Course 2, Module 5
Planning, Learning & Acting

CMPUT 397

Fall 2019

Announcements

- We start C3 next week! Watch eclass announcements for the link to our private session

- We removed the assignment on Policy Gradient at the end of the course. So that means your assignment mark is from 11 notebooks (graded quiz).

Any questions about course admin?

- Link for questions:
  - **http://www.tricider.com/brainstorming/3LEf4A3lOPB**

Review of Course 2, Module 5
Learning a Model AND Planning

Video 1: What is a Model?

- **All about models.** What they are? How they might be useful? How would we use one if we had it?
- Goals:
  - Describe a model and how it can be **used**.
  - Know the different model types: **distribution** models or **sample** models
  - identify when to use a distribution model or sample model

Video 2: Comparing Sample and Distribution Models

- If you could have either, which one might you **prefer**? It depends ....
- Goals:
  - Describe the **advantages** and **disadvantages** of sample models and distribution models
  - explain why sample models can be represented more **compactly** than distribution models.

Video 3: Random Tabular, Q-planning

- **A simple planning method.** Assumes access to a sample model. Does Q-learning updates
- Goals:
  - **You will be able to explain how planning is used to improve policies**
  - And describe one-step tabular Q-planning

Video 4: The Dyna Architecture

- **Introducing Dyna!** An architecture that mixes (1) learning a model, (2) updating the value function and policy as usual, and (3) planning
- Goals:
  - understand how **simulating experience** from the model **differs** from i**nteracting with the environment.**
  - You will also understand how the **Dyna** architecture **mixes direct RL** updates, and **planning** updates.

Video 5: The Dyna Algorithm

- The details about one implementation of the Dyna Architecture: **Dyna-Q**
- Goals:
  - Describe how Tabular Dyna-Q works.
  - **Identify** the direct RL, planning, model learning, and search control parts of Dyna-Q.

Tabular Dyna-Q

Video 6: Dyna & Q-learning in a Simple Maze

- Use a small gridworld to compare Tabular Dyna-Q and model-free Q-learning. **We run an experiment!**
- Goals:
  - describe how learning from both real and model experience impacts performance
  - explain how a model allows the agent to learn from **fewer interactions** with the environment.

Video 7: What if the model is inaccurate?

- How do we handle if the model is **wrong in some way**? How could that happen? What would be the impact of trying to plan with an inaccurate model?
- Goals:
  - Identify **ways** in which models can be inaccurate,
  - Explain the **effects** of planning with an inaccurate model
  - Describe how Dyna can plan successfully with an incomplete model

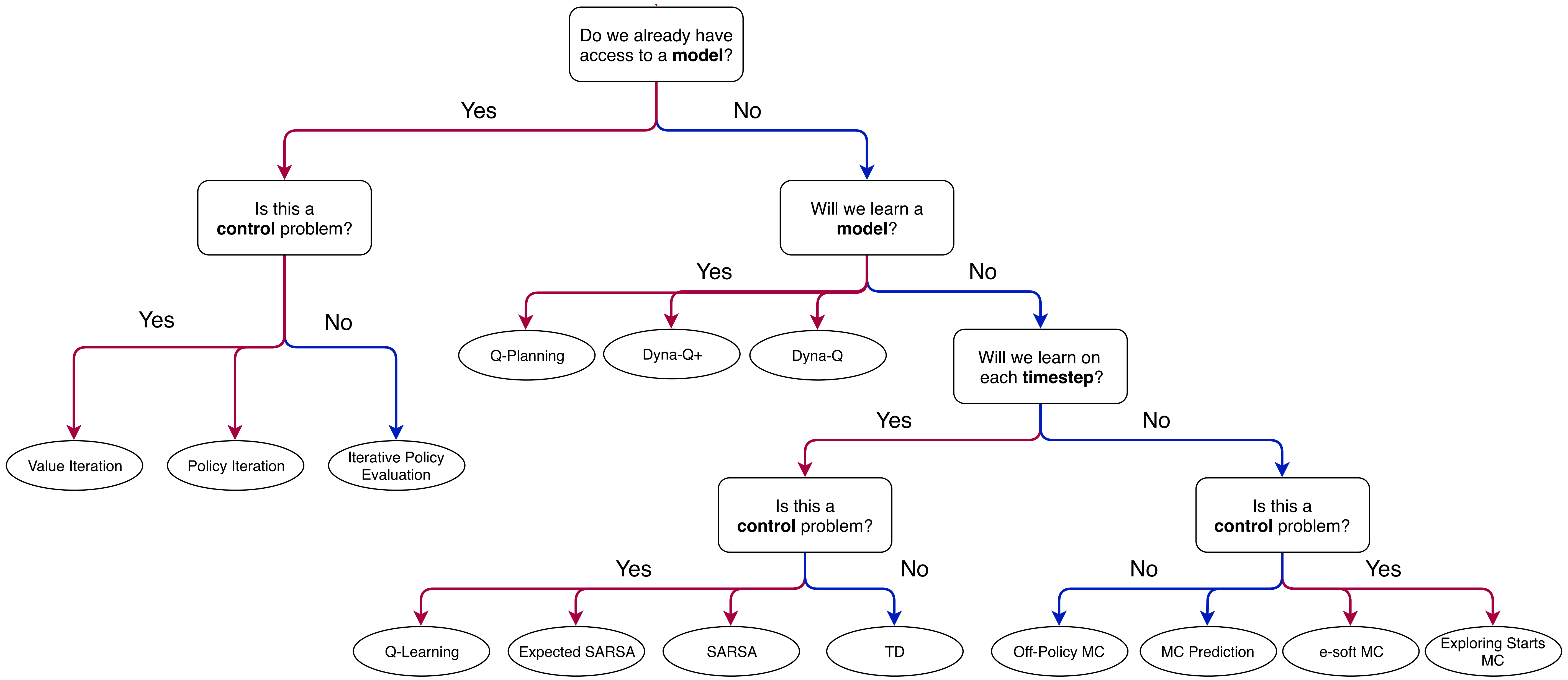
Video 8: In-depth with changing environments

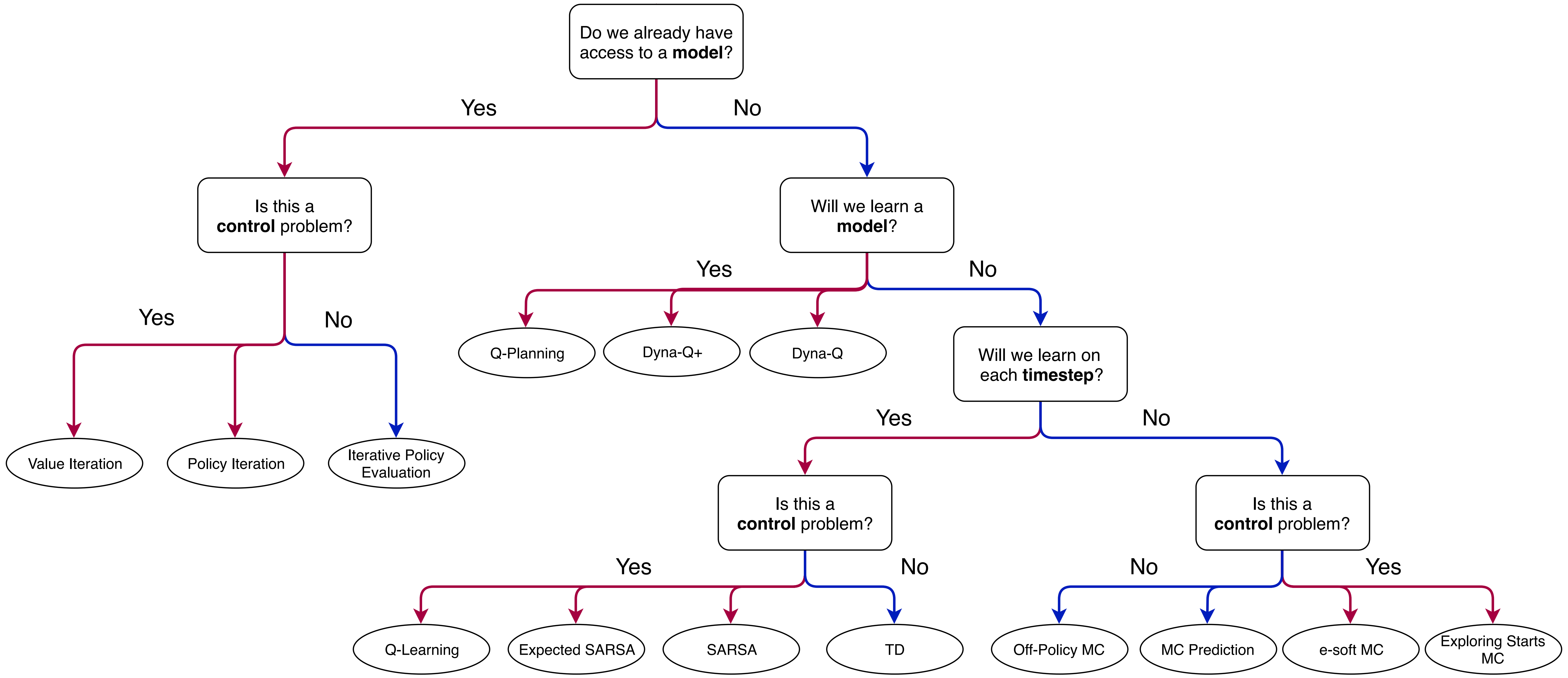
- We focus on a specific way the model can be inaccurate: **the world changes** and the model is out of date. **New Algorithm!**
- Goals:
  - Explain how model inaccuracies produce **another exploration-exploitation trade-off**
  - Describe how **Dyna-Q+** addresses this trade-off.

So many algorithms! What is a student to do?

- Introducing the **Course Map**

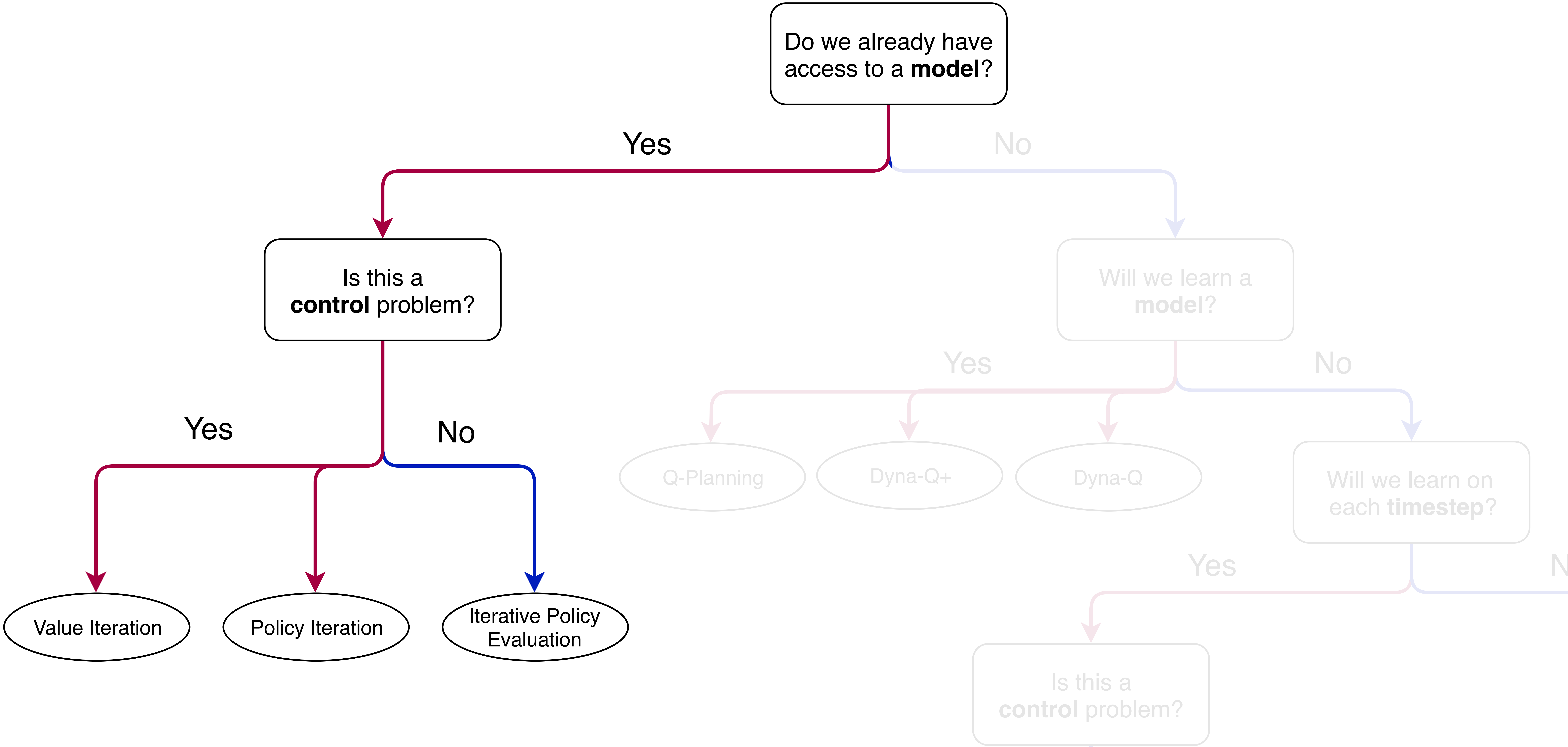

Can we represent the value function with a table?
No
Yes
Are we using average reward?
Do we already have access to a model?
No
Yes
Yes
No
Will we learn on each timestep?
Are the actions continuous?
Is this a control problem?
Will we learn a model?
Yes
No
Yes
No
Yes
No
Yes
No
Is this a control problem?
Gradient Monte Carlo
Gaussian Actor-Critic
Softmax Actor-Critic
Differential Semi-Gradient SARSA
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each timestep?
No
Yes
Semi-Gradient TD
Expected SARSA
Q-Learning
SARSA
Value Iteration
Policy Iteration
Iterative Policy Evaluation
Yes
No
Is this a control problem?
Is this a control problem?
Yes
No
No
Yes
Q-Learning
Expected SARSA
SARSA
TD
Off-Policy MC
MC Prediction
e-soft MC
Exploring Starts MC


Do we already have access to a **model**?
Yes
No
Is this a **control** problem?
Will we learn a **model**?
Yes
No
Yes
No
Value Iteration
Policy Iteration
Iterative Policy Evaluation
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Yes
No
Is this a **control** problem?
Is this a **control** problem?
Yes
No
No
Yes
Q-Learning
Expected SARSA
SARSA
TD
Off-Policy MC
MC Prediction
e-soft MC
Exploring Starts MC



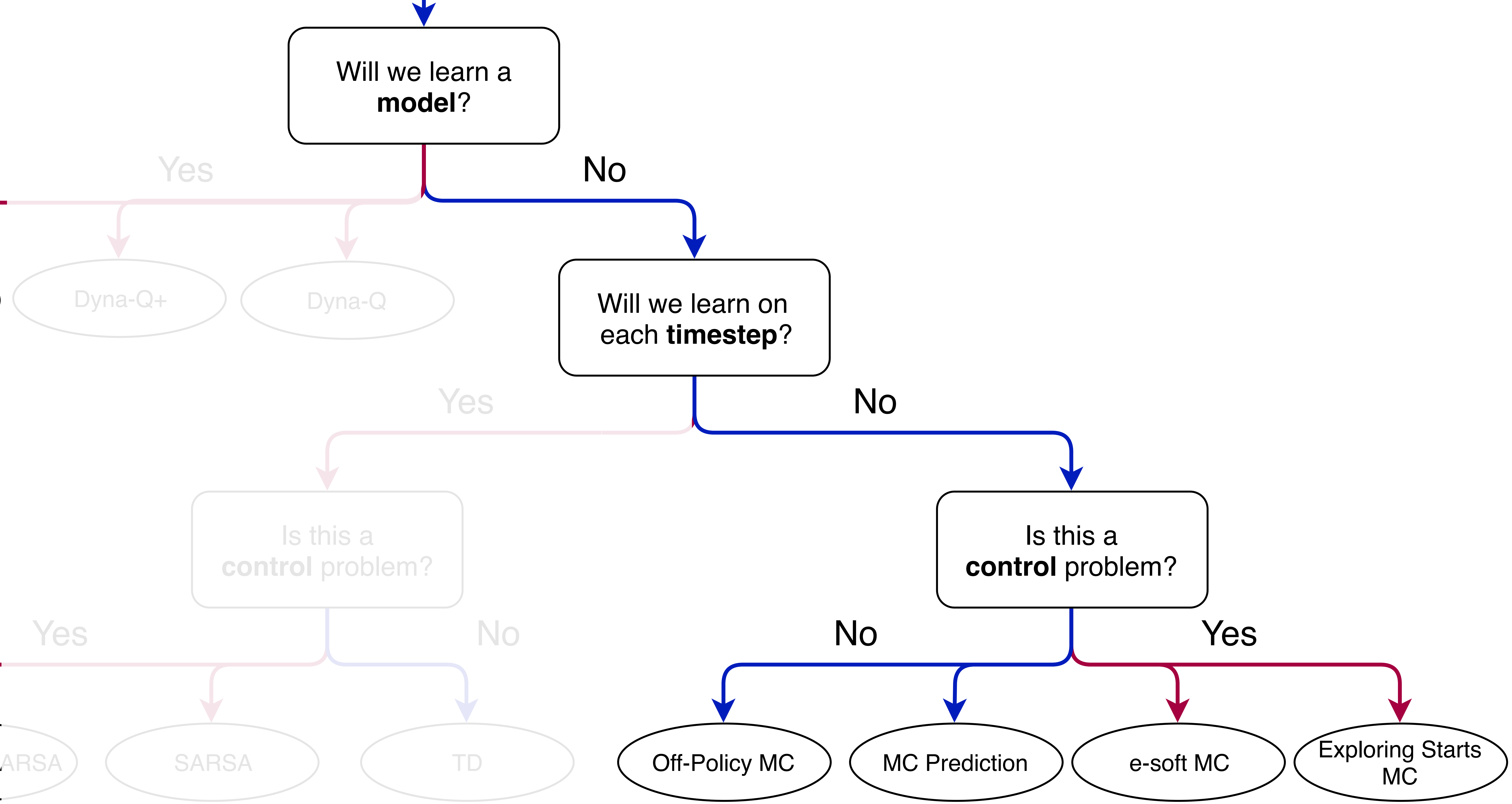
Do we already have access to a **model**?
Yes
No
Is this a **control** problem?
Will we learn a **model**?
Yes
No
Yes
No
Value Iteration
Policy Iteration
Iterative Policy Evaluation
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Yes
No
Is this a **control** problem?
Is this a **control** problem?
Yes
No
No
Yes
Q-Learning
Expected SARSA
SARSA
TD
Off-Policy MC
MC Prediction
e-soft MC
Exploring Starts MC


Do we already have access to a **model**?
Yes
No
Is this a **control** problem?
Will we learn a **model**?
Yes
No
Yes
No
Value Iteration
Policy Iteration
Iterative Policy Evaluation
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Yes
No
Is this a **control** problem?
Is this a **control** problem?
Yes
No
No
Yes
Q-Learning
Expected SARSA
SARSA
TD
Off-Policy MC
MC Prediction
e-soft MC
Exploring Starts MC



Do we already have access to a **model**?
Yes
No
Is this a **control** problem?
Yes
No
Value Iteration
Policy Iteration
Iterative Policy Evaluation
Will we learn a **model**?
Yes
No
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Yes
Is this a **control** problem?

Do we already have access to a **model**?
Yes
No
Is this a **control** problem?
Will we learn a **model**?
Yes
No
Yes
No
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Value Iteration
Policy Iteration
Iterative Policy Evaluation
Yes
No
Is this a **control** problem?
Is this a **control** problem?
Yes
No
No
Yes
Q-Learning
Expected SARSA
SARSA
TD
Off-Policy MC
MC Prediction
e-soft MC
Exploring Starts MC



Will we learn a **model**?
Yes
No
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Yes
No
Is this a **control** problem?
Yes
No
ARSA
SARSA
TD
Is this a **control** problem?
No
Yes
Off-Policy MC
MC Prediction
e-soft MC
Exploring Starts MC


Will we learn a **model**?
Yes
No
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Yes
No
Is this a **control** problem?
Yes
No
Q-Learning
Expected SARSA
SARSA
TD
Is this a **control** problem?
No
Yes
Off-Policy MC
MC Prediction
e-soft MC
Is this a **control** problem?
No
icy Iteration
Iterative Policy Evaluation


Do we already have access to a **model**?
Yes
No
Is this a **control** problem?
Will we learn a **model**?
Yes
No
Yes
No
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each **timestep**?
Policy Iteration
Iterative Policy Evaluation
Yes
No
Is this a **control** problem?
Is this a **control** problem?
Yes
No
No
Q-Learning
Expected SARSA
SARSA
TD
Off-Policy MC
MC Prediction

Terminology Review

- **Model:** a model of the environment. Anything that can predict how the environment will respond to the agent's actions: M(S,A) --->S',R
- **Planning**: the computational process that takes the model as input and produces or improves the policy
- **Sample Model**: a model that can produce a possible next state and reward, in agreement with the underlying transition probabilities of the world. We need not store all the probabilities to do this (think about epsilon-greedy)
- **Simulate:** sample a transition from the model. Given an S and A, ask the model for a possible next state S' and reward R
- **Simulated Experience:** samples generated by a sample model. Like dreaming or imagining things that could happen
- **Real Experience:** the states, actions, and rewards that are produced when an agent interacts with the real world.
- **Search Control:** the computational process that selects the state and action in the planning loop

Finish Variance derivation

- Back to understanding expectation and variance of updates

5. In this question we compare the variance of the target for Sarsa and Expected Sarsa. Recall the update for Sarsa is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$$

and for Expected Sarsa is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a') - Q(S_t, A_t)\right].$$

(a) Start by comparing the part of the update that is different: $Q(S_{t+1}, A_{t+1})$ compared to $\sum_{a' \in \mathcal{A}} \pi(a'|S_{t+1}) Q(S_{t+1}, a')$. Write down the variance for these two terms, given $S_{t+1} = s'$.

$$\mathrm{Var}(Q(s', A_{t+1})) \qquad \text{and} \qquad \mathrm{Var}\left(\sum_{a' \in \mathcal{A}} \pi(a'|s') Q(s', a')\right)$$

Conclude that the variance is zero for Expected Sarsa, but likely non-zero for Sarsa. Notice that the only random variable is A_{t+1}, which is the action selected according to the target policy π with distribution $\pi(\cdot|S_{t+1})$.