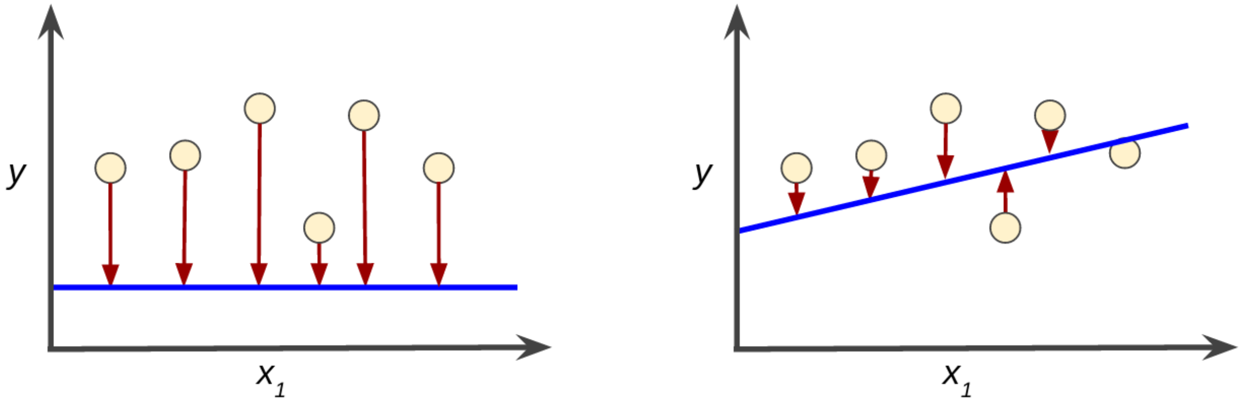
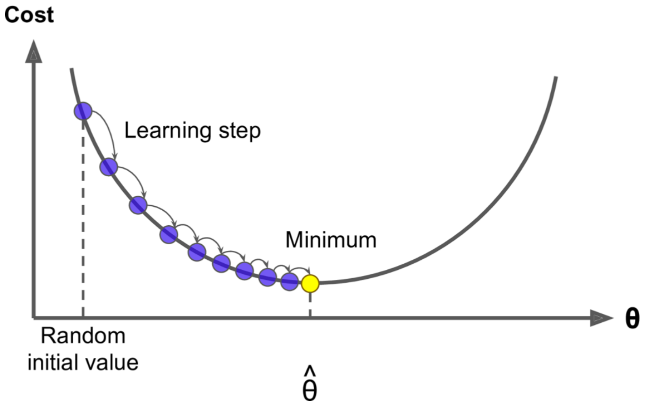
LLM Reasoning

❖ But what is the training objective?

In supervised learning:



- ➢ We attempt to minimize the loss between prediction and label.
- ➢ Minimize the loss function.



In reinforcement learning:

$$R_T = \sum_{i=0}^{T} r_{t+1} = r_t + r_{t+1} + \ldots + r_T$$

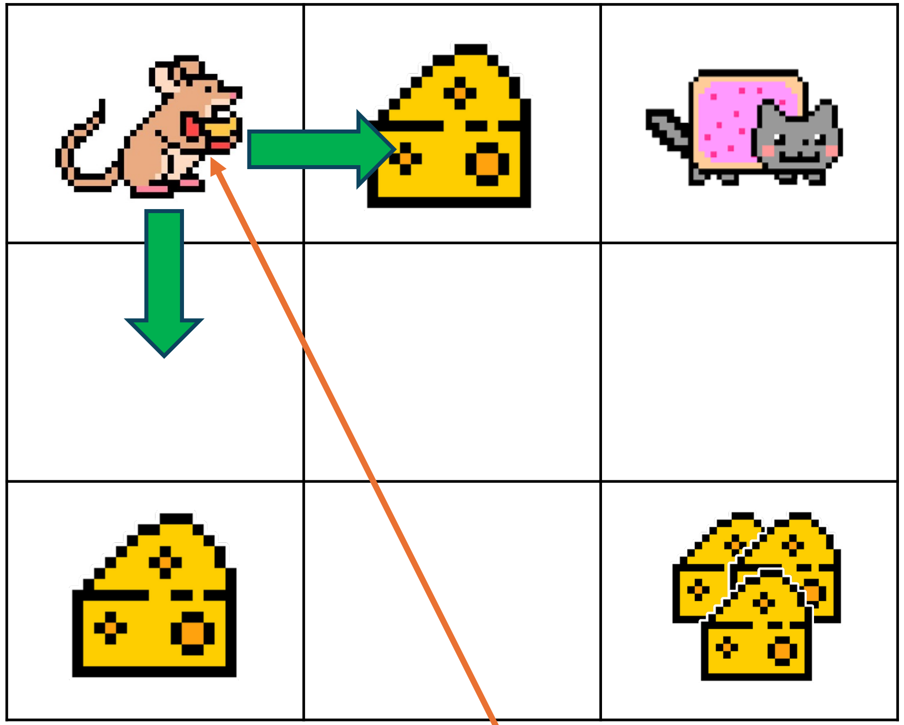
- ➢ We attempt to maximize the expected cumulative reward.
- ➢ Find optimal **policy $\boldsymbol{\pi}$**.

LLM Reasoning

❖ Policy

Given state S, our agent will have **many possible actions** A.

Points: 0



Possible actions at S_0: Right, Down.

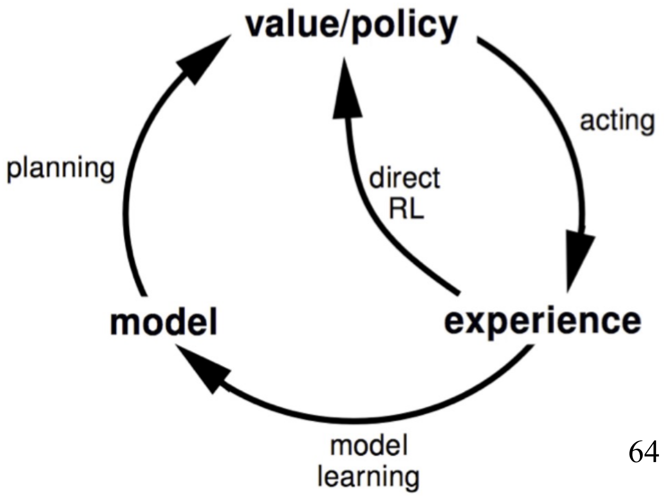
$$R_T = \sum_{i=0}^{T} r_{t+1} = r_t + r_{t+1} + \ldots + r_T$$

- ➢ In RL, we attempt to maximize the expected cumulative reward.

Need a way so that at every state, the agent could **be able to choose action that leads to the highest expected cumulative reward**.

π

Policy


value/policy
acting
planning
direct RL
model
experience
model learning