1. *How many states do we have for n possible mutations? What is it for n=10? And n=30?*
2. *What is the natural choice for the space of actions? How many of them do we have?*
3. *Can you propose a model of the transition probability ?*

The transition probability of the next state (k+1) depends only on the current state (memorylessness). We thus can use the Markov chain model to determine the transition probability.

In nucleotides mutation, the mutation of a nucleotide is regardless of its history mutation. The transition rate determines the transition probability by Markov chain model, e.g. state transition matrix JC69; TN93.

A picture containing table

Description automatically generated

1. *How can we estimate the experts' policy ?*

Policy: agent’s behaviour function

A policy is a distribution over actions given the states

Text, letter

Description automatically generated

A reward function is a weighted combination of features

Forward RL: Reward function 🡪(learn) policy

IRL: policy 🡪(learn) reward function: find the reward function that maximize the expert trajectory

Data: an agent performing some task that we want to learn the reward for.

**MaxEnt IRL**

The reward function is expressed as linear combinations with feature counts.

The constraint to solve the problem:

Schematic

Description automatically generated

To allow for computation with states, we use so-called features.

Feature matching:

Text

Description automatically generated

The larger the entropy and less information we can get. We thus The objective: maximize the likelihood of the observed data (max entropy)

Text

Description automatically generated

Pseudocode for MaxEnt:

Text

Description automatically generated

Disadvantages:

1. No contextual factors (e.g. time of day, weather) considered
2. It assumes linear reward function
3. Assumes state transitions are known

A MDP is a 5-tuple: states S, actions A; transition model T; Reward function and reward discount factor.