1. Which of the following algorithms are appropriate in a control setting in which updates will be made at every time step? [Select all that apply] **1 / 1 point**

- [x] Q-learning

Correct

Correct! Q-Learning uses temporal difference learning updates that are done at every time step with (state, action, next state, reward) transition tuples where the target is the sum of the reward and the max over the action values at the next state.

- [x] SARSA

Correct

Correct! SARSA uses temporal difference learning updates that are done at every time step with (state, action, next state, reward, next action) transition tuples where the target is the sum of the reward and the action value of the next action at the next state.

- [x] Expected SARSA

Correct

Correct! Expected SARSA uses temporal difference learning updates that are done at every time step with a (state, action, next state, reward) transition tuples where the target is the sum of the reward and the expected action value of the next state.

2. Which of the following algorithms are appropriate in a prediction setting in which updates will be made at the end of each episode? [Select all that apply] **1 / 1 point**

- [x] Off-Policy Monte-Carlo

Correct

Correct! Off-Policy Monte Carlo can be used to estimate the value function with respect to a target policy with experience from some behavior policy. The targets are empirically observed returns by waiting till the end of episodes.

- [x] Monte-Carlo Prediction

Correct

Correct! Monte Carlo can be used to estimate the value function with respect to a given policy with experience from the same policy. Thus, it solves a prediction problem. The targets are empirically observed returns by waiting till the end of episodes.

- [ ] Exploring Starts Monte-Carlo

3. Which of the following algorithms are appropriate in a tabular setting in which we will be learning a model and using it for planning? [Select all that apply] **1 / 1 point**

- [ ] Expected SARSA
- [x] Dyna-Q

Correct

Correct! Dyna-Q uses a model to learn from both simulated and real experience and planning is done by making queries to the model.

- [x] Dyna-Q+

Correct

Correct! Dyna-Q+ uses a model to learn from both simulated and real experience and planning is done by making queries to the model. In addition, Dyna-Q+ can handle non-stationarity in environment well by making use of an exploration bonus to visit long unvisited states and ensure that action-values are up-to-date across the MDP.

4. Which of the following algorithms are appropriate in a control setting in which we are given access to a model? [Select all that apply] **0 / 1 point**

- [x] Policy Iteration

Correct

Correct! Policy iteration is a method of computing an optimal policy by iteratively finding the value function corresponding to a given policy and then improving that policy. In order to do so, it makes use of the transition probabilities and reward function of the MDP or, equivalently, access to a model. Thus, it is an appropriate algorithm for a control setting with access to a model.

- [ ] Iterative Policy Evaluation
- [x] Value Iteration

Correct

Correct! Value iteration is a method of computing an optimal policy and its value by first finding an optimal value function first and then extracting a policy. In order to do so, it makes use of the transition probabilities and reward function of the MDP or, equivalently, access to a model. Thus, it is an appropriate algorithm for a control setting with access to a model.

- [ ] Dyna-Q

You didn't select all the correct answers

5. Which of the following algorithms are appropriate in a continuing control setting with a discrete action space and function approximation? [Select all that apply] **1 / 1 point**

- [ ] Gaussian Actor-Critic
- [x] Differential Softmax Actor-Critic

✓ **Correct**

Correct! Differential softmax actor-critic uses function approximation to parameterize a state-conditional categorical distribution over actions. Thus, it is appropriate for a discrete action space setting with function approximation. Differential actor-critic methods are also appropriate for the continuing setting and aim to find a good policy that maximizes average reward.

- [x] Differential Semi-Gradient SARSA

✓ **Correct**

Correct! In the average reward setting, differential semi-gradient SARSA finds a near optimal action-value function and hence policy with function approximation. Hence, it is appropriate for a continuing, control setting with function approximation. With linear function approximation with action-values being learned for all the discrete actions, differential semi-gradient SARSA is also appropriate for discrete-action spaces.

6. Which of the following algorithms are appropriate in an online prediction setting with linear function approximation? [Select all that apply] **1 / 1 point**

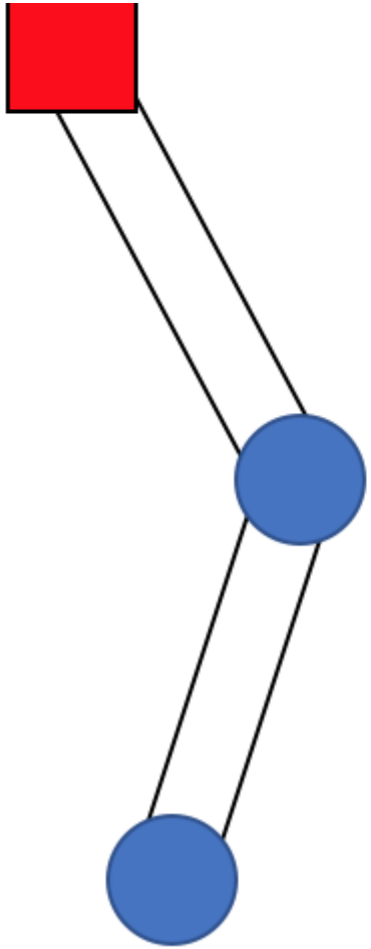
- [x] Semi-Gradient TD

✓ **Correct**

Correct! Semi-gradient TD can use linear function approximation and temporal difference learning style updates at every time step.

- [ ] Gradient Monte Carlo
- [ ] SARSA

7. In the **continuing** acrobot system (shown below), a double pendulum is fixed to the red square. The goal is to swing the double pendulum such that the height of the mass on the end of the lower pendulum exceeds the height of the black line. When the goal is reached, a reward of one is given and the double pendulum transitions to a vertical position. Otherwise, the reward is zero. Which of the following algorithms are appropriate for control in this context? [Select all that apply] **1 / 1 point**



- [ ] Expected SARSA
- [ ] Q-learning
- [x] Average Reward Actor-Critic

Correct

Correct! Acrobot (as we described it) is a continuing task, which means that we should be using average reward.

8. Which of the following algorithms are appropriate for control in the lunar lander MDP, as it is described in the lecture “Initial Project Meeting with Martha: Formalizing the Problem”? [Select all that apply] **1 / 1 point**

- [x] Expected SARSA

Correct
Correct! Expected SARSA can be used in an episodic setting.

- [ ] Average Reward Actor-Critic
- [x] Q-learning

Correct
Correct! Q-Learning can be used in an episodic setting.