States of States of Knowledge I

Reading for this lecture:
CMR articles *PRATISP, IACP, IACPLCOCS*

States of States of Knowledge

Causal states:

Conditions of knowledge that lead to optimal prediction

Today:

A hierarchy of states for optimal prediction

States of States of Knowledge

Agenda:

- Review
- Conditional Probabilities
- Mixed State Presentations
- Examples
- How to Calculate Mixed State Presentations
- Complexity Measures and Efficient Block Entropies
- Synchronization Information

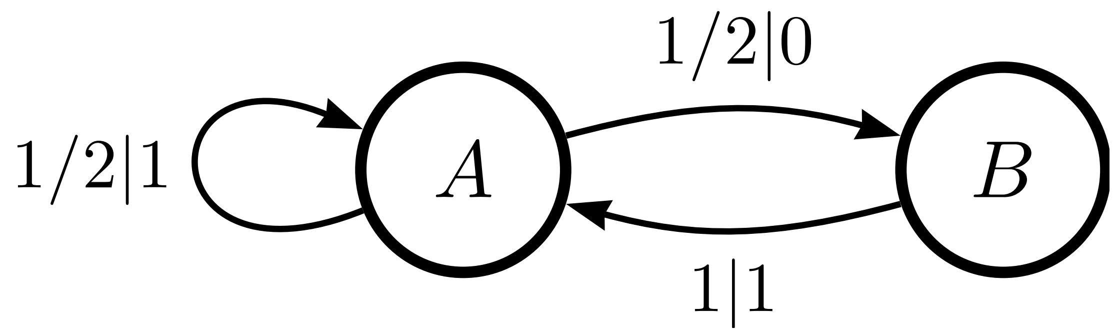
This lecture

Next

States of States of Knowledge

Review:

Golden Mean Process and its ε-Machine:



$$\mathcal{A} = \{0, 1\}$$
$$\mathcal{V} = \{A, B\}$$

$$T^{(0)} = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix} \qquad T^{(1)} = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix}$$

$$T \equiv T^{(0)} + T^{(1)}$$

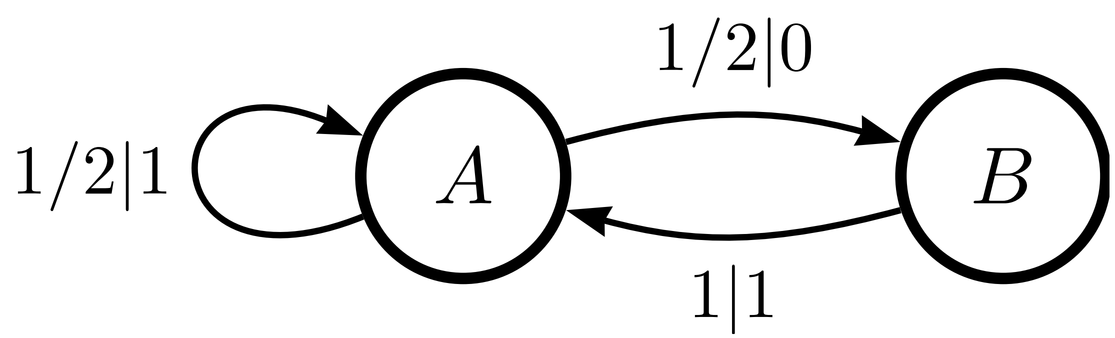
$$\pi = \pi T$$

$$\Rightarrow \pi = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

States of States of Knowledge

Review ...

Golden Mean Process and its ε-Machine:



Word distribution:

$$\Pr(w) = \pi T^{w} \mathbf{1}$$

$$\Pr(0) = \pi T^{0} \mathbf{1} = 1/3 \qquad \Pr(1) = \pi T^{1} \mathbf{1} = 2/3$$

$$\Pr(00) = \pi T^{0} T^{0} \mathbf{1} = 0 \qquad \Pr(10) = \pi T^{1} T^{0} \mathbf{1} = 1/3$$

$$\Pr(01) = \pi T^{0} T^{1} \mathbf{1} = 1/3 \qquad \Pr(11) = \pi T^{1} T^{1} \mathbf{1} = 1/3$$

States of States of Knowledge

Review ... Processes

Probability theory:

- Discrete alphabet $\mathcal{A}$
- Random variable X_t at time t
- Instance $x_t \in \mathcal{A}$ at time t

Random variable's probability mass function:

$$X_t \text{ is distributed as } \Pr(X_t = x)$$

$$X_t \sim \Pr(X_t = x)$$

Types of RVs:

- Quantitative: Age, voltage, ...
- Categorical: Names, colors, ...

Expectation value?

$$\text{red} \Pr(\text{red}) + \text{blue} \Pr(\text{blue})?$$

$$\text{fog} = \text{sunny} \Pr(\text{sunny}) + \text{rain} \Pr(\text{rain})?$$

States of States of Knowledge

Review ... Processes

Observed Process: $\Pr(\overleftrightarrow{X})$

Stationary: Probabilities time-independent

ε-Machines: Convenient representation

Typically, one of many possible presentations

Stationary probability of word:

$$\Pr(X^L = w) = \pi T^{(w)} \mathbf{1}$$

Observed process: Marginal distribution from machine "process":

$$\ldots (\sigma, x)_{-1} (\sigma, x)_0 (\sigma, x)_1 \ldots \qquad \sigma \in \boldsymbol{\mathcal{S}}, x \in \mathcal{A}$$

Implied: π is state distribution before w observed

Nonstationary state distributions? Sure thing!

More explicitly:

$$\Pr(X_t^L = w) = \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(X_t^L = w | S_t = \sigma) \Pr(S_t = \sigma)$$

States of States of Knowledge

Review ... Calculation?

State set: $\mathcal{V}$

Number of words grows $|\mathcal{A}|^L$

Efficient way to calculate word probabilities?

$$\Pr(X^L = w) = \pi T^{(w)} \mathbf{1}$$

Two ways:

$$\Pr(X^3 = abc) \equiv \pi T^a T^b T^c \mathbf{1} = \sum_{ijkl} \pi_i T^a_{ij} T^b_{jk} T^c_{kl} \quad \text{(1)}$$

$$= (((\pi T^a) T^b) T^c) \mathbf{1} \quad \text{(2)}$$

1. Sum probability of each path generating w:

   Time complexity: $O(|\mathcal{V}|^L)$

2. Track how path probabilities change as w is generated:

$$[\pi T^w]_\sigma = \Pr(X_t^L = w, S_{t+L} = \sigma)$$

   Time complexity: $O(|\mathcal{V}|^2 L)$

States of States of Knowledge

Conditional probability: Condition on an event

Conditional distribution is *still* a distribution

Conditional Random Variables:

Example: $\mathcal{A} = \{a, b\}$

$$Y_0 = \begin{cases} a & \text{if } X_0 = a \\ b & \text{if } X_0 = b \end{cases} \quad \Pr(Y_0 = y) \equiv \Pr(X_0 = y | S_0 = \sigma)$$

In words, Y_0 is a RV representing the distribution of RV X_0 conditioned on the event $S_0 = \sigma$.

Note: $H[Y_0] = H[X_0|S_0 = \sigma] \neq H[X_0|S_0]$

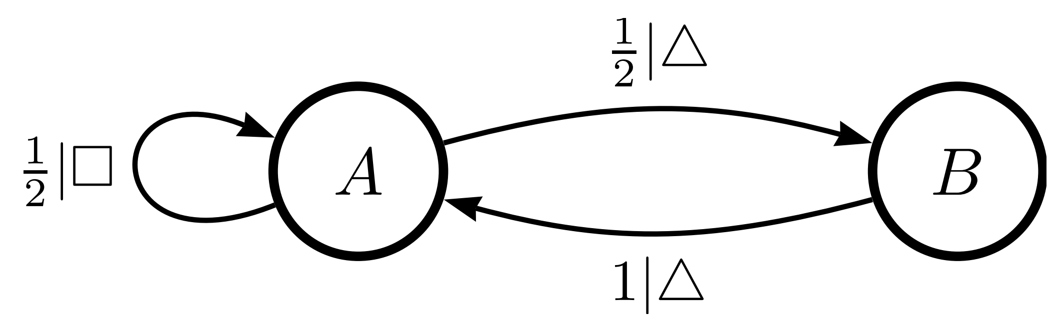
States of States of Knowledge

Conditional Random Variables ...

S_0 $\quad$ S_1 $\quad$ S_2 $\quad$ S_3 $\quad$ S_4

X_0 $\quad$ X_1 $\quad$ X_2 $\quad$ X_3

Example 1: $J_0 \sim \Pr(X_0 | S_0 = A)$ for the Even process.



Note: $\Pr(J_0 = \triangle) \neq \Pr(X_0 = \triangle)$

$$\Pr(J_0 = \triangle) = \Pr(X_0 = \triangle | S_0 = A) = \frac{1}{2}$$

$$\Pr(X_0 = \triangle) = \pi T^{\triangle} \mathbf{1} = \frac{2}{3}$$

States of States of Knowledge

Conditional Random Variables ...

$$\begin{array}{ccccccccc} S_0 & & S_1 & & S_2 & & S_3 & & S_4 \\ & X_0 & & X_1 & & X_2 & & X_3 & \end{array}$$

Example 2: $F_2 \sim \Pr(S_2 | S_0 = \sigma_0)$

$$\begin{aligned} \Pr(X_2 = x_2 | F_2 = \sigma_2) &= \Pr\big(X_2 = x_2 | \left(S_2 = \sigma_2 | S_0 = \sigma_0\right)\big) \\ &= \Pr(X_2 = x_2 | S_2 = \sigma_2, S_0 = \sigma_0) \\ &= \Pr(X_2 = x_2 | S_2 = \sigma_2) \end{aligned}$$

a condition of a condition is still a condition

State shielding

So, $H[X_2 | F_2 = \sigma_2] = H[X_2 | S_2 = \sigma_2]$

Conditional Random Variables ...

S_0 S_1 S_2 S_3 S_4

X_0 X_1 X_2 X_3

Example 2: $F_2 \sim \Pr(S_2 | S_0 = \sigma_0)$

$$\Pr(X_2 = x_2 | F_2 = \sigma_2) = \Pr(X_2 = x_2 | S_2 = \sigma_2)$$

$$H[X_2 | F_2 = \sigma_2] = H[X_2 | S_2 = \sigma_2]$$

But,

$$\begin{aligned} H[X_2 | F_2] &= \sum_{\sigma_2} \Pr(F_2 = \sigma_2) H[X_2 | F_2 = \sigma_2] \\ &= \sum_{\sigma_2} \Pr(S_2 = \sigma_2 | S_0 = \sigma_0) H[X_2 | S_2 = \sigma_2] \\ &\neq H[X_2 | S_2] \end{aligned}$$

States of States of Knowledge

Conditional Random Variables ...

$$S_0 \quad X_0 \quad S_1 \quad X_1 \quad S_2 \quad X_2 \quad S_3 \quad X_3 \quad S_4$$

Example 3: $G_2 \sim \Pr(S_2 | X_2 = x_2)$

$$\Pr(S_3 = \sigma_3 | G_2 = \sigma_2) = \Pr\big(S_3 = \sigma_3 | \left(S_2 = \sigma_2 | X_2 = x_2\right)\big)$$

$$= \Pr\left(S_3 = \sigma_3 | S_2 = \sigma_2, X_2 = x_2\right)$$

$$H[S_3 | G_2 = \sigma_2] = H[S_3 | S_2 = \sigma_2, X_2 = x_2] \quad (= 0 \text{ if unifilar})$$

$$H[S_3 | G_2] = \sum_{\sigma_2} \Pr(G_2 = \sigma_2) H[S_3 | G_2 = \sigma_2]$$

$$= \sum_{\sigma_2} \Pr(S_2 = \sigma_2 | X_2 = x_2) H[S_3 | S_2 = \sigma_2, X_2 = x_2]$$

$$\neq H[S_3 | S_2, X_2 = x_2]$$

States of States of Knowledge

Conditional Random Variables ... Summary

Nested conditions are important until everything is an event.

Define: $A' \sim \Pr(A|B = b)$

$$\begin{aligned} H[C|A' = a] &= H[C|(A = a|B = b)] \\ &= H[C|A = a, B = b] \end{aligned}$$

$$\begin{aligned} H[C|A'] &= H[C|(A|B = b)] \\ &= \textstyle\sum_a \Pr(A = a|B = b) H[C|(A = a|B = b)] \\ &= \textstyle\sum_a \Pr(A = a|B = b) H[C|A = a, B = b] \end{aligned}$$

$$H[C|A, B = b] = \textstyle\sum_a \Pr(A = a, B = b) H[C|A = a, B = b]$$

Conditional Random Variables ...

Recall: $$\Pr(X_t^L = w) = \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(X_t^L = w | S_t = \sigma) \Pr(S_t = \sigma)$$

- Previously: Conditioned on events.
- Now, condition on random variable (~ state distribution).
- Must make dependence on S_t explicit:
  - RHS looks like an expectation value.
  - Think of each $\Pr(X_t^L = w | S_t = \sigma)$ as instance of a RV.

States of States of Knowledge

Conditional Random Variables ...

Recall: $\Pr(X_t^L = w) = \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(X_t^L = w | S_t = \sigma) \Pr(S_t = \sigma)$

Define a quantitative random variable:

$$Z \equiv \Pr(X_t^L = w | S_t = \sigma)$$

$$= \begin{cases} \Pr(X_t^L = w | S_t = A) & \text{if } S_t = A \\ \Pr(X_t^L = w | S_t = B) & \text{if } S_t = B \end{cases}$$

$$\Pr(Z = z) = \Pr\left(Z = \Pr(X_t^L = w | S_t = \sigma)\right)$$

$$\equiv \Pr(S_t = \sigma)$$

Then, $\langle Z \rangle = \sum_z z \Pr(Z = z)$

$$= \sum_\sigma \Pr(X_t^L = w | S_t = \sigma) \Pr(S_t = \sigma)$$

States of States of Knowledge

Conditional Random Variables ...

Recall: $\Pr(X_t^L = w) = \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(X_t^L = w | S_t = \sigma) \Pr(S_t = \sigma)$

Thinking about Z is tedious.

Adopt a shorthand.

$\langle Z \rangle$

Definition:

$$\Pr(X_t^L = w | S_t) \equiv \Big\langle \Pr(X_t^L = w | S_t = \sigma) \Big\rangle$$

$$= \sum_{\sigma} \Pr(X_t^L = w | S_t = \sigma) \Pr(S_t = \sigma)$$

Conditional Random Variables ...

Conflate RV with vector of probabilities:

Example: Word probabilities from stationary distribution:

$$\Pr(X_t^L = w | S_t \sim \pi) = \pi T^{(w)} \mathbf{1}$$

Example: Word probabilities from uniform distribution:

$$U_t \sim \Pr(U_t = \sigma) \equiv 1/|\mathcal{V}|$$

$$\Pr(X_t^L = w | U_t) = \frac{1}{|\mathcal{V}|} \sum_\sigma \Pr(X_t^L = w | U_t = \sigma)$$

Example: Word probabilities from arbitrary vector μ:

$$\Pr\left(X_t^L = w | S_t \sim \mu\right) = \mu T^{(w)} \mathbf{1}$$

States of States of Knowledge

Conditioning on Random Variables

Working with probabilities, either is fine, unambiguous:

$$\Pr(X_1 = x | S_1)$$

$$\Pr(X_1 = x | S_1 \sim \mu)$$

Working with distributions, either is fine, unambiguous:

$$\Pr(X_1 | S_1)$$

$$\Pr(X_1 | S_1 \sim \mu)$$

States of States of Knowledge

Conditioning on Random Variables ...

Working with functionals: $H[X_1|S_1]$ is ambiguous in new notation.

$$H[X_1|S_1] \stackrel{?}{=} \sum_\sigma \Pr(S_1 = \sigma) H[X_1|S_1 = \sigma)$$

Average entropy of X_1 conditioned on events of S_1

$$= - \sum_\sigma \Pr(\sigma) \sum_x \Pr(x|\sigma) \log_2 \Pr(x|\sigma)$$

$$= - \sum_x \sum_\sigma \Pr(\sigma, x) \log_2 \Pr(x|\sigma)$$

$$H[X_1|S_1] \stackrel{?}{=} - \sum_x \Pr(X_1 = x|S_1) \log_2 \Pr(X_1 = x|S_1)$$

Entropy of X_1 for various S_1 distributions

$$= - \sum_x \sum_\sigma \Pr(x|\sigma) \Pr(\sigma) \log_2 \sum_\sigma \Pr(x|\sigma) \Pr(\sigma)$$

$$= - \sum_x \sum_\sigma \Pr(\sigma, x) \log_2 \sum_\sigma \Pr(\sigma, x)$$

States of States of Knowledge

Conditioning on Random Variables ...

Working with functionals: $H[X_1|S_1]$ is ambiguous in new notation:

$H[X_1|S_1]$ for average entropy of X_1 conditioned on events of S_1.

$H[X_1|S_1 \sim \mu]$ for entropy of X_1 for various S_1 distributions.

Old notation: Latter would be $H[X_1]$, corresponding to the viewpoint that all we care about is the observed process.

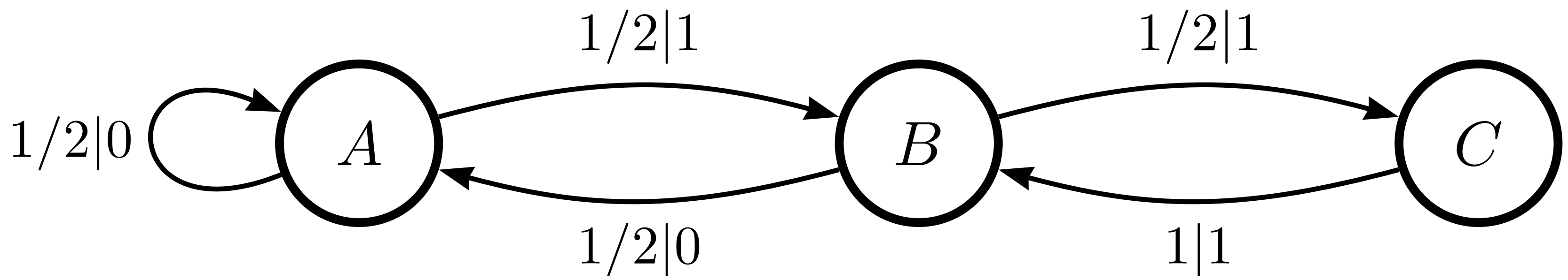
New notation: Can ask about multiple observed processes (from machine perspective), some stationary and some nonstationary.

If we use $\mu = \pi$, then we have the observed process.

States of States of Knowledge

Conditioning on Random Variables ...

Example: Odd Process



$$\pi = \begin{bmatrix} 2/5 & 2/5 & 1/5 \end{bmatrix}$$

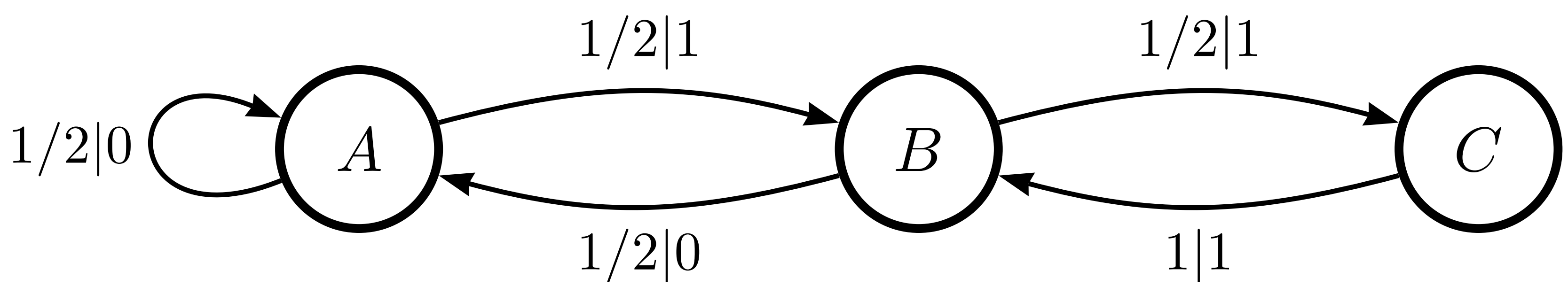
$$\Pr(X_3 = 0 | S_3 = A, S_0 \sim \pi) = \sum_\sigma \Pr(X_3 = 0 | S_3 = A, S_0 = \sigma) \Pr(S_0 = \sigma)$$

states are shielding $\rightarrow$
$$= \sum_\sigma \Pr(X_3 = 0 | S_3 = A) \Pr(S_0 = \sigma)$$
$$= \Pr(X_3 = 0 | S_3 = A)$$
$$= 1/2$$

States of States of Knowledge

Conditioning on Random Variables ...

Example: Odd Process



Take: $\mu = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$

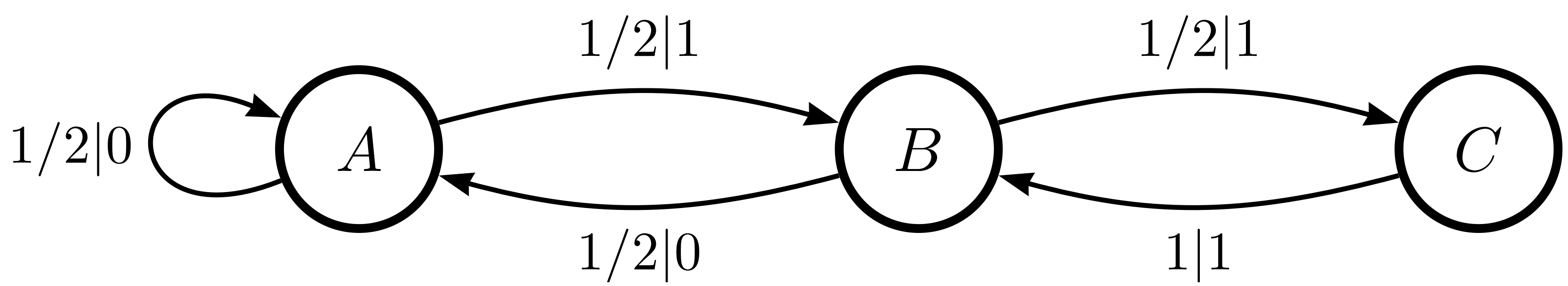
Recall:
$T = T^{(0)} + T^{(1)}$

$$\begin{aligned} \Pr(X_3 = x | S_1 \sim \mu) &= \mu T T T^{(x)} \mathbf{1} \\ &= \Pr(X_3 = x | S_2 \sim \mu T) \\ &= \Pr(X_3 = x | S_3 \sim \mu T T) \end{aligned}$$

States of States of Knowledge

Conditioning on Random Variables ...

Example: Odd Process



Take: $\mu = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$

$$\begin{aligned} \Pr(X_3 = x | S_1 \sim \mu) &= \mu T T T^{(x)} \mathbf{1} \\ &= \Pr(X_3 = x | S_2 \sim \mu T) \\ &= \Pr(X_3 = x | S_3 \sim \mu T T) \\ \mu \neq \pi : \text{ shift only } X_t \text{ is nonstationary} \rightarrow &\neq \Pr(X_4 = x | S_1 \sim \mu) \end{aligned}$$

States of States of Knowledge

Cautions!

Shorthands:

$$\Pr(X_t = x|S_t) \leftrightarrow \{\Pr(X_t = x|S_t = \sigma) : \sigma \in \mathcal{V}\}$$

$$\Pr(X_t|S_t) \leftrightarrow \{\Pr(X_t = x|S_t = \sigma) : \sigma \in \mathcal{V}, x \in \mathcal{A}\}$$

So far, we defined:

$$\Pr(X_t = x|S_t) = \langle \Pr(X_t = x|S_t = \sigma) \rangle \quad \text{Expected probability}$$

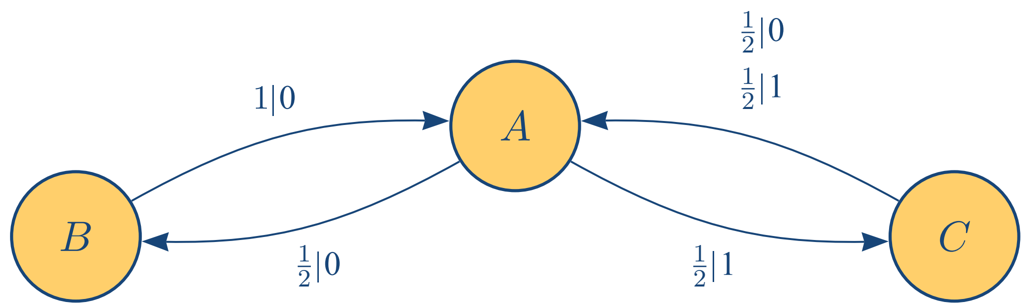
$$\Pr(X_t|S_t) = \langle \Pr(X_t|S_t = \sigma) \rangle \quad \text{Expected distribution}$$

Context determines whether we refer to:

- "Set of conditional probabilities" versus "expected probability"
- "Set of conditional probabilities" versus "expected distribution"

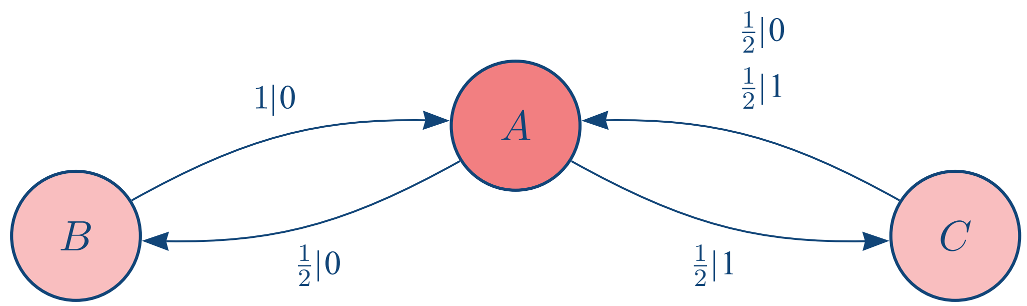
States of States of Knowledge

Mixed State Presentations: Evolving state distributions

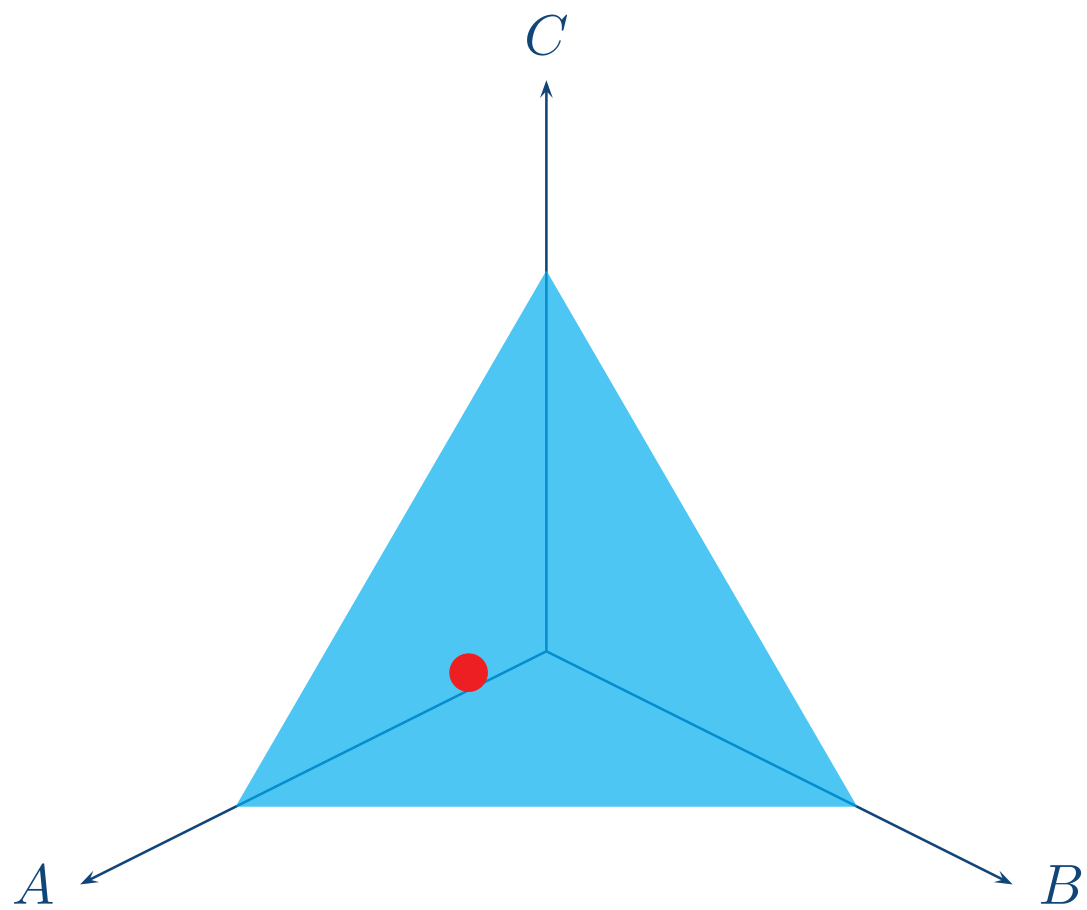


States of States of Knowledge

Mixed State Presentations ...

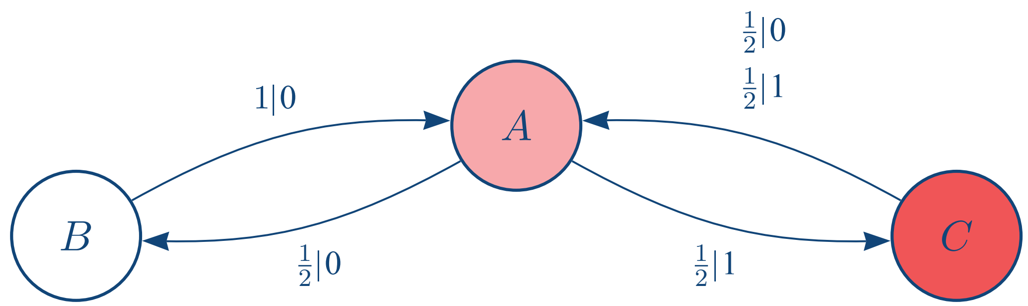


$$\mu = \begin{bmatrix} 1/2 & 1/4 & 1/4 \end{bmatrix}$$
$$w = \lambda$$

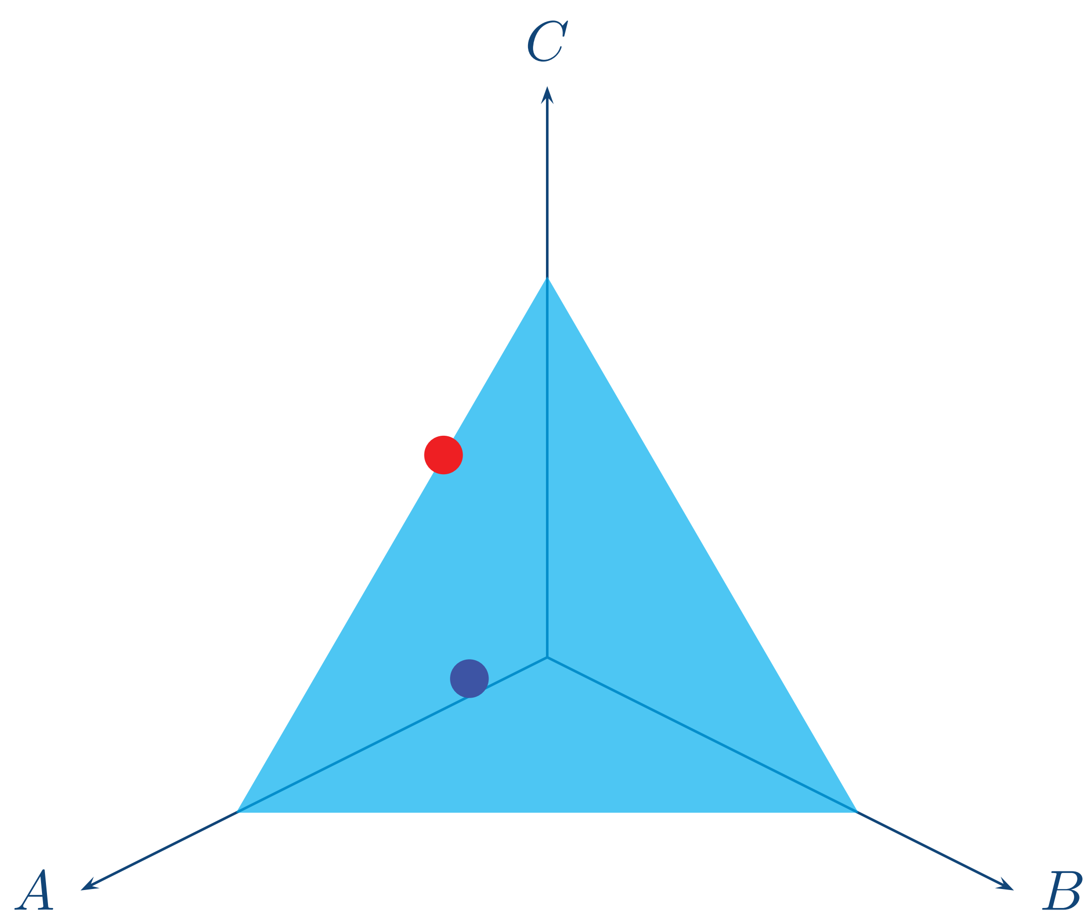


States of States of Knowledge

Mixed State Presentations ...

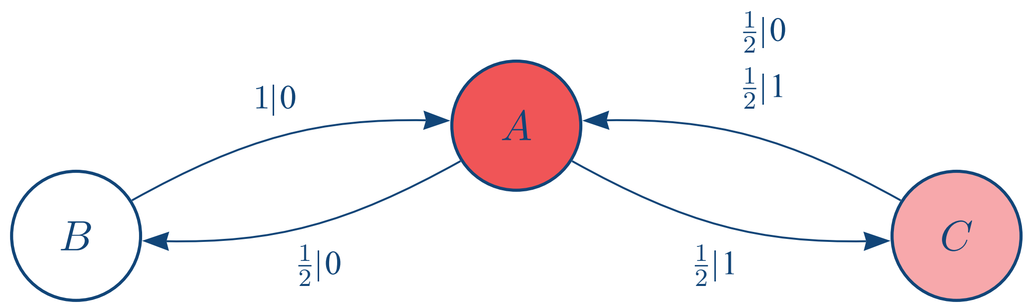


$$\mu = \begin{bmatrix} 1/3 & 0 & 2/3 \end{bmatrix}$$
$$w = 1$$

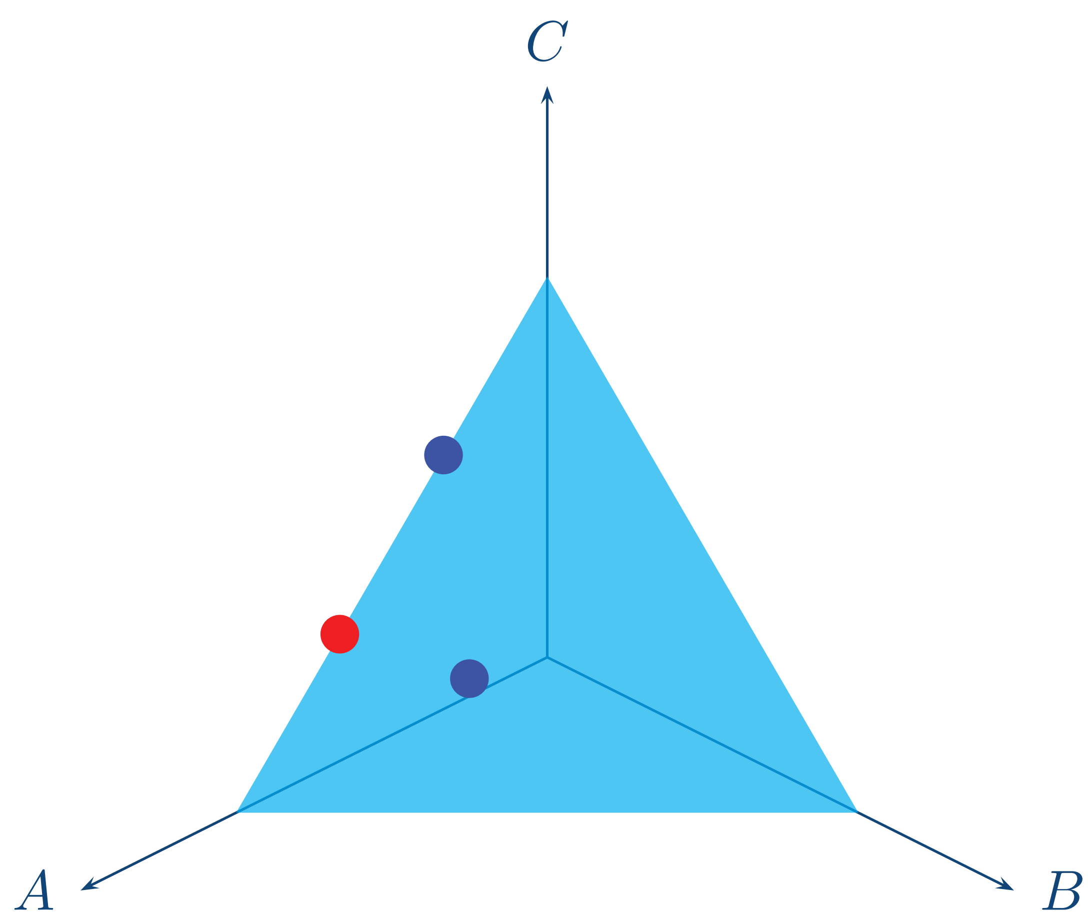


States of States of Knowledge

Mixed State Presentations ...

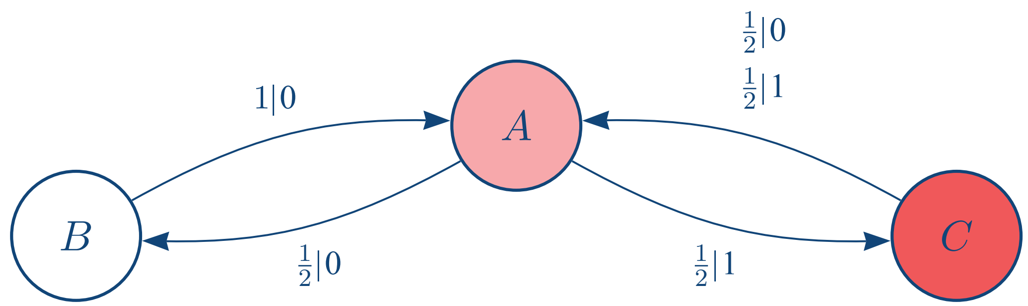


$\mu = \begin{bmatrix} 2/3 & 0 & 1/3 \end{bmatrix}$
$w = 11$

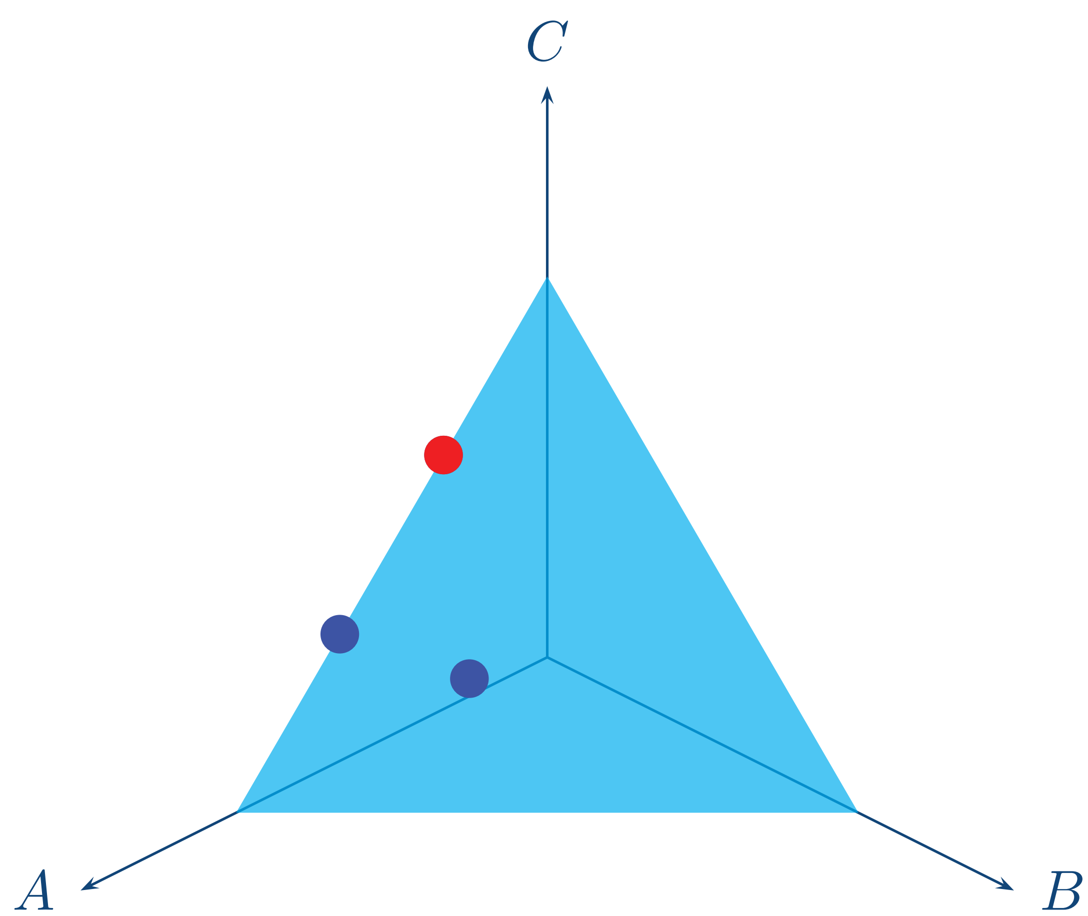


States of States of Knowledge

Mixed State Presentations ...

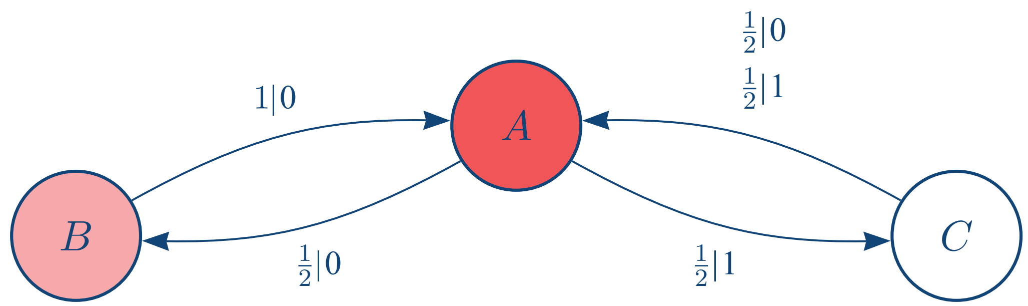


$$\mu = \begin{bmatrix} 1/3 & 0 & 2/3 \end{bmatrix}$$
$$w = 111$$

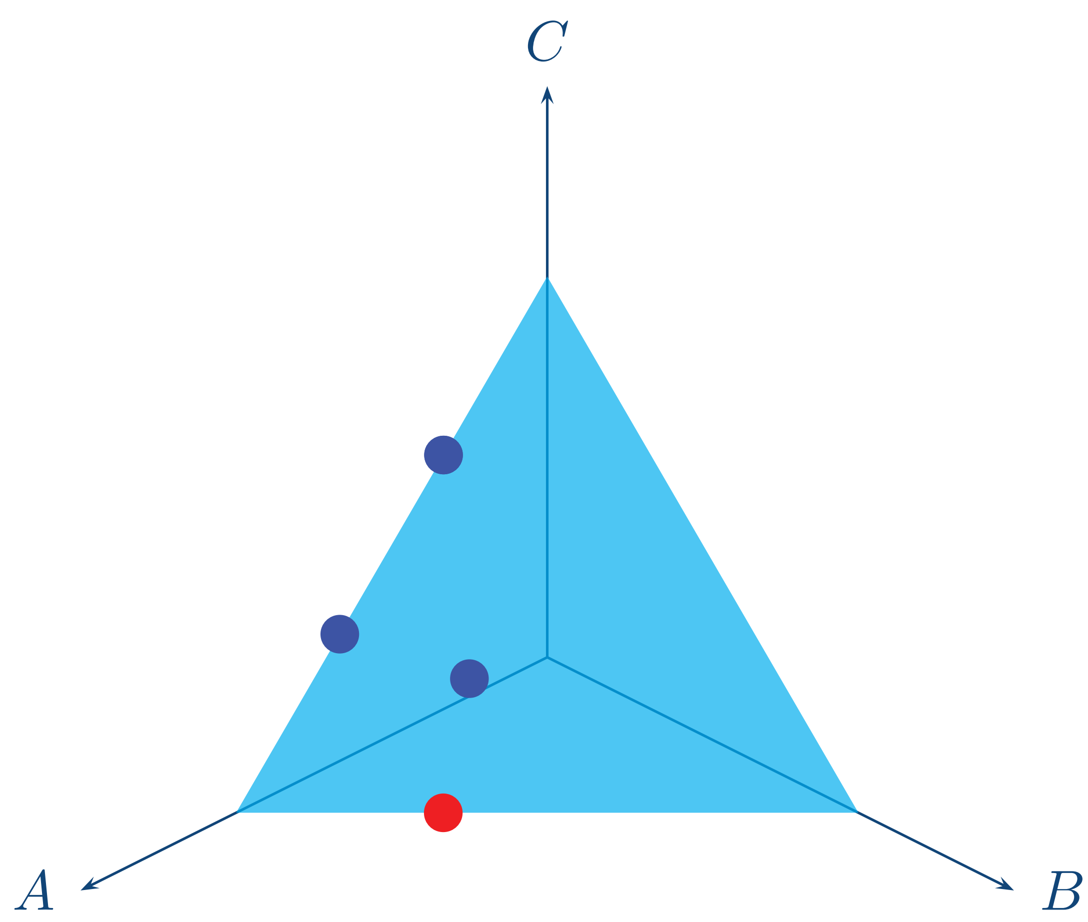


States of States of Knowledge

Mixed State Presentations ...

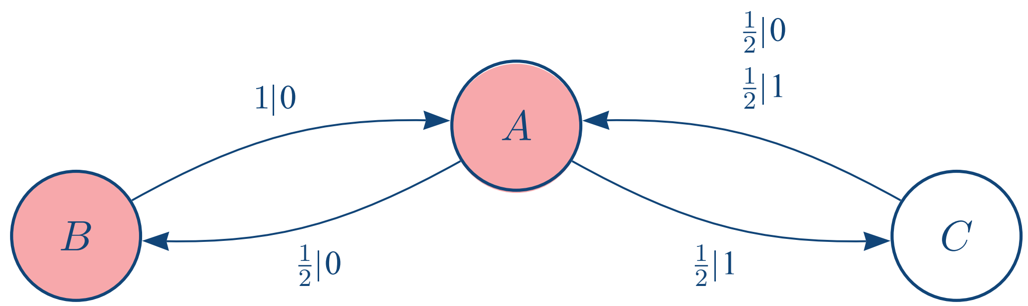


$$\mu = \begin{bmatrix} 2/3 & 1/3 & 0 \end{bmatrix}$$
$$w = 1110$$

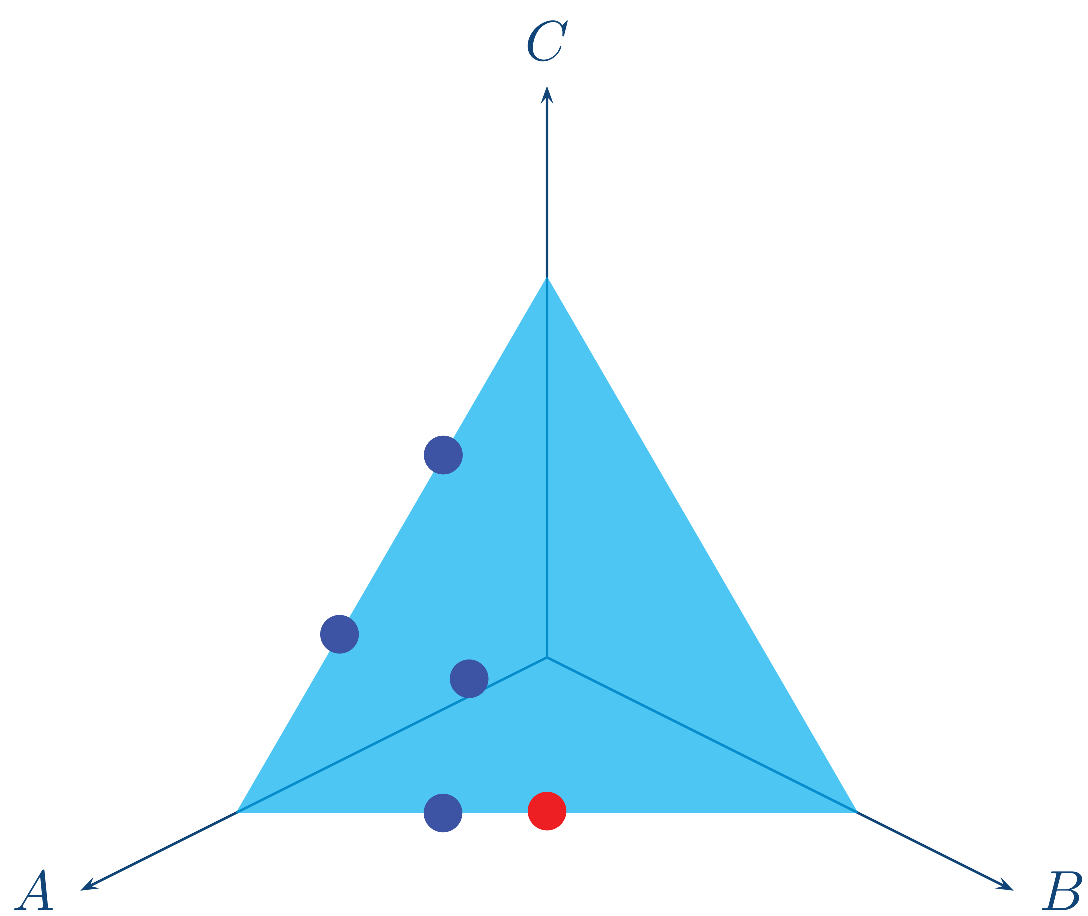


States of States of Knowledge

Mixed State Presentations ...

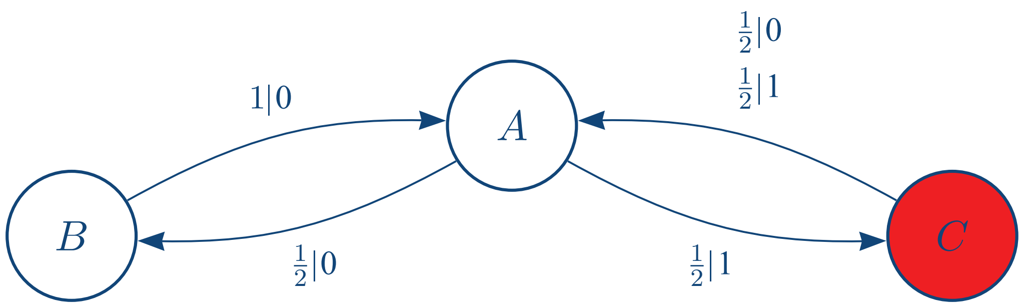


$$\mu = \begin{bmatrix} 1/2 & 1/2 & 0 \end{bmatrix}$$
$$w = 11100$$



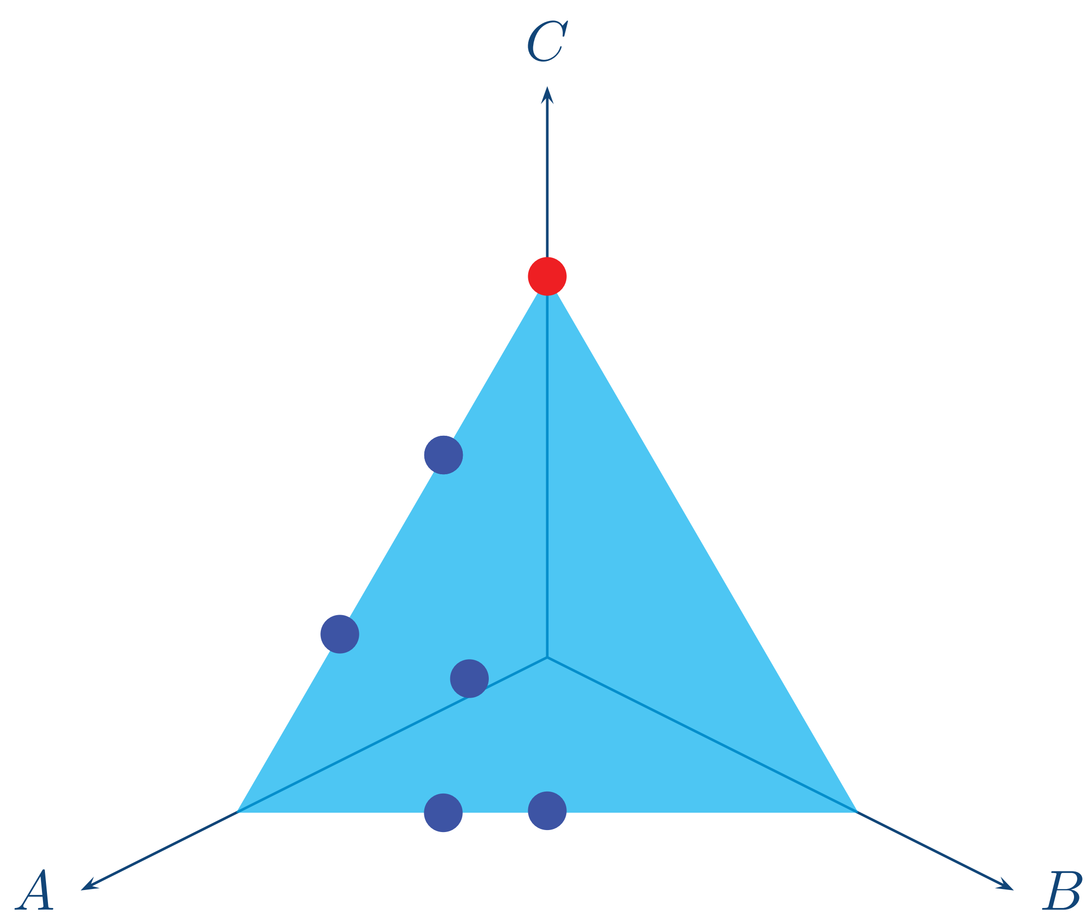
States of States of Knowledge

Mixed State Presentations ...



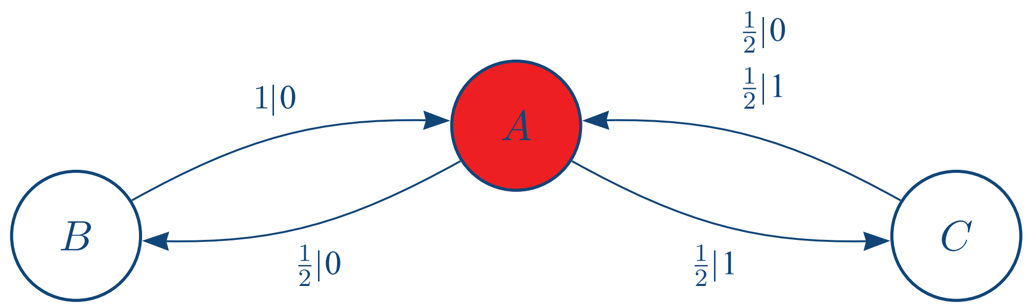
$\mu = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$

$w = 111001$

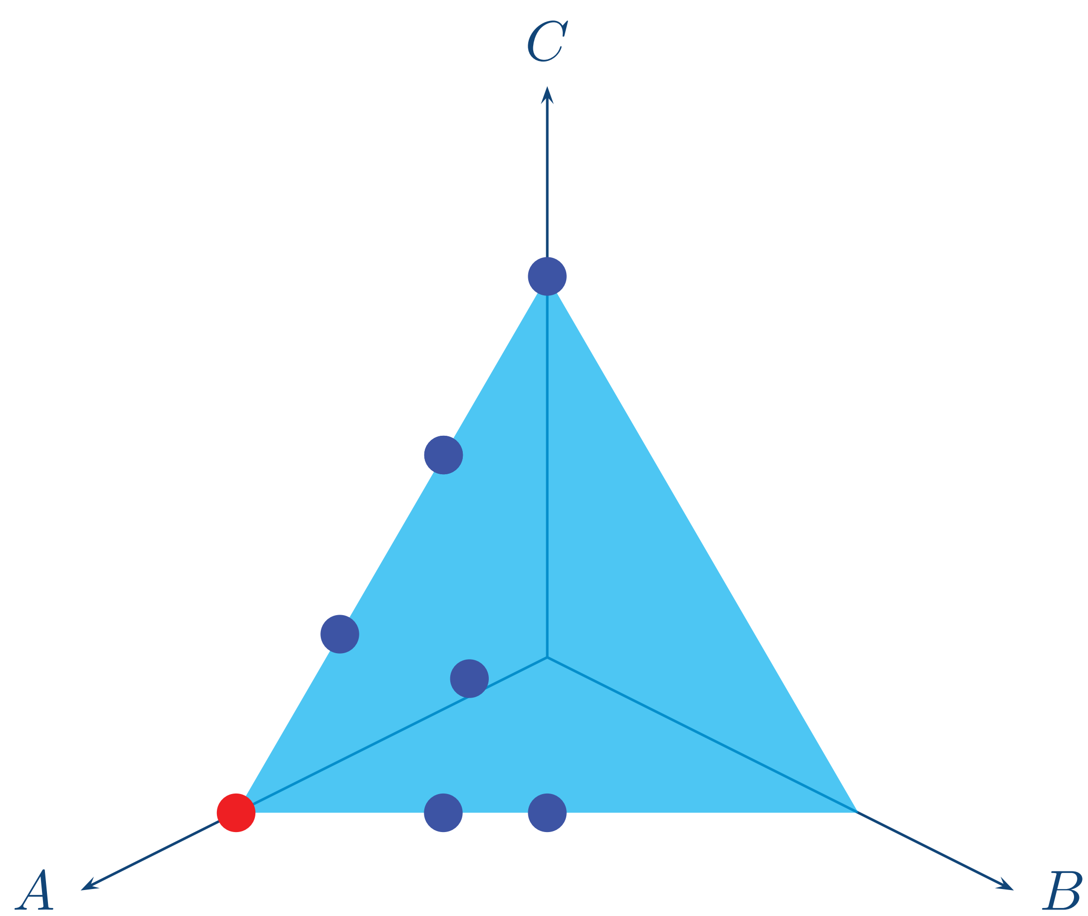


States of States of Knowledge

Mixed State Presentations ...

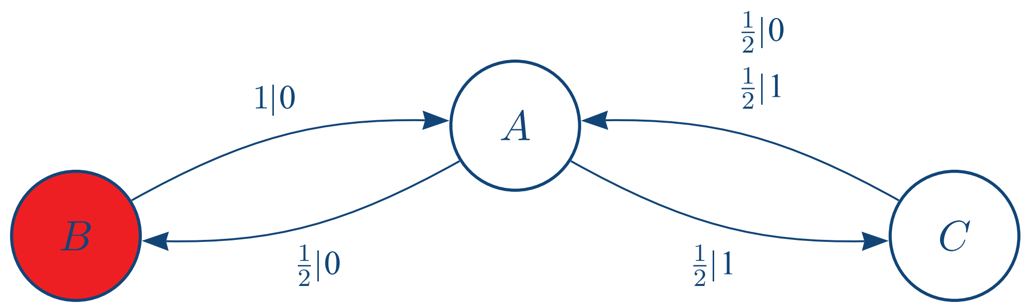


$$\mu = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$
$$w = 1110011$$

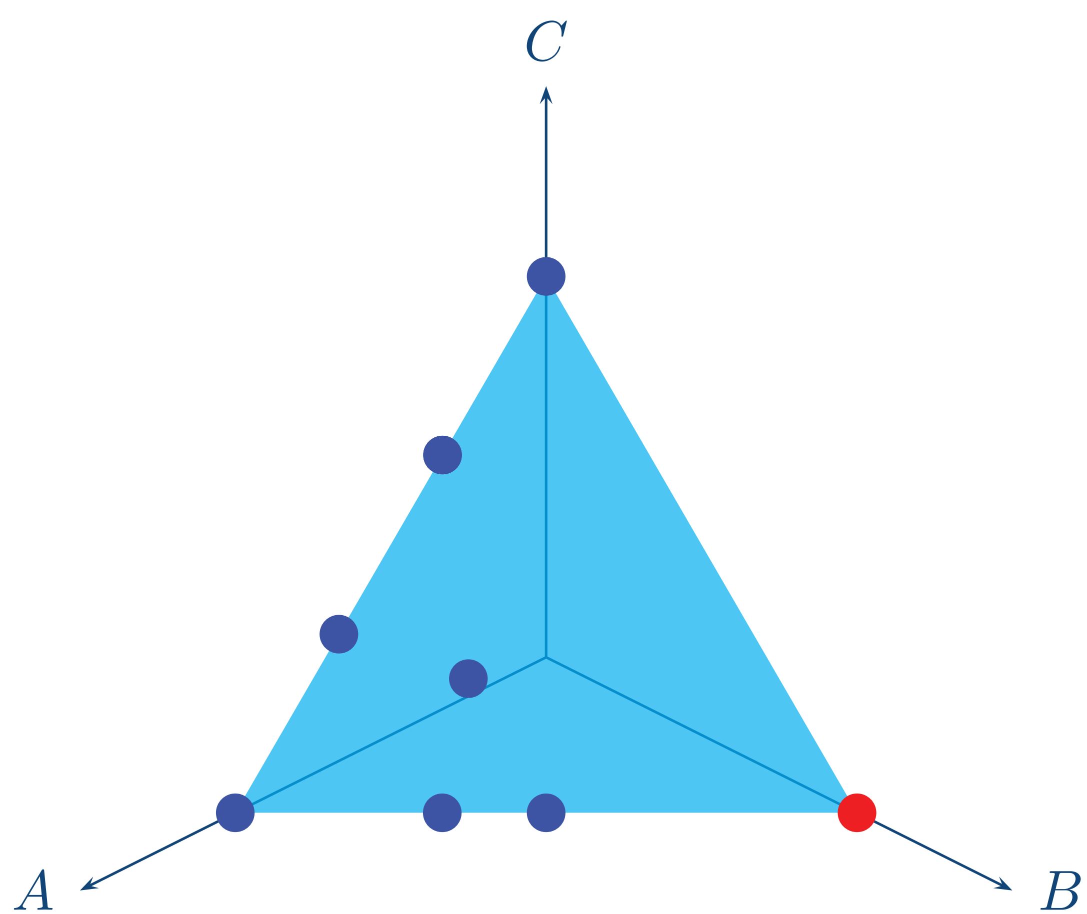


States of States of Knowledge

Mixed State Presentations ...

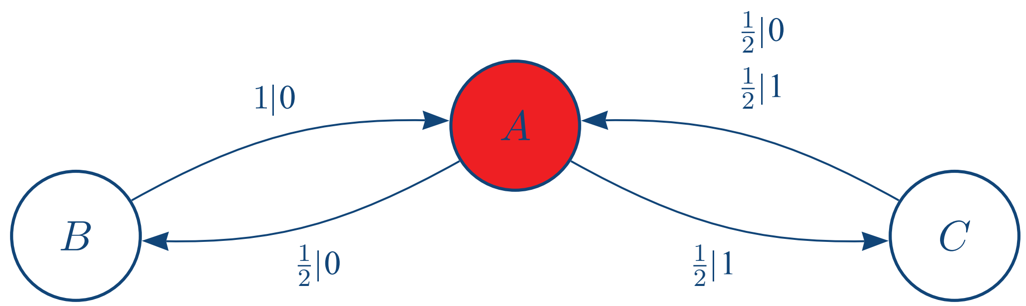


$$\mu = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$
$$w = 11100110$$

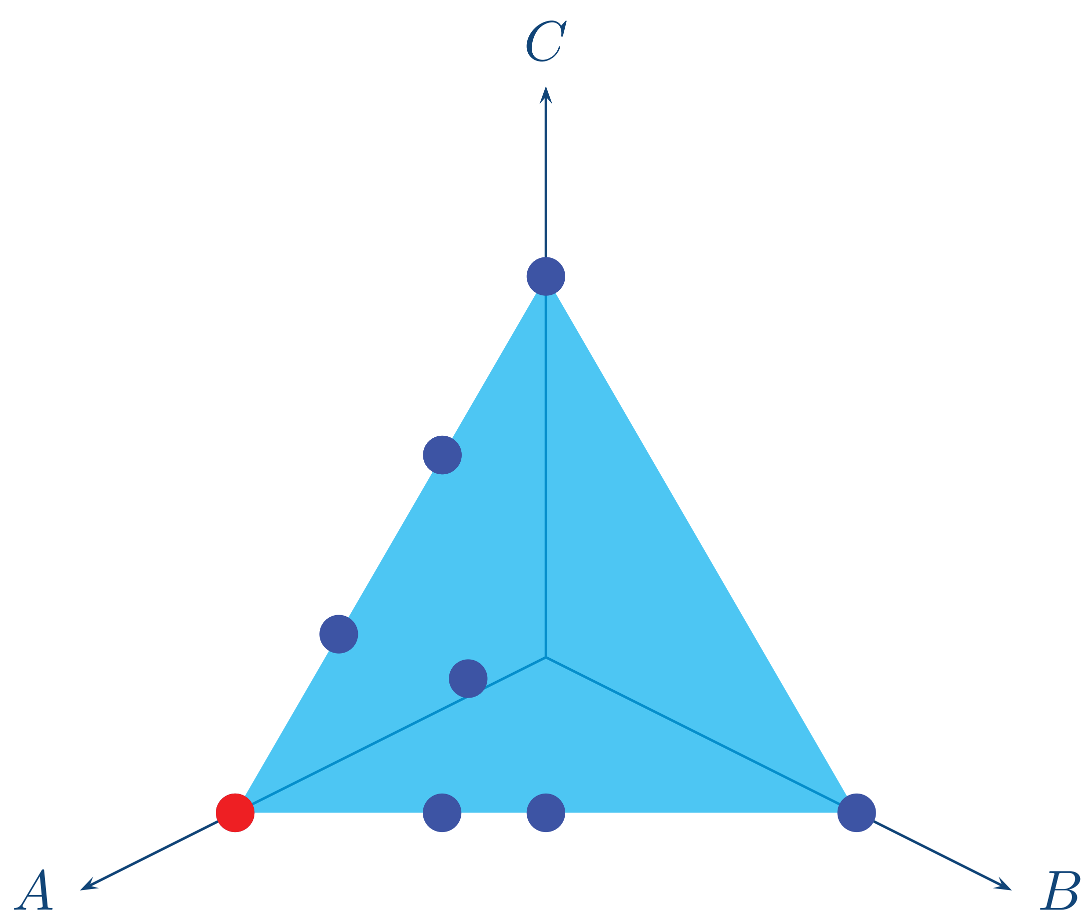


States of States of Knowledge

Mixed State Presentations ...



$$\mu = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$
$$w = 111001100$$



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:

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

Comments:

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

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:

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

$$= \frac{\Pr\left(X_t^L = w, S_{t+L} |\, \mu_t(\lambda)\right)}{\Pr\left(X_t^L = w |\, \mu_t(\lambda)\right)}$$

$$= \frac{\mu_t(\lambda) T^{(w)}}{\mu_t(\lambda) T^{(w)} \mathbf{1}}$$

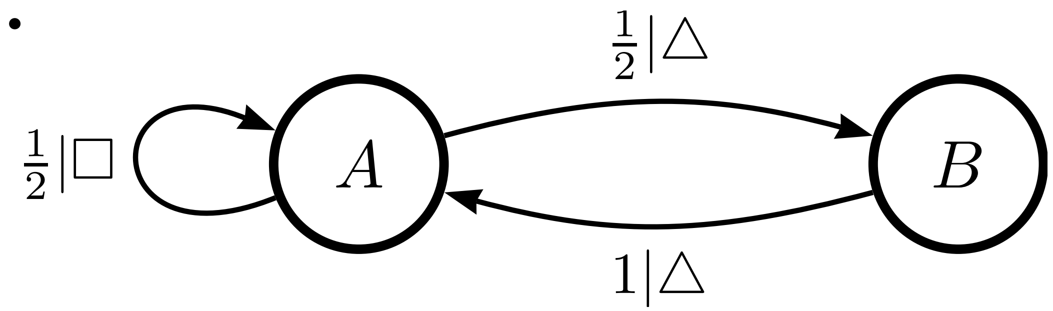
With $t = 0$ and $\mu_t(\lambda) = \pi$, we have:

$$\mu_L(w) = \frac{\pi T^{(w)}}{\pi T^{(w)} \mathbf{1}}$$

States of States of Knowledge

Mixed State Presentations ...

Example: Even Process



Have: $\mu_0(\lambda) = \pi = \begin{bmatrix} 2/3 & 1/3 \end{bmatrix}$

Mixed state for $w = \triangle$:

$$\begin{aligned} \mu_1(\triangle) &= \Pr(S_1 | X_0 = \triangle, S_0 \sim \pi) \\ &= \frac{\pi T^{\triangle}}{\pi T^{\triangle} \mathbf{1}} \\ &= \frac{\begin{bmatrix} 1/3 & 1/3 \end{bmatrix}}{2/3} \\ &= \begin{bmatrix} 1/2 & 1/2 \end{bmatrix} \end{aligned}$$

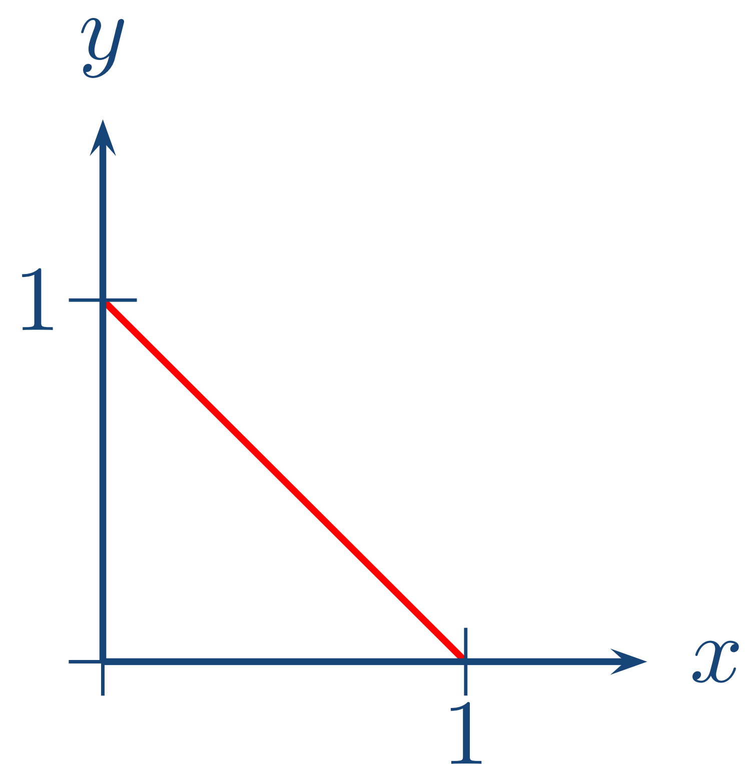
After seeing a $\triangle$, equally likely to be in A or B.

States of States of Knowledge

Mixed State Presentations ...

Simplicies:

- $|\mathcal{V}|$-dimensional vectors are L_1-normalized (in probability)
- Visualize on a $(|\mathcal{V}| - 1)$-dimensional simplex



1-simplex

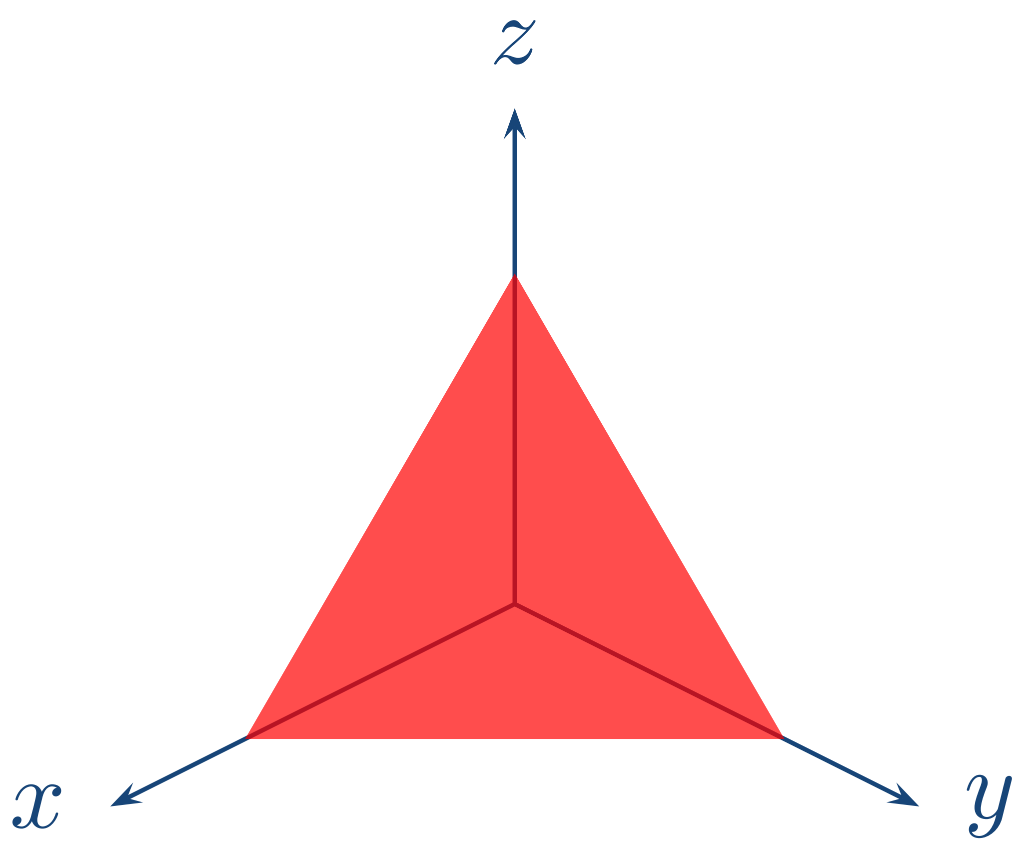
States of States of Knowledge

Mixed State Presentations ...

Simplicies:

- $|\mathcal{V}|$-dimensional vectors are L_1-normalized (in probability)

- Visualize on a $(|\mathcal{V}| - 1)$-dimensional simplex



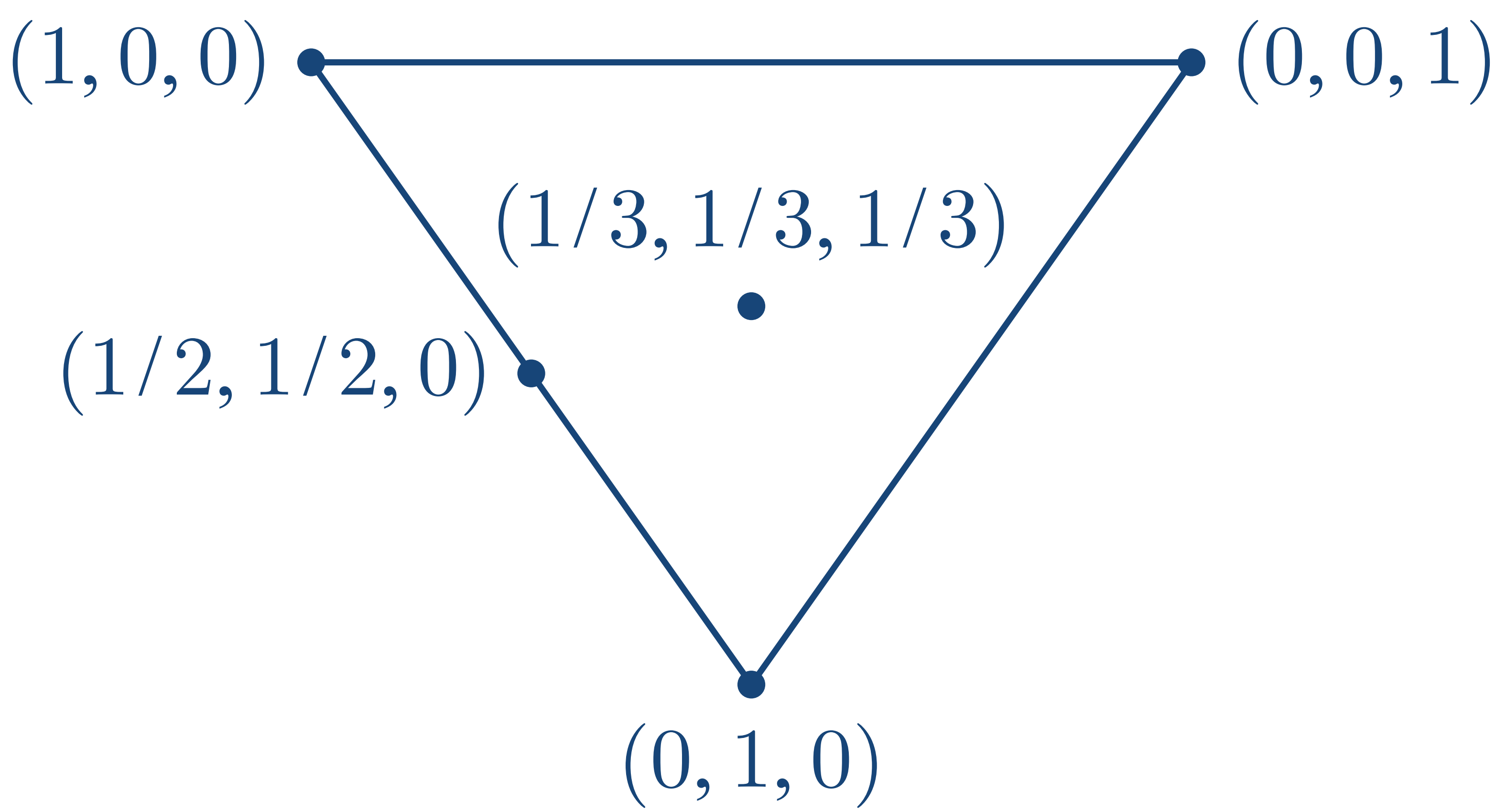
2-simplex

Mixed State Presentations ...

Simplicies:

- $|\mathcal{V}|$-dimensional vectors are L_1-normalized (in probability)

- Visualize on a $(|\mathcal{V}| - 1)$-dimensional simplex



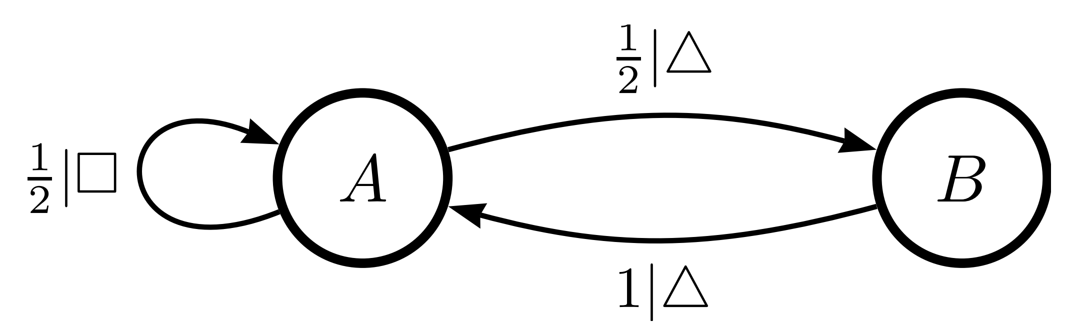
States of States of Knowledge

Mixed State Presentations ...

Interpretation of mixed states:

- Uncertainty in the state given a word and starting distribution.
- $H[\mu_L(w)] = 0$: State of machine is known with probability 1.
- Mixed states with zero entropy are basis vectors: "pure" states.
- Distributions (or mixtures) over "pure" states.
- Points in a geometric space.
- Different words can lead to same state uncertainty:

Example: Even Process ...



$$\mu_1(\triangle) = \mu_3(\triangle\triangle\triangle) = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$$

States of States of Knowledge

Mixed State Presentations ...

Idea of mixed states:

- Words have a natural dynamic (concatenation): $w \xrightarrow{s} ws$
- Equivalence relation: $w \sim_{\mathrm{MSP}} w' \iff \mu(w) = \mu(w')$
- Each mixed state is a "state" of state uncertainty.
- Dynamic over mixed states gives evolution of uncertainty:

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

- Unifilarity inherited from dynamic over words.

States of States of Knowledge

Reading for next lecture: CMR articles

TBA

PRATISP

IACP

IACPLCOCS