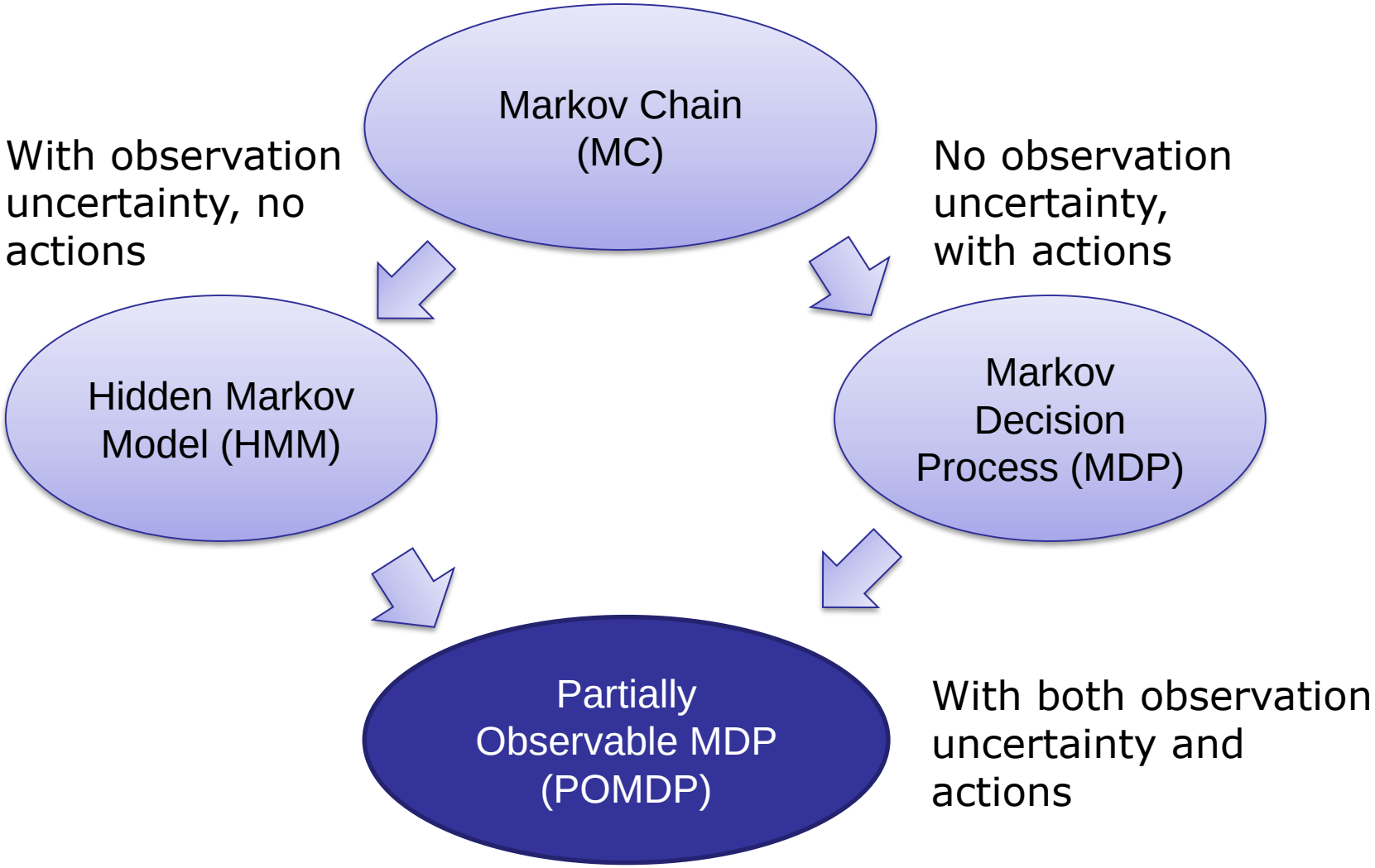
Today

Partially Observable Markov Decision Processes (POMDP)

- Problem Formulation
- Sufficient Statistics
- Algorithms

Markov Models



Review: Markov Decision Process

Key Elements of a Markov Decision Process:

- Decision Epochs, $\mathcal{T} = \{1,2, \dots, T\}$
- States, $\mathcal{S} = \{1,2, \dots, S\}$
- Actions, $\mathcal{A} = \{1,2, \dots, A\}$
- Rewards, r: $\mathcal{S} \times \mathcal{A} \mapsto \Re$
- Transition Probability Matrix, P

Partially Observable Markov Decision Process

Key Elements of a Partially Observable Markov Decision Process:

- Decision Epochs
- Core States
- Actions
- Observations (denoted o)
- Rewards
- Transition Probability Matrix P
- Observation Probability Matrix, Q

Partially Observable Markov Decision Process

System is in some core state s_0 (which is unknown to the decision-maker)

Decision-maker takes action a_0

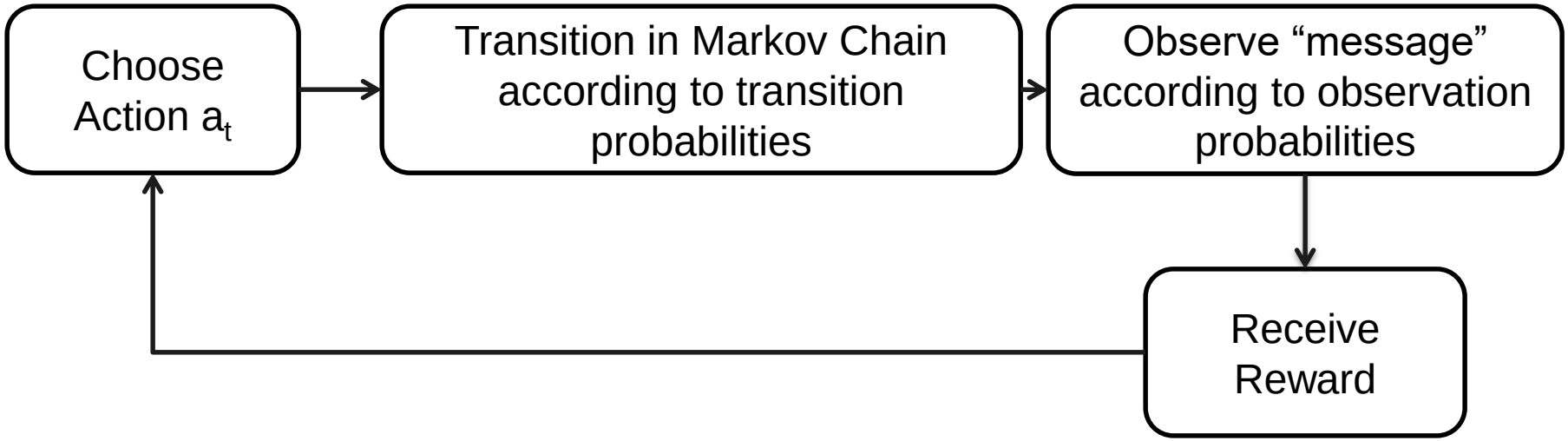
System transitions to a new core state, s_1, according to the transition probabilities $p(s_1|s_0, a_0)$

Decision-maker receives an observation o_1 which occurs with probability $p(o_1|s_1)$ given that the system is in state s_1

Decision-maker takes action a_1

…

Partially Observable Markov Decision Process



Breast Cancer Example - Background

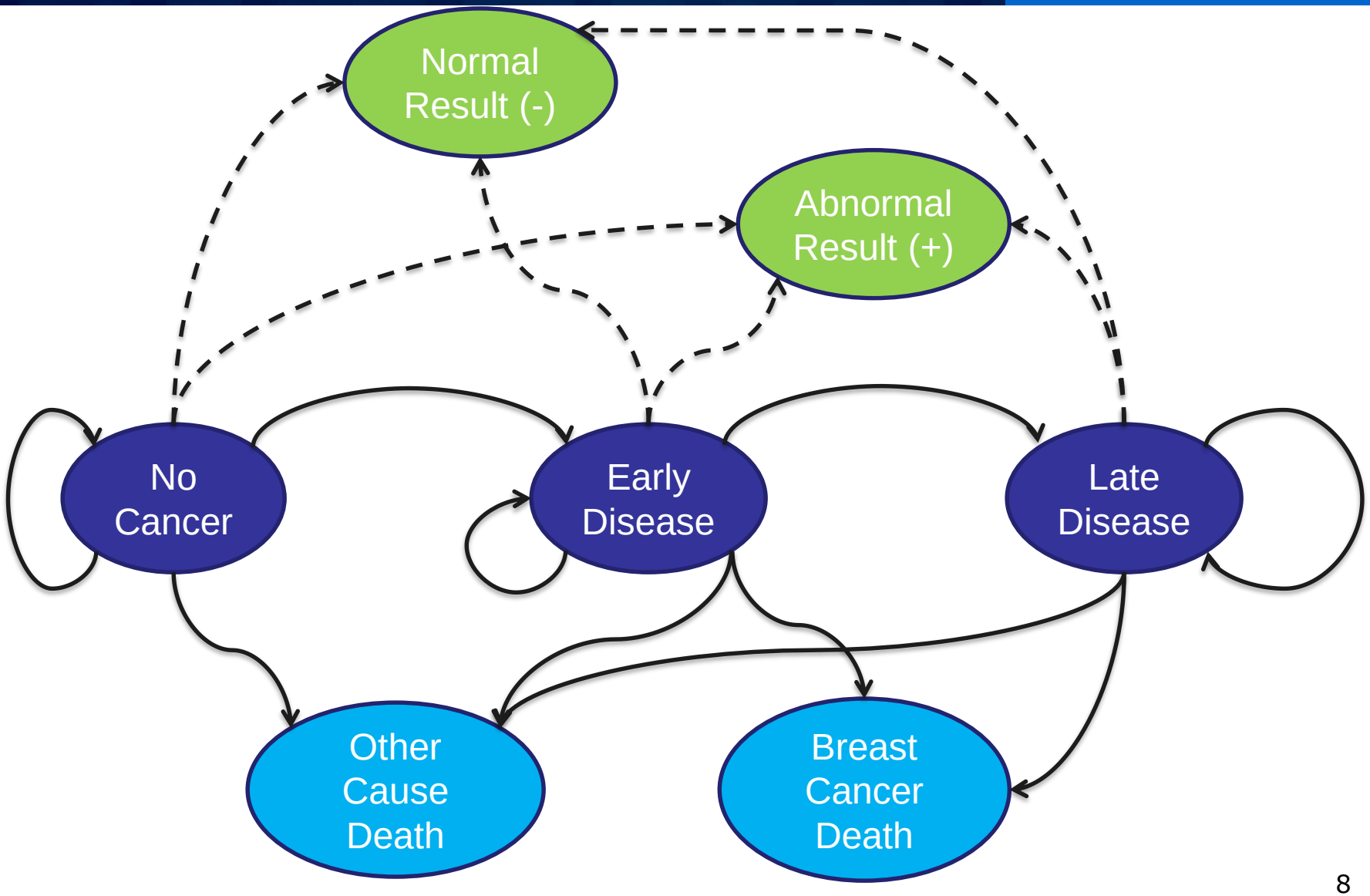
Mammography:

- An x-ray picture of the breast
- Can be used to check if a woman has breast cancer
- Imperfect test: It is possible that a woman may receive an “abnormal” result even if she is cancer-free or that she receives a “normal” result even if she has cancer

Consequences:

- Not treating someone with cancer can decrease their chance of survival
- Treating someone with cancer causes unnecessary stress and pain associated with treatment

Breast Cancer Example



Breast Cancer Example

States:

- Unobservable (core) states: {No Cancer, Early Disease, Late Disease}
- Observable states: {Other Cause Death, Breast Cancer Death}

Actions:

- {Mammography, Wait to Mammography}

Observations:

- {Normal Result, Abnormal Result}

Transition Probabilities:

- Describe the natural disease progression

Breast Cancer Example

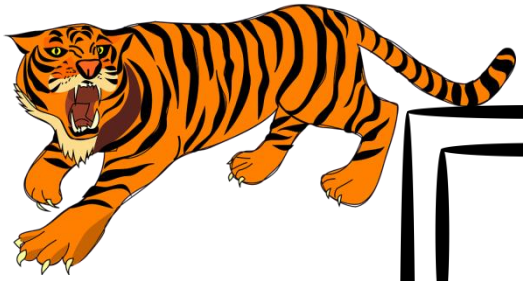
Observation Matrix, Q

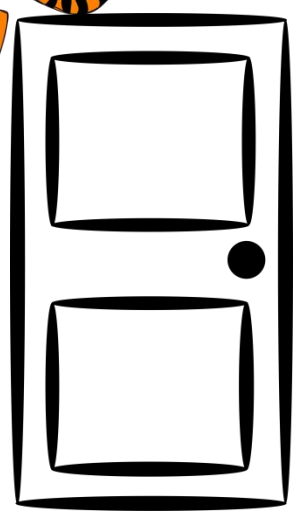
Matrix of conditional probabilities of observing observation $o \in \mathcal{O}$ given system is in state $s \in \mathcal{S}$

	No Cancer	Early Disease	Late Disease
Normal	$P(Normal \mid No\ Cancer)$	$P(Normal \mid Early)$	$P(Normal \mid Late)$
Abnormal	$P(Abnormal \mid No\ Cancer)$	$P(Abnormal \mid Early)$	$P(Abnormal \mid Late)$

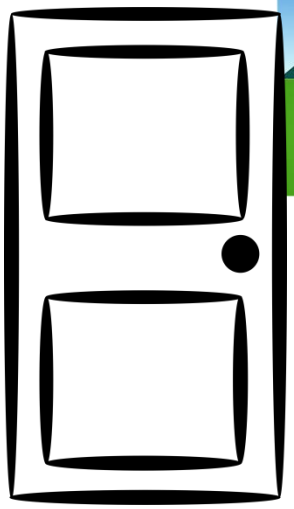
Maillart, L. M., Ivy, J. S., Ransom, S., & Diehl, K. (2008). Assessing dynamic breast cancer screening policies. *Operations Research*, *56*(6), 1411-1427.

In-Class Activity: The Tiger Problem





Reward if you open the door with Tiger: -100



Reward if you open the door to Freedom: +10

The Tiger is somewhat lazy and only changes location 25% of the time.

Your Options:

- Listen
- Open Left Door
- Open Right Door

If you choose to listen, you will hear a sound from the left or the right. Due to echoes in the room, you can only be 85% sure you heard the roar from the correct direction.

In-Class Activity: The Tiger Problem

Core States: Describes the location of the Tiger (and end of game)

Unobservable: {Behind Left Door , Behind Right Door}

Observable: {Eaten, Escaped}

Actions:

$\mathcal{A}$ ={Listen, Open Left Door, Open Right Door}

In-Class Activity: The Tiger Problem

Rewards:

$$r(s,a,s') = \begin{cases} -100, & s' = Eaten, s \in \{BLD, BRD\}, a \in \{OLD, ORD\} \\ 10, & s' = Escaped, s \in \{BLD, BRD\}, a \in \{OLD, ORD\} \\ -1, & listen \end{cases}$$

In-Class Activity: The Tiger Problem

Transition Probabilities

$$P(Listen) = \begin{matrix} \\ \text{BLD} \\ \text{BRD} \\ \text{Eaten} \\ \text{Escaped} \end{matrix} \begin{matrix} \begin{matrix} \text{BLD} & \text{BRD} & \text{Eaten} & \text{Escaped} \end{matrix} \\ \begin{bmatrix} 0.75 & 0.25 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{matrix}$$

$$P(Open\ Left\ Door) = \begin{bmatrix} 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 0.25 & 0.75 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P(Open\ Right\ Door) = \begin{bmatrix} 0 & 0 & 0.25 & 0.75 \\ 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The Tiger Problem: Observations

Observations:

{Hear Roar From Left , Hear Roar From Right}

Observation Matrix: Describes conditional probability of observation given state

	Behind Left Door	Behind Right Door
Hear Roar Left	0.85	0.15
Hear Roar Right	0.15	0.85

What should we do?

1) What ideas do you have to help solve this problem?

2) What information would be most useful to the decision maker?

Sufficient Statistic

Sufficient statistic: Information that is sufficient for decision-making in a sequential decision-making problem.

In POMDPs, we define a *belief vector*, ***b***, to be the sufficient statistic. ***b*** = $\{p_1,..,p_n\}$ where p_i represents the probability of state i.

The belief vector is defined as a vector with one element for each state

$$b_t(s_t) = P(s_t | \underbrace{o_t, a_{t-1}, o_{t-1}, a_{t-2}, \dots, o_1, a_0}_{h_t})$$

Complete history of observations up to and including t as well as history of actions up to and including time t-1

Sufficient Statistic: Bayesian Updating

Goal: To have an efficient way to update the belief vector

$$b_{t+1} = T(b_t, a_t, o_{t+1})$$

Facts about Conditional Probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ (1) Bayes' Rule

$$\mathrm{P}(A, B) = P(A|B)P(B)$$ (2) Conditional Probability

If B_1, B_2, …, B_n are mutually exclusive and collectively exhaustive

$$P(A) = \sum_{i=1}^{n} P(A, B_i)$$ (3) Law of Total Probability

Sufficient Statistic: Bayesian Updating

Bayesian Update of Sufficient Statistic:

$$b_t(s_t) = P(s_t|o_t, a_{t-1}, o_{t-1}, a_{t-2}, \dots, o_1, a_0)$$

$$= \frac{P(s_t, o_t, a_{t-1}|\, o_{t-1}, a_{t-2}, \dots, o_1, a_0)}{P(o_t, a_{t-1}|\, o_{t-1}, a_{t-2}, \dots, o_1, a_0)}$$

Numerator

$$P(s_t, o_t, a_{t-1}|\, \overbrace{o_{t-1}, a_{t-2}, \dots, o_0, a_0}^{h_{t-1}}) = \sum_{s_{t-1}\in\mathcal{S}} P(s_t, o_t, a_{t-1}, s_{t-1}|\, h_{t-1})$$

$$= \sum_{s_{t-1}\in\mathcal{S}} P(o_t|s_t, a_{t-1}, s_{t-1}, h_{t-1})\, P(s_t|a_{t-1}, s_{t-1}, h_{t-1}) P(a_{t-1}|s_{t-1}, h_{t-1}) P(s_{t-1}|h_{t-1})$$

$$= P(a_{t-1}|h_{t-1}) P(o_t|s_t) \sum_{s_{t-1}\in\mathcal{S}} P(s_t|a_{t-1}, s_{t-1})\, b_{t-1}(s_{t-1})$$

Sufficient Statistic: Bayesian Updating

Bayesian Update of Sufficient Statistic:

$$b_t(s_t) = P(s_t|o_t, a_{t-1}, o_{t-1}, a_{t-2}, \dots, o_1, a_0)$$

$$= \frac{P(s_t, o_t, a_{t-1}|\, o_{t-1}, a_{t-2}, \dots, o_1, a_0)}{P(o_t, a_{t-1}|\, o_{t-1}, a_{t-2}, \dots, o_1, a_0)}$$

Applying (2)

Numerator

$$P(s_t, o_t, a_{t-1}|\, \overbrace{o_{t-1}, a_{t-2}, \dots, o_0, a_0}^{h_{t-1}}) = \sum_{s_{t-1}\in\mathcal{S}} P(s_t, o_t, a_{t-1}, s_{t-1}|\, h_{t-1})$$

Applying (3)

$$= \sum_{s_{t-1}\in\mathcal{S}} P(o_t|s_t, a_{t-1}, s_{t-1}, h_{t-1})\, P(s_t|a_{t-1}, s_{t-1}, h_{t-1}) P(a_{t-1}|s_{t-1}, h_{t-1}) P(s_{t-1}|h_{t-1})$$

Repeatedly applying (2)

$$= P(a_{t-1}|h_{t-1}) P(o_t|s_t) \sum_{s_{t-1}\in\mathcal{S}} P(s_t|a_{t-1}, s_{t-1})\, b_{t-1}(s_{t-1})$$

Independence and Definitions

Sufficient Statistic: Bayesian Updating

Bayesian Update of Sufficient Statistic:

$$b_t(s_t) = P(s_t|o_t, a_{t-1}, o_{t-2}, a_{t-2}, \dots, o_0, a_0) = \frac{P(s_t, o_t, a_{t-1}|\, o_{t-2}, a_{t-2}, \dots, o_0, a_0)}{P(o_t, a_{t-1}|\, o_{t-2}, a_{t-2}, \dots, o_0, a_0)}$$

Denominator

$$P(o_t, a_{t-1}|\, \overbrace{o_{t-1}, a_{t-2}, \dots, o_0, a_0}^{h_{t-1}}) = \sum_{s_t' \in \mathcal{S}} \sum_{s_{t-1} \in \mathcal{S}} P(s_t', o_t, a_{t-1}, s_{t-1}|\, h_{t-1})$$

$$= \sum_{s_t' \in \mathcal{S}} \sum_{s_{t-1} \in \mathcal{S}} P(o_t|s_t', a_{t-1}, s_{t-1}, h_{t-1})\, P(s_t'|a_{t-1}, s_{t-1}, h_{t-1}) P(a_{t-1}|s_{t-1}, h_{t-1}) P(s_{t-1}|h_{t-1})$$

$$= P(a_{t-1}|h_{t-1}) \sum_{s_t' \in \mathcal{S}} P(o_t|s_t') \sum_{s_{t-1} \in \mathcal{S}} P(s_t'|a_{t-1}, s_{t-1})\, b_{t-1}(s_{t-1})$$

Sufficient Statistic: Bayesian Updating

Bayesian Update of Sufficient Statistic:

$$b_t(s_t) = P(s_t|o_t, a_{t-1}, o_{t-2}, a_{t-2}, \dots, o_0, a_0) = \frac{P(s_t, o_t, a_{t-1}|\, o_{t-2}, a_{t-2}, \dots, o_0, a_0)}{P(o_t, a_{t-1}|\, o_{t-2}, a_{t-2}, \dots, o_0, a_0)}$$

Denominator

$$P(o_t, a_{t-1}|\, \overbrace{o_{t-1}, a_{t-2}, \dots, o_0, a_0}^{h_{t-1}}) = \sum_{s_t' \in \mathcal{S}} \sum_{s_{t-1} \in \mathcal{S}} P(s_t', o_t, a_{t-1}, s_{t-1}|\, h_{t-1})$$

Applying (3)

$$= \sum_{s_t' \in \mathcal{S}} \sum_{s_{t-1} \in \mathcal{S}} P(o_t|s_t', a_{t-1}, s_{t-1}, h_{t-1})\, P(s_t'|a_{t-1}, s_{t-1}, h_{t-1}) P(a_{t-1}|s_{t-1}, h_{t-1}) P(s_{t-1}|h_{t-1})$$

Repeatedly applying (2)

$$= P(a_{t-1}|h_{t-1}) \sum_{s_t' \in \mathcal{S}} P(o_t|s_t') \sum_{s_{t-1} \in \mathcal{S}} P(s_t'|a_{t-1}, s_{t-1})\, b_{t-1}(s_{t-1})$$

Independence and Definitions

Sufficient Statistic: Bayesian Updating

Goal: To have an efficient way to update the belief vector

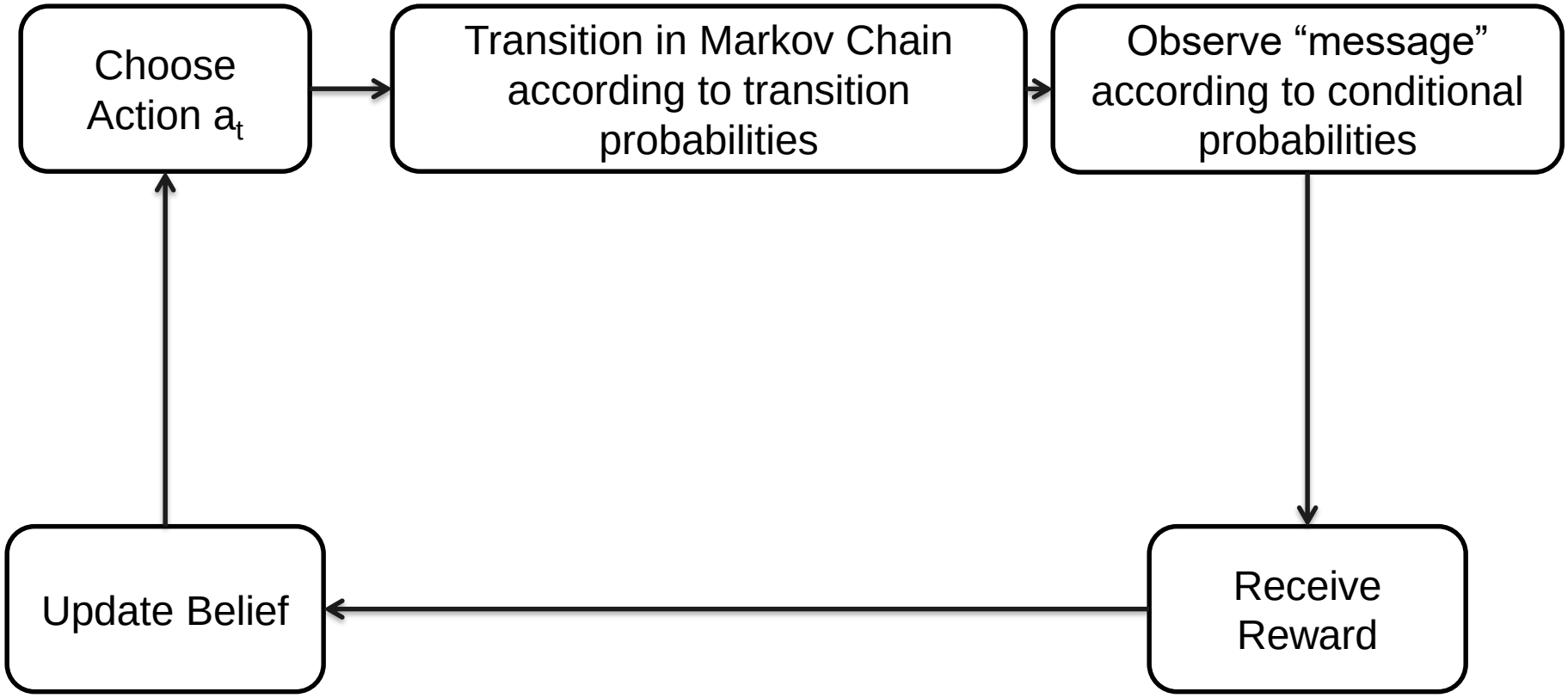
$$b_{t+1} = T(b_t, a_t, o_{t+1})$$

$$b_t(s_t) = P(s_t|o_t, a_{t-1}, o_{t-2}, a_{t-2}, \dots, o_0, a_0) = \frac{P(s_t, o_t, a_{t-1}|\, o_{t-2}, a_{t-2}, \dots, o_0, a_0)}{P(o_t, a_{t-1}|\, o_{t-2}, a_{t-2}, \dots, o_0, a_0)}$$

$$= \frac{P(o_t|s_t)\sum_{s_{t-1}\in\mathcal{S}} P(s_t|s_{t-1}, a_{t-1})\, b_{t-1}(s_{t-1})}{\sum_{s_t'\in\mathcal{S}} P(o_t|s_t')\sum_{s_{t-1}\in\mathcal{S}} P(s_t'|s_{t-1}, a_{t-1})\, b_{t-1}(s_{t-1})}$$

Now everything is in terms of transition probabilities, observation probabilities, and last belief vector.

POMDP Sequence of Events



Value Functions in POMDPs

Rewards Vector: $r(a_t) = [r_1(a_t), \dots, r_S(a_t)]'$ denotes the expected rewards under transitions and observations.

$$r_{s_t}(a_t) = \sum_{o_{t+1} \in \mathcal{O}} \sum_{s_{t+1} \in \mathcal{S}} r(s_t, a_t, s_{t+1}, o_{t+1}) p(s_{t+1}|s_t, a_t) p(o_{t+1}|s_{t+1})$$

Value Function: In POMDPs, the value function is defined on the belief space.

- Immediate Expected Rewards associated with taking action a_t can be written as the vector product $b_t \cdot r(a_t)$

Finite horizon value function:

Probability of receiving observation o_{t+1} given belief vector b_t and action a_t

$$v_t(b_t) = \max_{a_t \in \mathcal{A}} \left\{ b_t \cdot r_t(a_t) + \lambda \sum_{o_{t+1} \in \mathcal{O}} P(o_{t+1}|b_t, a_t) v_{t+1}(T(b_t, a_t, o_{t+1})) \right\}$$

$$v_{T+1}(b_{T+1}) = b_{T+1} \cdot r_{T+1}$$

Algorithm for Solving POMDPs

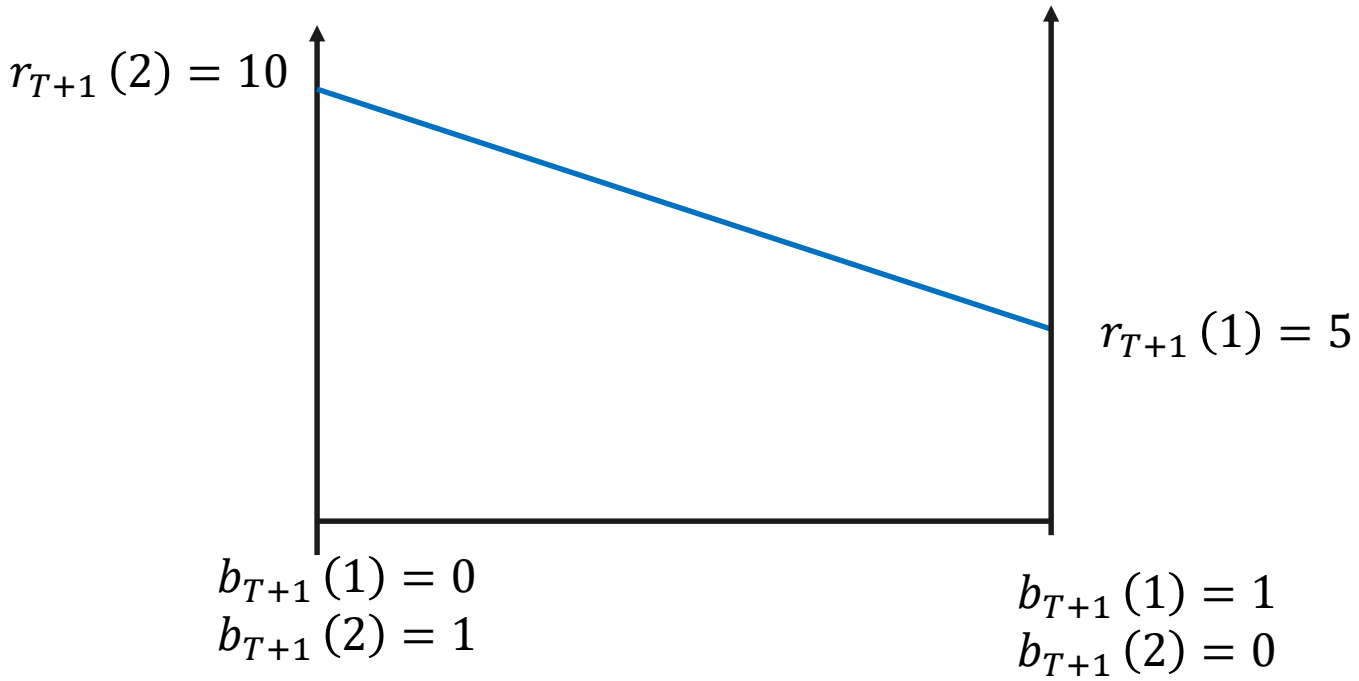
Suppose there are two core states: {1, 2}

We can represent the *belief space* with a line segment

Algorithm for Solving POMDPs

Suppose there are two core states: {1, 2}

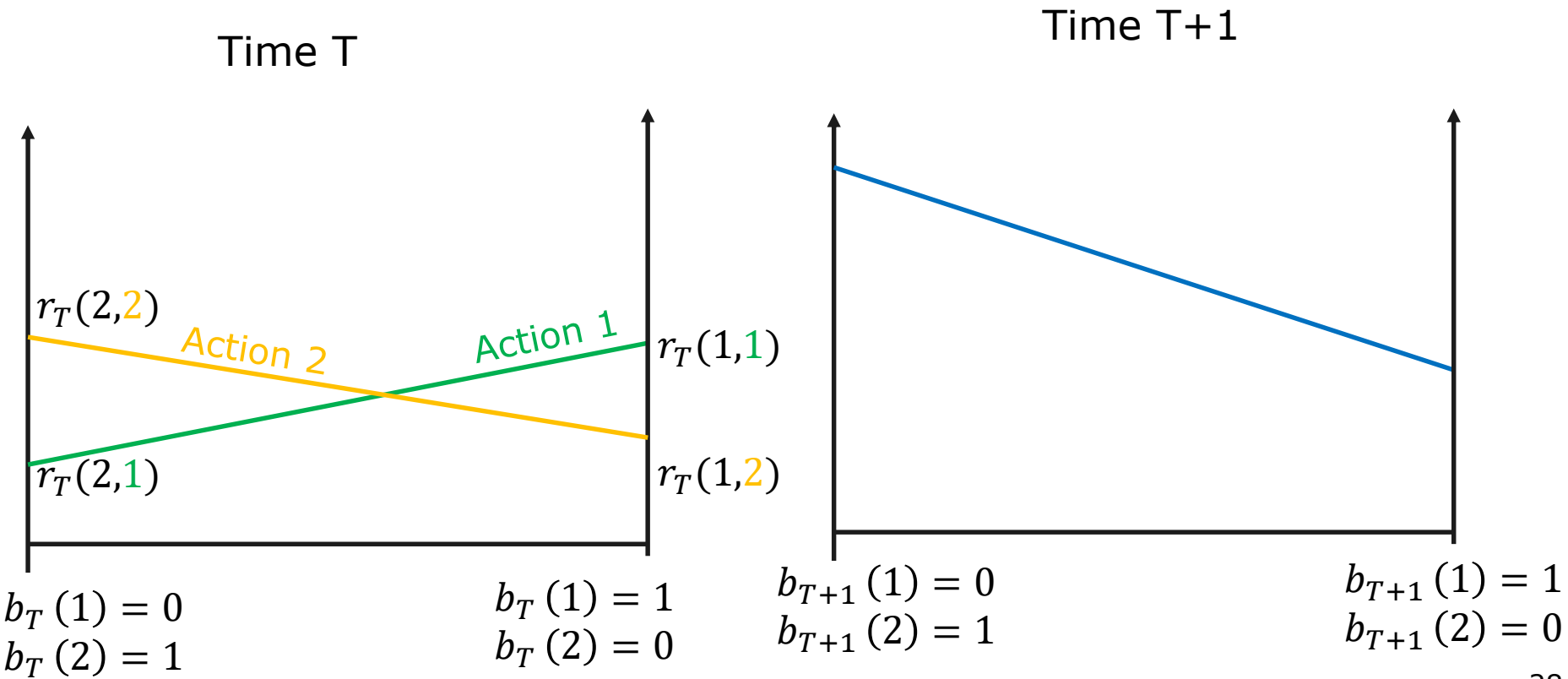
At time T+1, we can represent the value function with a vector



Algorithms

Suppose there are two core states: {1, 2}

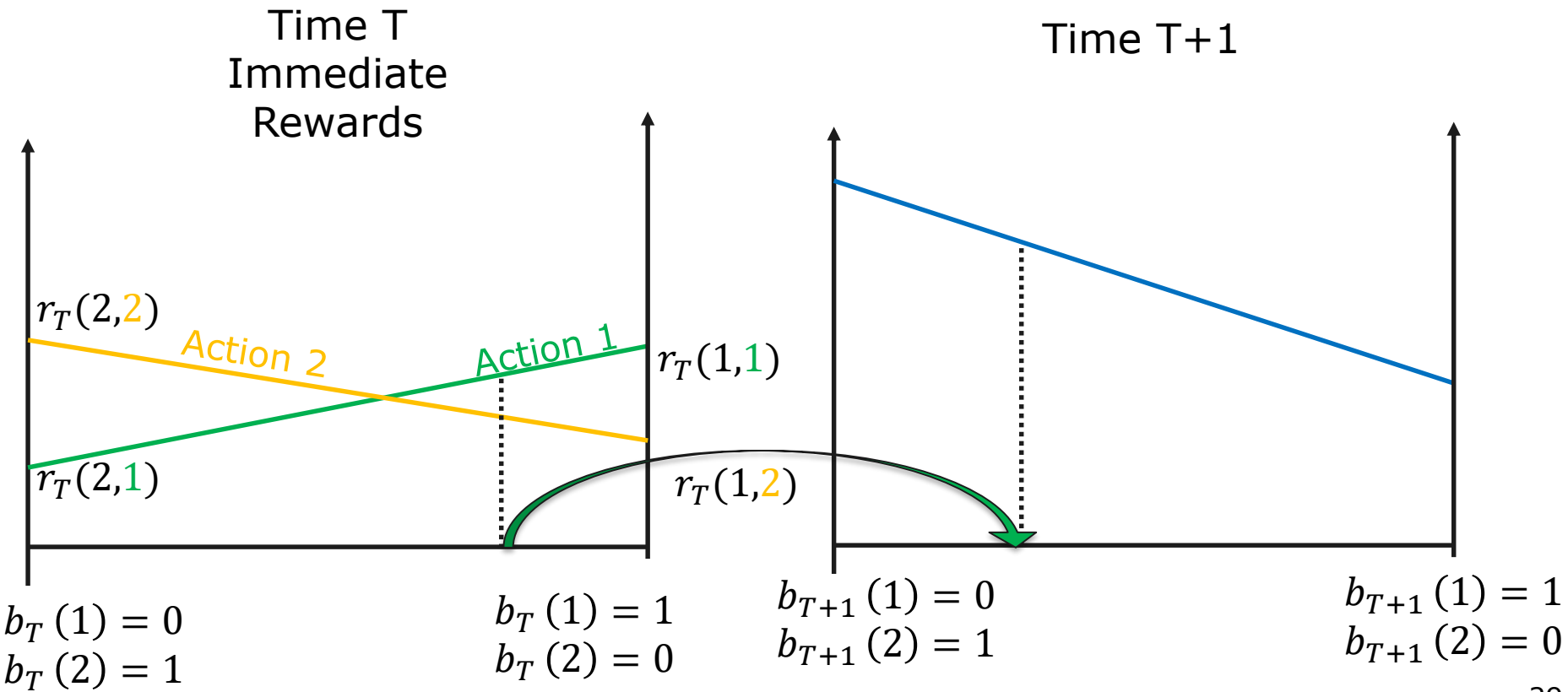
At time T, we can represent the immediate rewards with a set of vectors



Algorithms

Suppose there are two core states: {1, 2}

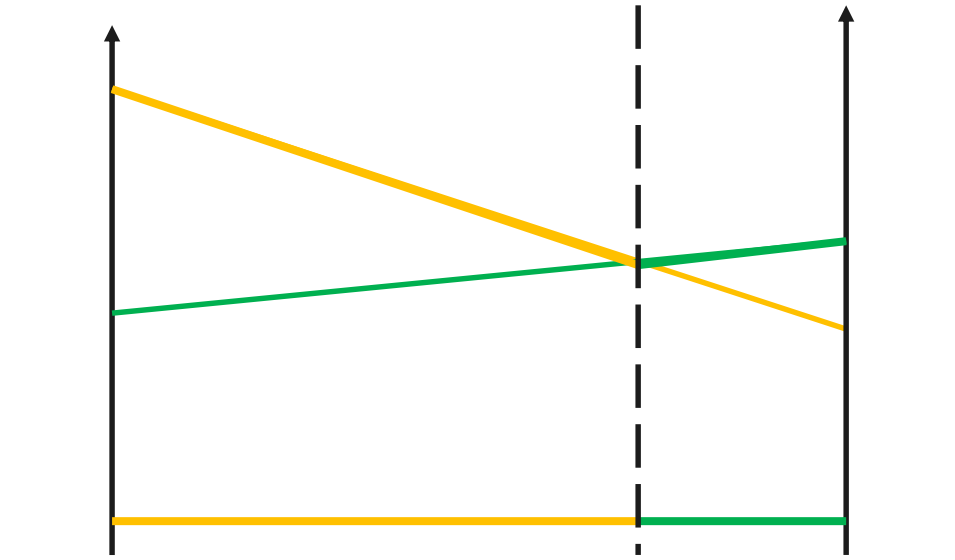
At time T, we can represent the immediate rewards with a set of vectors



Algorithm for Solving POMDPs

Suppose there are two core states: {1, 2}

At any time t, we can represent the value function with a set of vectors (called α-vectors)



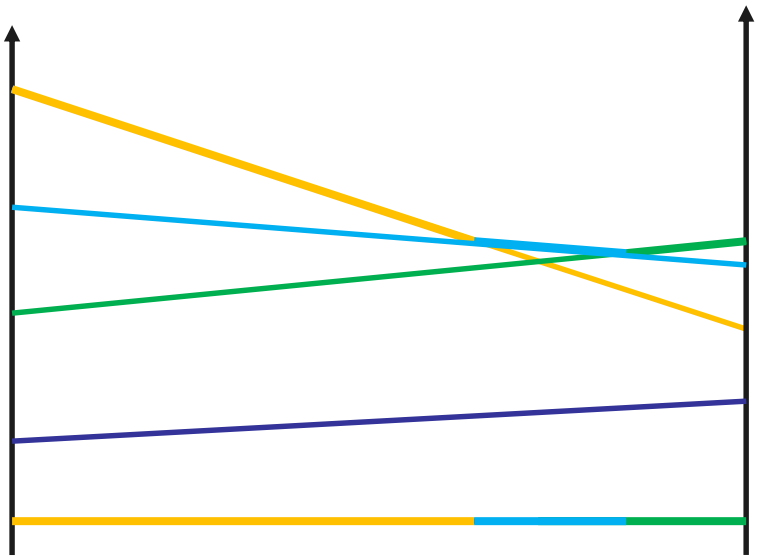
$b_T(1) = 0$
$b_T(2) = 1$

$b_T(1) = 1$
$b_T(2) = 0$

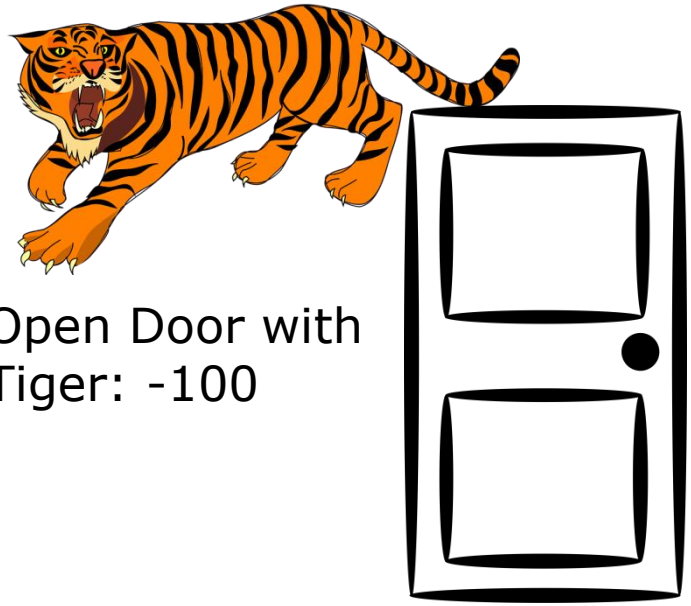
Algorithm for Solving POMDPs

Suppose there are two core states: {1, 2}

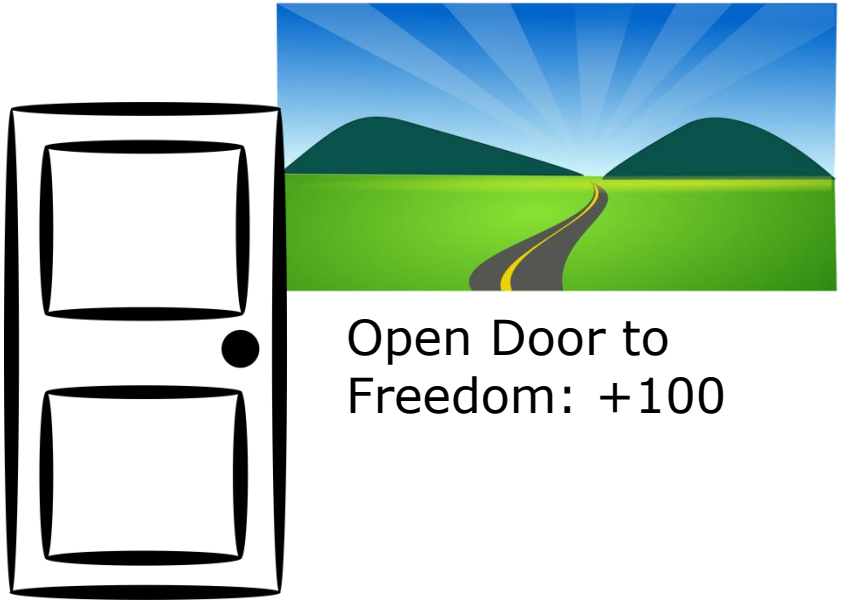
As we perform the backwards induction, the number of α-vectors grow exponentially.



The Tiger Problem: Bayesian Updating



Open Door with Tiger: -100



Open Door to Freedom: +100

Your Options:

- Listen
- Open Left Door
- Open Right Door

The Tiger is somewhat lazy and only changes location 25% of the time.

If you choose to listen, you will hear a sound from the left or the right. Due to echoes in the room, you can only be 85% sure you heard the roar from the correct direction.

End of Class Activity: Tiger Problem

At time *t*, you believe there is a 50% chance that the tiger is behind the left door and a 50% chance the tiger is behind the right door.

You choose to listen and hear a roar from the left.

What is the belief vector at time *t+1*? What probability would you assign to the event "The tiger is behind the left door at time *t+1*"?

POMDP Summary

POMDPs compute the optimal action in partially observable, stochastic domains.

For finite horizon problems, the resulting value functions are piecewise linear and convex.

In each iteration the number of linear constraints grows exponentially.

POMDPs so far have only been applied successfully to very small state spaces with small numbers of possible observations and actions.

POMDP Learning Resources http://www.pomdp.org/