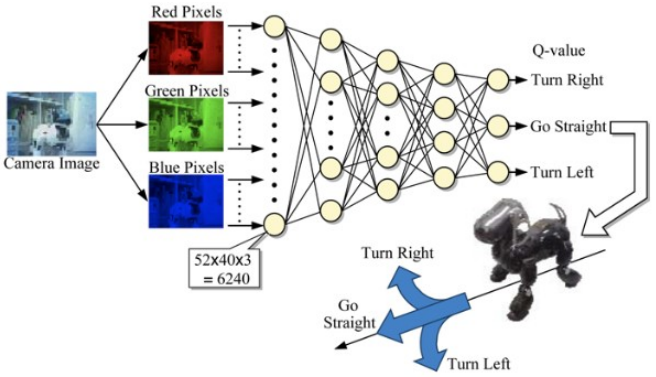
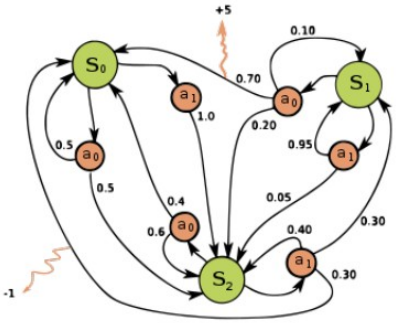
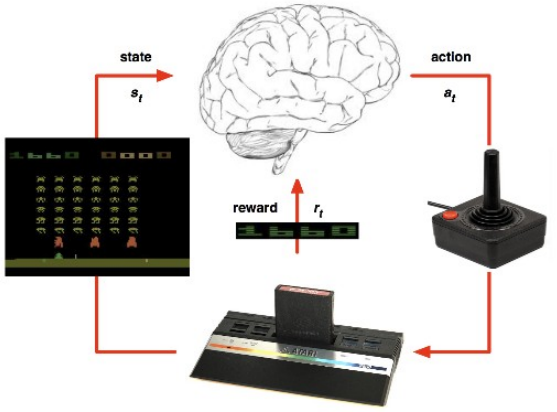
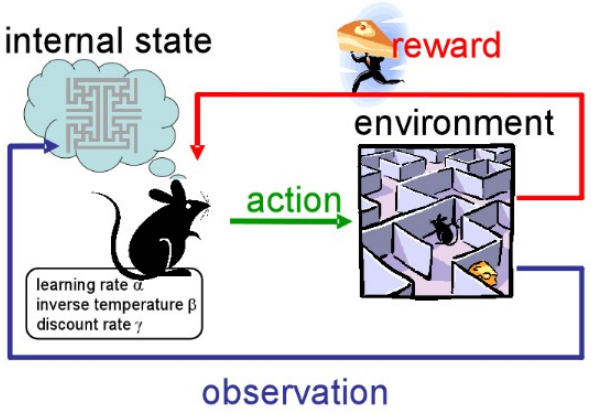
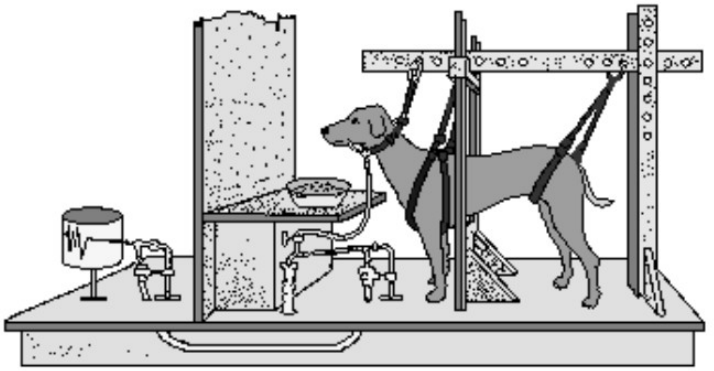
# COMP 138: Reinforcement Learning



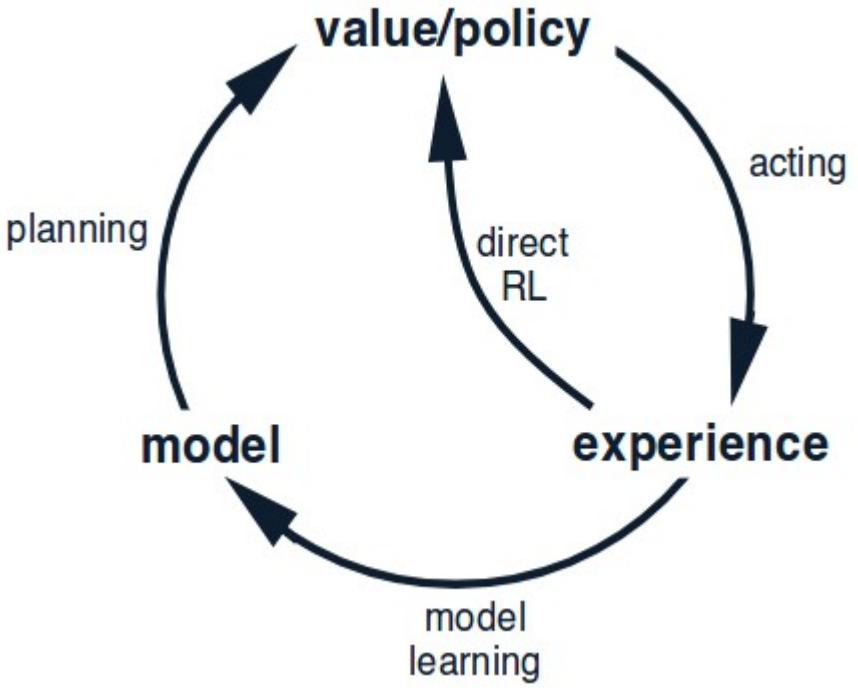



**Instructor**: Jivko Sinapov

# Today


value/policy
acting
planning
direct
RL
model
experience
model
learning

# Announcements

# Upcoming Project Due Dates

- Team Formation – Oct 17
- Project Proposal ~~– Oct 31st~~ Nov 3rd

# Overview of 8.1 and 8.2

# How do we make Dyna-Q handle stochastic environments?

- Small group activity
- Re-write the pseudocode and produce an algorithm which handles stochastic environments
- Now, modify the algorithm you wrote to handle “gradual” non-stationarity (or if it already does, discuss why)

# Reading Discussion

# Reading Discussion

“In the algorithm block in page.161, how should we understand this “sample next reward and sample next state”?? Does this mean that in this case, we actually have a model?”

“It is said that “This policy is built by the planning process while the agent is still wandering near the start state”. Before this planning happen, there is only one episode, does this mean that only one episode is enough to make the planning happen?”

# Reading Discussion

“How does the Dyna architecture balance between direct learning from real experience and indirect learning from simulated experiences using a model?”

- Mingjia

# Reading Discussion

“Could there be more elaboration on how prioritized sweeping is implemented in practice? How is the prioritization of state-action pairs determined, and how does it impact the convergence and efficiency of learning?”

- Yinkai

# Reading Discussion

“If my understanding is correct, Rollout Algorithms are algorithms that select actions based on values of next states (using the combination of those values and simulated actions from the current state to estimate action values), but how exactly are the values determined? If the values are all initialized arbitrarily, wouldn't Rollout Algorithms perform very poorly initially? Or is the intention for Rollout Algorithms to improve as they are run more times?”

- Randy

# Reading Discussion

“When planning is done online, while interacting with the environment, a number of issues arise. New information may change the model and thus the planning. How do we divide computing resources between decision making and model learning?”

- Prithvi

# Reading Discussion

“How does Real-time Dynamic Programming (RTDP) differ from traditional dynamic programming, and what types of tasks or scenarios are better suited for RTDP?”

- YuanYuan

# Research Article

# Research Article

“The article mentions that curriculum learning can be useful when an agent has converged on a suboptimal policy. How would the agent know that the policy is suboptimal and that it should trigger a curriculum?”

- Brennan

# Research Article

“If we disregard the economic costs of getting real experience in real-world robotic scenarios, would algorithms utilizing real experience outperform those using model experience?”

- Qidi

# Research Article

“Is it possible to come up with a generic approach to deriving a curriculum for a particular problem? Or can this only be done empirically, sampling different subtasks as shown in the paper.”

- Channi

# Research Article

“I was really curious about how hands-off the curriculum generation could be. The algorithms appeared like they could work mostly independently, but the examples provided for Ms. PacMan and HFO seemed to be tuned or defined by humans. What is the current progress of making curriculum learning autonomous?”

- Grayson

# Research Article

“Can you explain more about the “Promising Initializations”? Does this mean that for task M’, all states with higher rewards are set? If that is the case, then it is still a method of minimizing the observation space. But I thought “task simplification” already minimize the space.”

- Qing

# Research Article

“I have some questions for the paper. What might be the reasons that performance of mistake learning agent becomes worse than baseline as game steps increase? For these task transforming functions, are they problem specific so that we need to redefine them if there is slight change in rules of the environment?”

- Zixiao

# Research Article

“Were you able to successfully develop an automated method for selecting subtasks? Is this an area which you are still working on?”

- Tyler

# Research Article

“I have some confusion about the subgoal curriculum "classes". I believe we learned that it is quite dangerous to incentivize agents to solve subtasks, as it can lead to the agent not actually having an incentive to solve the main task. Is this only the case for reward functions, or is it still applicable here?”

- Andrew

# Research Article

“I'm intrigued by the concept of curriculum learning as presented by Narvekar et al. How can the principles of curriculum learning be extended and optimized for more complex, real-world applications? What are the limitations and challenges in designing and implementing a curriculum for RL agents?”

- Yinkai

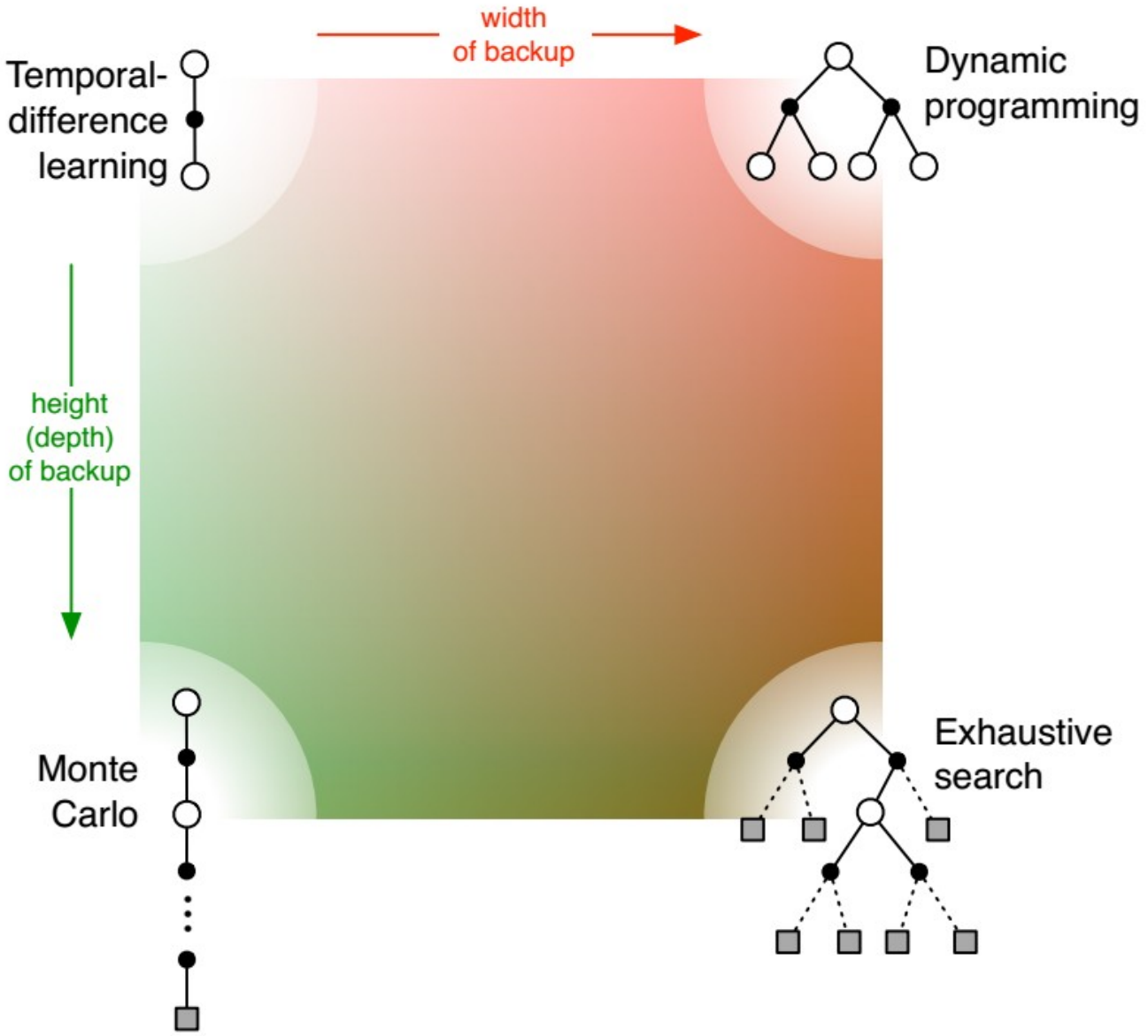
# Research Article

“Is there any ideas/starting points on how we can automatically generate curriculum tasks for a class of problems for curriculum learning, or does it typically require an expert to choose the tasks? If not, is using LLMs a prospective solution for this in terms of automatically processing a real world problem and the generated subtasks to generating a curriculum?”

- Prithvi

# Moderated Discussion

# Unified View




[source: Sutton]

# Planning and Learning

- Model vs. Model-Free RL
- Types of Models:
  - Distributional
  - Sample
- Q-planning and Dyna-Q

THE END