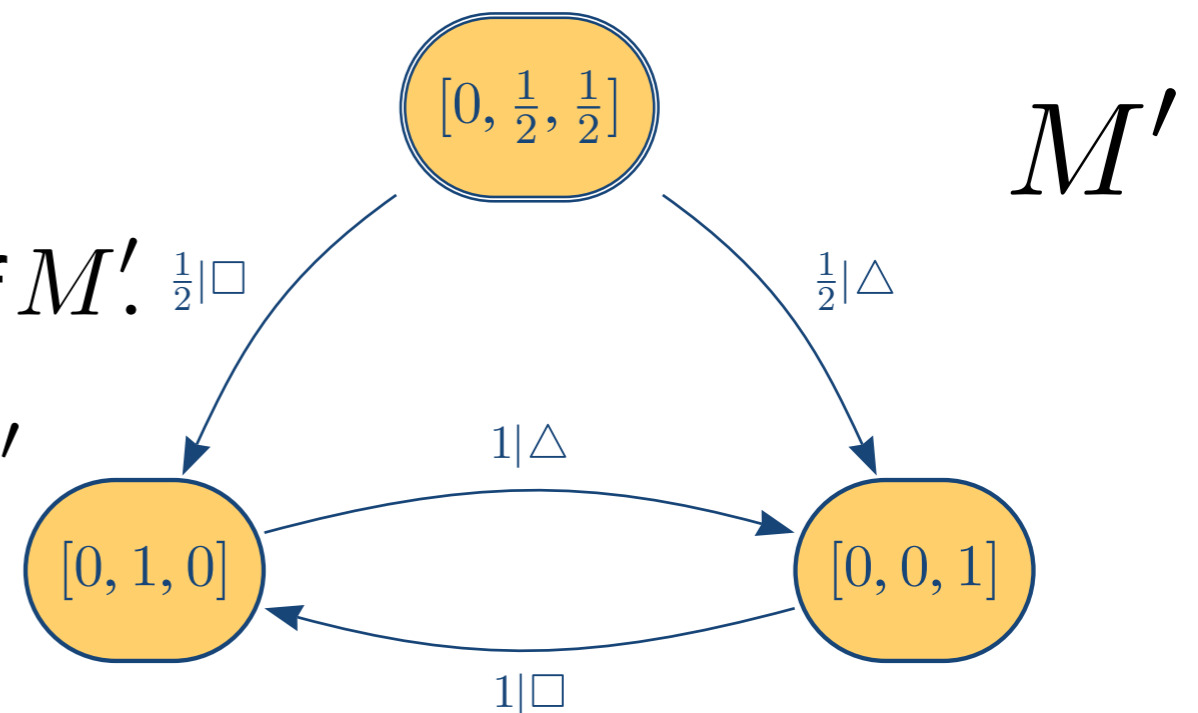
- Do it again:  
States of states of state uncertainty.

- Let  $M' = \mathcal{U}(M, \mu_t(\lambda))$ .

- Let  $\pi$  be the stationary distribution of  $M'$ .

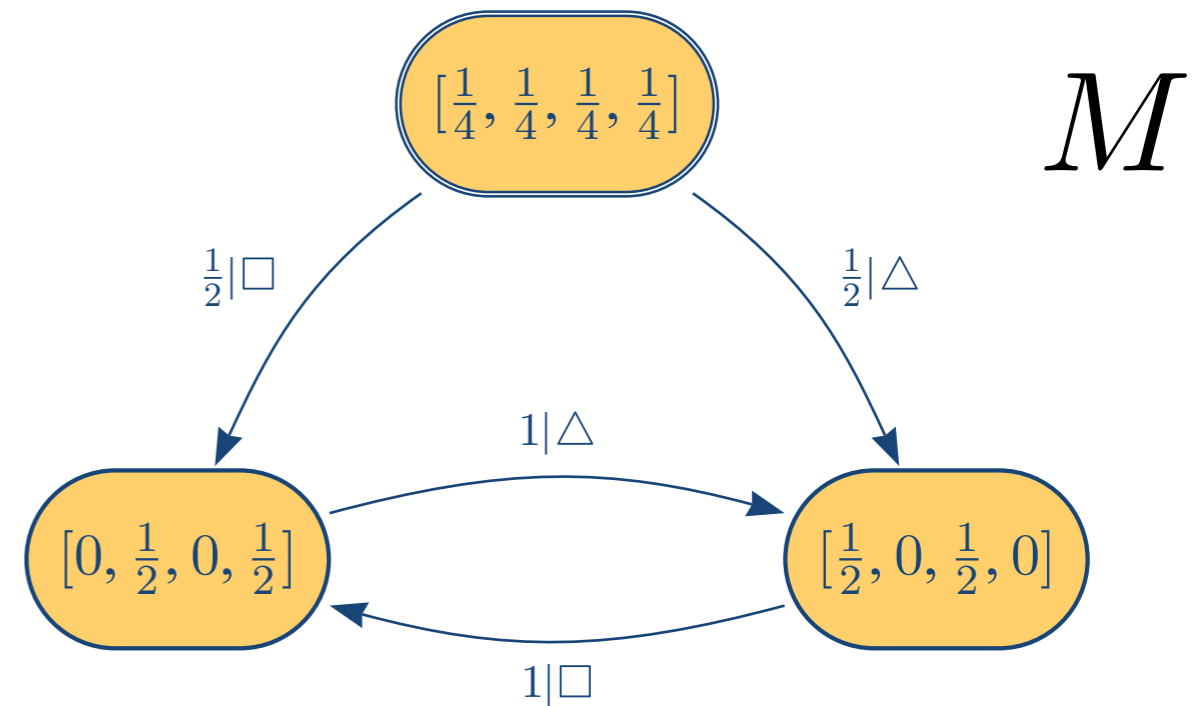
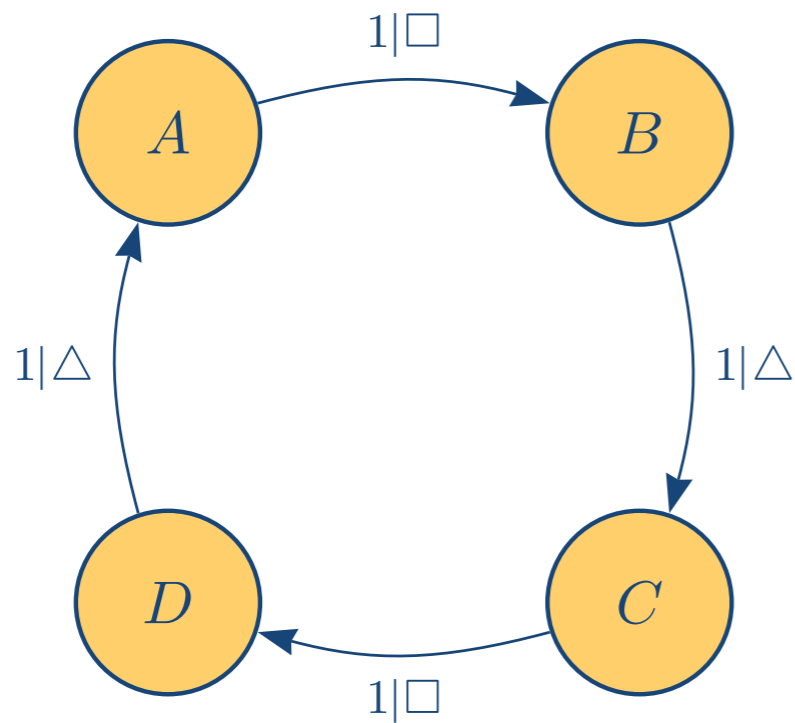
- Exercise: Prove

$$\pi = \mu_t(\lambda) \Rightarrow \mathcal{U}(M', \pi) = M'$$



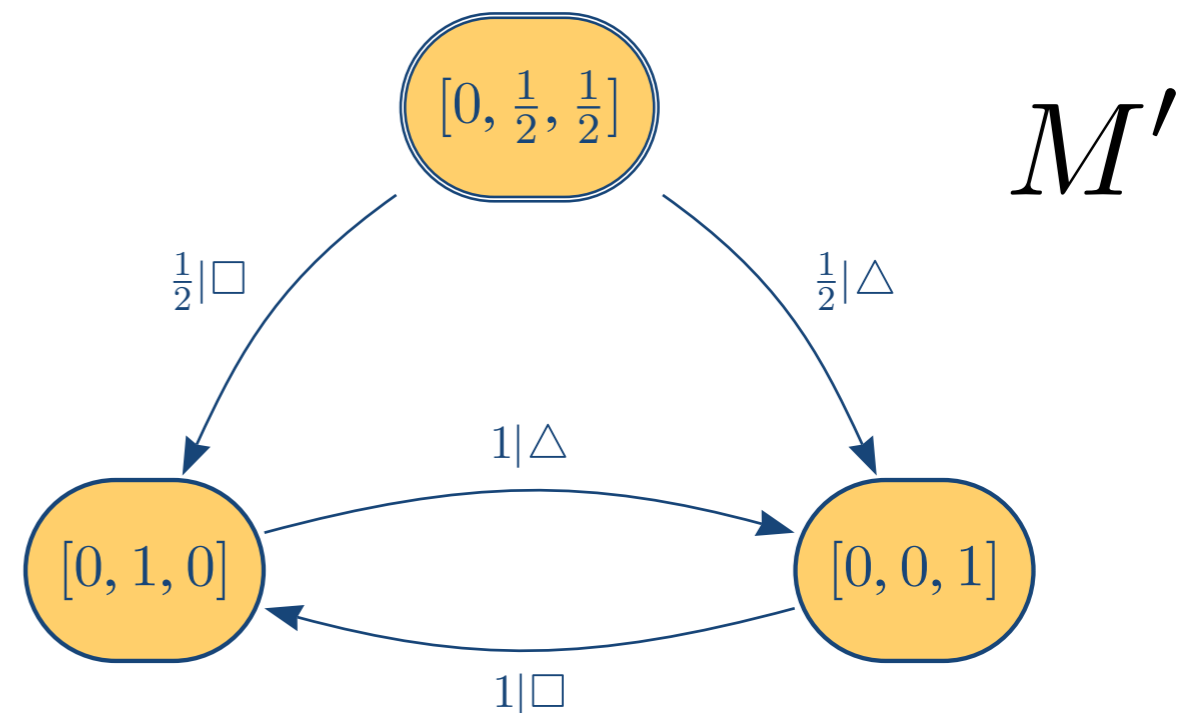
# States of States of Knowledge

Transient mixed states ...



- Meaning of  $[0, \frac{1}{2}, \frac{1}{2}]$ :

$$\begin{aligned} \frac{1}{2} [0, \frac{1}{2}, 0, \frac{1}{2}] + \frac{1}{2} [\frac{1}{2}, 0, \frac{1}{2}, 0] \\ = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}] \end{aligned}$$



# States of States of Knowledge

## Mixed State Presentation ...

### Relationship to $\epsilon$ -machine:

- In general, MSP is not the  $\epsilon$ -machine, but it is unifilar.
- MSP is a prescient rival.
- MSP history partition is a refinement of causal-state partition.
- Minimize MSP to get causal-state partition.
- MSP of causal-state-partition is  $\epsilon$ -machine.
- $\epsilon(w) = \mu_{t+L}(w)$  if  $\mathcal{U}(M)$  is the  $\epsilon$ -machine.
- However, mixed states contain additional information.

# States of States of Knowledge

## Mixed State Presentation ...

To construct the  $\epsilon$ -machine from any HMM  $M$ :

$$M \rightarrow \mathcal{U}(M) \rightarrow \min_{\text{unifilar}} (\mathcal{U}(M)) \rightarrow \epsilon\text{-machine}$$

# States of States of Knowledge

## Information Measures

### Block entropy

$$H(L) \equiv H[X_1, \dots, X_L] = - \sum_{x^L \in \mathcal{A}^L} \text{Pr}(x^L) \log_2 \text{Pr}(x^L)$$

### Entropy rate

$$h_\mu = \lim_{L \rightarrow \infty} [H(L+1) - H(L)]$$

### Excess Entropy

$$\mathbf{E} = \lim_{L \rightarrow \infty} [H(L) - h_\mu L]$$

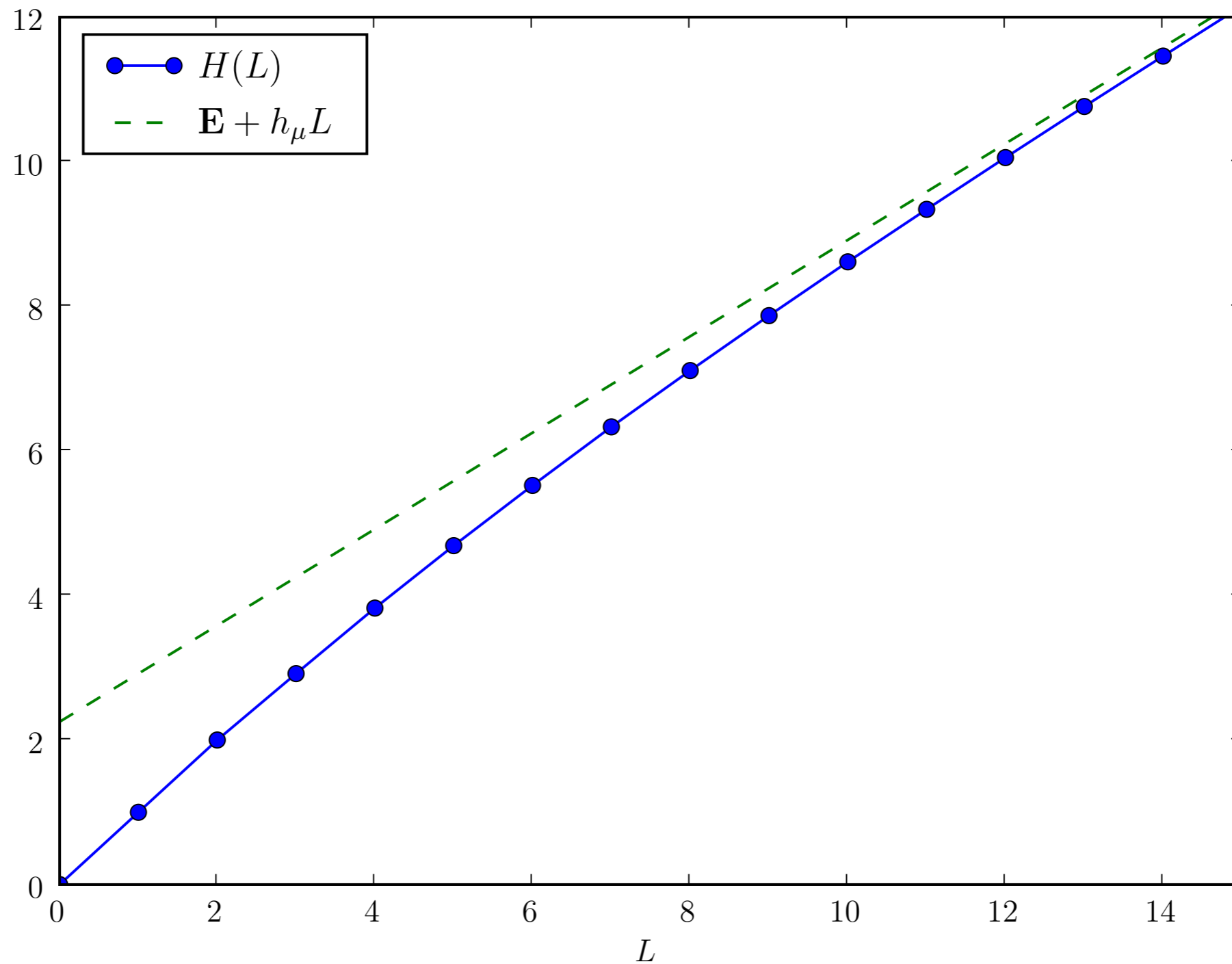
Block entropy is key in the study of many information measures.



# States of States of Knowledge

Information Measures ...

## RRXOR Process



# States of States of Knowledge

## Computational Complexity

### Word Probabilities:

$$\Pr(010) = \sum_{ijkl} \pi_i T_{ij}^0 T_{jk}^1 T_{kl}^0 \mathbf{1} \quad \text{Sum over all paths} \quad \Rightarrow \Pr(w) \sim O(|\mathcal{V}|^{|w|})$$
$$= \left( ((\pi T^0) T^1) T^0 \right) \mathbf{1} \quad \text{Update state distribution} \quad \Rightarrow \Pr(w) \sim O(|\mathcal{V}|^2 |w|)$$

### Block entropy:

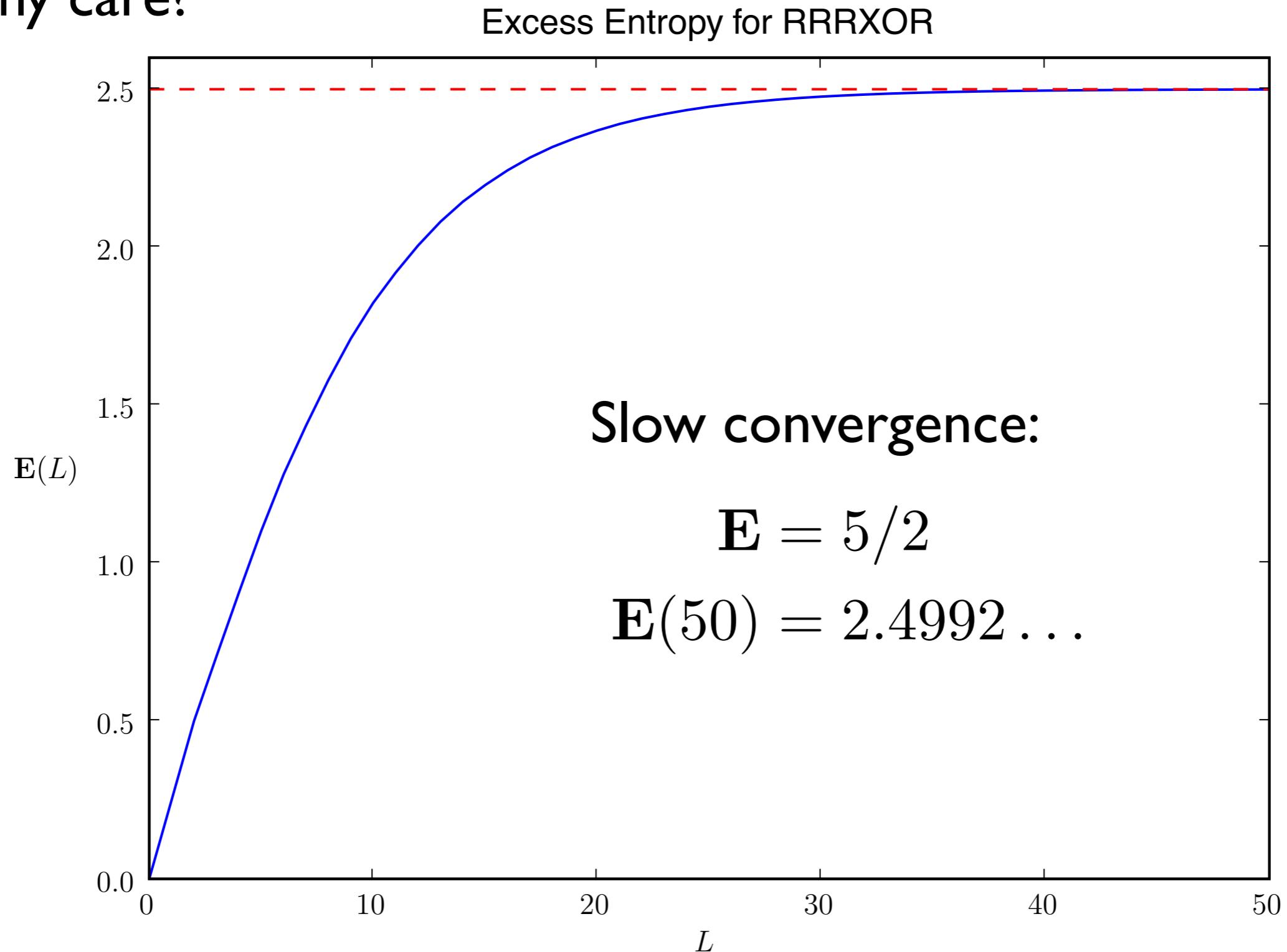
$$H(L) = - \sum_{x^L \in \mathcal{A}^L} \Pr(x^L) \log_2 \Pr(x^L) \quad \Rightarrow H(L) \sim O(|\mathcal{A}|^L |\mathcal{V}|^2 L)$$

Fundamentally, the problem is still exponential in  $L$ .

# States of States of Knowledge

## Information Measures ...

### Why care?



# States of States of Knowledge

## Block Entropy Chain Rule:

Implicit:

$$H(1) \equiv H[X_0]$$

$$H(2) \equiv H[X_0, X_1]$$

$$= H[X_1|X_0] + H[X_0]$$

$$= H[X_1|X_0] + H(1)$$

$$H(3) \equiv H[X_0, X_1, X_2]$$

$$= H[X_2|X_0, X_1] + H[X_0, X_1]$$

$$= H[X_2|X_0, X_1] + H(2)$$

...

$$H(L) \equiv H[X_0, \dots, X_{L-1}]$$

$$= H[X_{L-1}|X_0^{L-1}] + H(L-1)$$

# States of States of Knowledge

## Block Entropy Chain Rule:

Explicit:

$$H(1) \equiv H[X_0 | S_0 \sim \pi]$$

$$H(2) \equiv H[X_0, X_1 | S_0 \sim \pi]$$

$$= H[X_1 | X_0, S_0 \sim \pi] + H[X_0 | S_0 \sim \pi]$$

$$= H[X_1 | X_0, S_0 \sim \pi] + H(1)$$

$$H(3) \equiv H[X_0, X_1, X_2 | S_0 \sim \pi]$$

$$= H[X_2 | X_0, X_1, S_0 \sim \pi] + H[X_0, X_1 | S_0 \sim \pi]$$

$$= H[X_2 | X_0, X_1, S_0 \sim \pi] + H(2)$$

...

$$H(L) \equiv H[X_0, \dots, X_{L-1} | S_0 \sim \pi]$$

$$= H[X_{L-1} | X_0^{L-1}, S_0 \sim \pi] + H(L-1)$$

# States of States of Knowledge

## Efficient Block Entropy:

### Theorem:

$$H[X_{L-1}|X_0^{L-1}, S_0 \sim \pi] = H[X_{L-1}|(R_{L-1}|R_0 = \mu_0(\lambda))]$$

### Consequence:

$$H(L) = H[X_{L-1}|(R_{L-1}|R_0 = \mu_0(\lambda))] + H(L-1)$$

That is,  $H(L)$  can be calculated with single-step conditional block entropies using mixed states of  $\mathcal{U}(M)$ .

### Comments:

- Algorithm is linear in  $L$ .
- Distributions over states are a better representation.

# States of States of Knowledge

## Efficient Block Entropy ... useful results

Recall:  $\Pr(X_t = x, R_{t+1} = \eta | R_t = \mu_t(\lambda))$

$$\equiv \begin{cases} \mu_t(\lambda) T^x \mathbf{1} & \eta = \mu_{t+1}(x) \\ 0 & \eta \neq \mu_{t+1}(x) \end{cases}$$

# States of States of Knowledge

## Efficient Block Entropy ... useful results

$$\begin{aligned} \text{Recall: } \Pr(X_t = x, R_{t+1} = \eta | R_t = \mu_t(\lambda)) \\ \equiv \delta_{\eta, \mu_{t+1}(x)} \Pr(X_t = x | S_t \sim \mu_t(\lambda)) \end{aligned}$$



# States of States of Knowledge

## Efficient Block Entropy ... useful results

$$\begin{aligned} \text{Recall: } \Pr(X_t = x, R_{t+1} = \eta | R_t = \mu_t(\lambda)) \\ \equiv \delta_{\eta, \mu_{t+1}(x)} \Pr(X_t = x | S_t \sim \mu_t(\lambda)) \end{aligned}$$

**Result 1:**

$$\begin{aligned} \Pr(X_t = x | R_t = \mu_t(\lambda)) &= \sum_{\eta} \Pr(X_t = x, R_{t+1} = \eta | R_t = \mu_t(\lambda)) \\ &= \sum_{\eta} \delta_{\eta, \mu_{t+1}(x)} \Pr(X_t = x | S_t \sim \mu_t(\lambda)) \\ &= \Pr(X_t = x | S_t \sim \mu_t(\lambda)) \end{aligned}$$

**For example:**

$$\Pr(X_0 = x | R_0 = \pi) = \Pr(X_0 = x | S_0 \sim \pi)$$

# States of States of Knowledge

## Efficient Block Entropy ... useful results

$$\begin{aligned} \text{Recall: } \Pr(X_t = x, R_{t+1} = \eta | R_t = \mu_t(\lambda)) \\ \equiv \delta_{\eta, \mu_{t+1}(x)} \Pr(X_t = x | S_t \sim \mu_t(\lambda)) \end{aligned}$$

### Result 2:

$$\begin{aligned} \Pr(R_{t+1} = \eta | R_t = \mu_t(\lambda)) &= \sum_x \Pr(X_t = x, R_{t+1} = \eta | R_t = \mu_t(\lambda)) \\ &= \sum_x \delta_{\eta, \mu_{t+1}(x)} \Pr(X_t = x | S_t \sim \mu_t(\lambda)) \\ &= \sum_x \delta_{\eta, \mu_{t+1}(x)} \Pr(X_t = x | R_t = \mu_t(\lambda)) \end{aligned}$$

**Obvious: Symbols that transition to  $\eta$  contribute to  $\eta$ 's probability.**

# States of States of Knowledge

## Efficient Block Entropy ... useful results

Previous result generalizes:

$$\Pr(R_L = \eta | R_0 = \mu_0(\lambda)) = \sum_{w \in \mathcal{A}^L} \delta_{\eta, \mu_L(w)} \Pr(X_0^L = w | R_0 = \mu_0(\lambda))$$

Prove this holds for  $L = 2$ :

$$\begin{aligned} & \Pr(R_2 = \eta | R_0 = \mu_0(\lambda)) \\ &= \sum_{\psi} \Pr(R_1 = \psi, R_2 = \eta | R_0 = \mu_0(\lambda)) \\ &= \sum_{\psi} \Pr(R_2 = \eta | R_1 = \psi, R_0 = \mu_0(\lambda)) \Pr(R_1 = \psi | R_0 = \mu_0(\lambda)) \\ &= \dots \\ &= \sum_{w \in \mathcal{A}^2} \delta_{\eta, \mu_2(w)} \Pr(X_0^2 = w | R_0 = \mu_0(\lambda)) \end{aligned}$$

# States of States of Knowledge

## Efficient Block Entropy ... Proof

$$H[X_{L-1} | X_0^{L-1}, S_0 \sim \mu_0(\lambda)]$$

# States of States of Knowledge

Synchronization Information:

$$\mathbf{S} = \sum_{L=0}^{\infty} H[S_L | X_0^L] = \sum_{L=0}^{\infty} \sum_w \Pr(X_0^L = w) H[S_L | X_0^L = w]$$

In principle, easy to compute now:

- Iterate through all words, calculate:

$$\mu_L(w) \equiv \Pr(S_L | X_0^L)$$

$$H[\mu_L(w)] = H[S_L | X_0^L = w]$$

- Desire: Make use of MSP equivalence relation
- Question: What does  $\mathbf{S}$  look like on the MSP?

# States of States of Knowledge

## Synchronization Information ...

Focus on individual  $L$ -terms:

$$\begin{aligned} H[S_L | X_0^L] &= \sum_w \Pr(X_0^L = w | S_0 \sim \mu_0(\lambda)) H[S_L | X_0^L = w] \\ &= \sum_w \Pr(X_0^L = w | S_0 \sim \mu_0(\lambda)) H[\mu_L(w)] \\ &= \sum_w \sum_\eta \delta_{\eta, \mu_L(w)} \Pr(X_0^L = w | S_0 \sim \mu_0(\lambda)) H[\eta] \\ &= \sum_\eta H[\eta] \sum_w \delta_{\eta, \mu_L(w)} \Pr(X_0^L = w | R_0 = \mu_0(\lambda)) \\ &= \sum_\eta H[\eta] \Pr(R_L = \eta | R_0 = \mu_0(\lambda)) \end{aligned}$$

# States of States of Knowledge

## Synchronization Information ...

Result:

$$\mathbf{S} = \sum_{L=0}^{\infty} H[S_L | X_0^L] = \sum_{L=0}^{\infty} \sum_{\eta} \Pr(R_L = \eta | R_0 = \mu_0(\lambda)) H[\eta]$$

At each  $L$ :

- Compute mixed-state entropy
- Weight it by the probability of being in each allowed mixed state.

If mixed-state presentation's basis is recurrent  $\varepsilon\mathbf{M}$ , then the entropy of each recurrent mixed state will be zero.

Note, here work with probabilities over transient causal states.

Synchronization information can now be written for the MSP of the recurrent  $\varepsilon\mathbf{M}$ .

# States of States of Knowledge

## Summary

- Proper representation (mixed state presentation) gave insight.
- Block entropy calculations now scale linearly in  $L$ .
- Made use of conditional random variables.
- Made use of nonstationary word distributions.



# States of States of Knowledge

Reading for next lecture: CMR articles

*TBA*

*PRATISP*